Department
for Work &
Pensions

GSR
GOVERNMENT SOCIAL RESEARCH
Social Science in Government

Evaluating the Impact of Universal Credit on the Labour Market in Live Service and the North West Expansion

July 2014

DWP ad hoc research report no. 7

A report of research carried out by the Department for Work and Pensions

This document/publication is also available on our website at: https://www.gov.uk/government/collections/research-reports

If you would like to know more about DWP research, please email: Socialresearch@dwp.gsi.gov.uk

First published 2014.

ISBN 978-1-78425-291-5

Views expressed in this report are not necessarily those of the Department for Work and Pensions or any other Government Department.

Introduction

Universal Credit marks a major shift in the State's approach to delivering welfare. To maximise our learning about what works and to understand the impact of a change on this scale requires a significant programme of robust research and analysis. Key to determining the overall success of Universal Credit is measuring its employment impact. While we are committed to evaluate and learn from all aspects of UC Live Service[1], this paper focuses solely on measuring our ambition to deliver better labour market outcomes.

Universal Credit roll-out plans for 2014-15 will test the impact of UC in a live environment by extending the live service across new claimant groups including couples and more complex families with children and across the whole of North West England. Increasing the scale and scope for UC eligibility in a controlled way gives us greater opportunity to learn more and more quickly by testing that processes and systems work for a range of different claimant groups. We want to use this opportunity to understand and measure the contribution of UC in supporting our ambitions to deliver better labour market outcomes. For that reason, we have invested significant time and effort into developing our labour market analysis by drawing on expert advice from leading academics and researchers.

This paper describes how we plan to evaluate the impact of Live Service and the North West expansion on the labour market. The proposed approach strongly reflects the independent peer review of these plans by researchers at the Institute for Fiscal Studies (IFS). IFS have provided a number of invaluable pointers towards refining our analytical approach through their report: "Evaluating the labour market impacts of Universal Credit: a feasibility study". As IFS note in their peer review of our proposed evaluation plans, our approach has "incorporated the major points and caveats raised in [our] feasibility study". And, within the boundaries of what we are hoping to evaluate under the current North West expansion plans for UC, IFS note that " in most respects the report is excellent, the proposed evaluation strategy is wholly appropriate…".

What are we evaluating?

Delivering better labour market outcomes is a core aim of UC. Through our analysis we want to find out what difference UC makes to new claimants':

a) chances of being in work;
b) time in work; and
c) earnings.

[1] https://www.gov.uk/government/publications/universal-credit-evaluation-framework

How are we evaluating this?

To find out what difference UC makes to these outcomes we need to know what new UC claimants' outcomes would have been had they claimed the equivalent legacy benefit instead of UC. The problem is that we do not know what they would have achieved under the legacy system. We only see their UC outcomes. So, we have to *estimate* what outcomes they would have achieved under the legacy system. We do this by looking at the outcomes of a comparison group, i.e. the outcomes achieved by similar people who claim legacy benefits rather than UC.

The phased introduction of UC by geography and time means we can construct a comparison group in two ways. First, we can look at similar people who make a similar new claim at the same time but who do not claim UC because of *where* they claim. Second, we can look at similar people who made a similar new claim in the same area but did not claim UC because of *when* they claimed, i.e. they claimed before UC was introduced in that area.

Our strategy focuses on identifying a comparison group using geographical variation, i.e. finding similar people from comparable areas who make equivalent new legacy claims at the same time. This reflects the fact that it will be difficult to identify truly comparable time periods because of seasonal and cyclical factors.

We will use various statistical matching methods to ensure that the comparison group we construct is as close as possible in all relevant respects to the UC claimants. That is, the comparison group will be the same in terms of observable personal and labour market characteristics that might affect their labour market outcomes. The better the quality of the matching the more certain we will be that any differences in outcomes are due to UC (i.e. because there would be no other differences between the groups that could lead to them achieving different outcomes).

Will it Work?

The phased roll-out across the North West combined with the rich administrative data we have on all new claimants means the chances of being able to identify a large number of high quality matches and therefore obtain reliable and precise estimates is extremely high. That is, we will have lots of new legacy claimants from which we can select only those who have the same characteristics and are facing similar local labour markets to the UC claimants. The only relevant difference between the comparison group and the treated UC group will be that they are claiming a legacy benefit because of where they make their claim. The large sample sizes also provide more scope to find out whether the impact of UC differs between different demographic groups. We will also combine matching with differences-in-differences to control for any differences between the UC and legacy claimants that we cannot observe or measure but which are constant over time.

Similar methods and data sources have been successfully applied to evaluate other DWP programmes in recent years. Whilst propensity score matching combined with differences-in-differences is our preferred method and the most appropriate given the roll-out and data available we will use alternative methods to explore how sensitive the results are to different assumptions and approaches.

Conclusions

The roll-out schedule and administrative data available means that we should be able to obtain reliable and precise estimates of what difference UC makes to the labour market outcomes of new UC claimants during the early phases of the UC roll-out. This reflects that there will be a lot of people who will remain outside UC for a relatively long time in a range of places. The administrative data we have will enable us to construct a large matched comparison group to isolate the impact of UC from other things that affect claimants' labour market outcomes. We will combine matched methods with other evaluation approaches both as part of our sensitivity analysis and to improve the precision of our estimates.

We are committed to developing a rigorous assessment of labour market outcomes. As we work through the detailed analysis, we will continue to draw on expert external advice to quality assure our work. We are confident that this approach will allow us to effectively and robustly evaluate the impact of UC on labour market outcomes.

Evaluating the Impact of Universal Credit on the Labour Market in Live Service and the North West Expansion of Live Service

Introduction

This paper describes how we plan to evaluate the impact of Universal Credit in Live Service and during the North West expansion on the labour market. It has 8 parts covering:

i) Policy Background
ii) Evaluation Strategy
iii) The Evaluation Question
iv) Evaluation Methods
v) Data Sources
vi) Other Issues
- Entry and Anticipation Effects
- Long and Short-Term Impacts
- Timescales
vii) Longer-term Evaluation Plans
viii) Conclusions

It draws extensively on a feasibility study by researchers at the Institute for Fiscal Studies. An earlier version of this paper was comprehensively peer reviewed by IFS. Their review concluded that our approach has "incorporated the major points and caveats raised in [our] feasibility study". And, within the boundaries of what we are hoping to evaluate under the current North West expansion plans for UC, IFS note that "in most respects the report is excellent, the proposed evaluation strategy is wholly appropriate…". We have updated our plans in light of all the issues IFS raised in their peer review. A copy of IFS's peer review of an earlier version of our evaluation plans is at Annex A.

i) Policy Background

Universal Credit is a major feature of the Welfare Reform Act. It aims to simplify the current benefits system to make work pay. It is an integrated, income-related, working age credit providing households with a basic allowance topped up by additional components to recognise the needs of families with children, housing costs, disability and health conditions that limit work, and caring responsibilities. It will be available both in and out of work and will replace six major means-tested benefits and tax credits for people of working age:- Working Tax Credits (WTC), Child Tax Credits (CTC), Housing Benefit (HB), Income Support (IS), income-based JSA (JSA) and income-related Employment and Support Allowances (ESA). The aim is to strengthen work incentives (both financial and non-financial) so that more people work and people work more hours.

Universal Credit will simplify the benefit and tax credit system to make the system easier to operate and understand helping to reduce fraud, error and administrative costs. As an integrated in and out of work credit, Universal Credit removes the previous distinction between in and out of work benefits, so anxiety on the part of the customer on moving around the wider benefit system is reduced, there will be no re-application involved, and there will be no administrative barrier to the customer entering employment or changing their hours.

Universal Credit was introduced for single people without children who met certain other conditions (e.g. no housing benefit claim) who made simple new claims to JSA in:

- Ashton-Under-Lyne from April 2013
- Wigan from 1 July 2013; and
- Warrington and Oldham from the end of July 2013.

This first phase of Universal Credit is called Pathfinder. The purpose of Pathfinder was to test processes and delivery before rolling out UC to more areas and more claimant types. Whilst the primary purpose was to ensure delivery works we still want to know what impact it has, if any, on *labour market outcomes*.

Pathfinder has become known as Live Service as it has been extended to six more offices:

- Hammersmith from October 2013, Rugby and Inverness from November 2013;
- Bath and Harrogate from February 2014; and
- Shotton from April 2014.

The Government is committed to expanding the delivery of UC on the current platform in more offices and more claimant types across the North West. We have already introduced UC for couples in some offices and we will continue to expand the scale across the North West and extend the scope by introducing families with children and more complex claims later this year. Key points to note:

- UC is only replacing new claims by jobseekers and lone parents making new Income Support (IS) claims during these early phases of the roll-out.
- Once a person makes a new claim to UC they will be subject to UC from that point onwards.
- UC is not being introduced at all over this period anywhere else outside the North West except for the original Live Service offices. Therefore, new claims in other places by similar people will remain under the full legacy system for some time.
- There will be variation in exposure to UC across the North West as some offices replace new JSA/ISLP claims sooner and/or more quickly than others.

ii) Evaluation Strategy

A policy and delivery change as large as Universal Credit requires a significant programme of monitoring, research, evaluation and analysis, both to enable us to report on the extent to which it has achieved its aims, but also to understand how we can improve on the design and delivery. Delivering on learning is a top priority for the expansion of Live Service.

DWP has developed and published an evaluation framework that represents the first step in the development of the full evaluation programme of Universal Credit.
http://www.dwp.gov.uk/docs/universal-credit-evaluation-framework.pdf

The framework sets out the Department's broad intentions for the evaluation, highlights the key aims and objectives and considers possible analytical approaches.

We are designing our approach to testing and evaluation on the back of a theory of change for UC. Theory of change starts with the assumption that a policy operates in a changing economic and social context, and that the people involved in delivering and experiencing the policy are subject to variable choices and a variable capacity to act. It is the combination of the behavioural context and the policy levers which drive outcomes. Applying theory of change to UC evaluation involves unpacking the underpinning theories behind the policy. We can use this to help us evaluate, shape and fine-tune UC with a better understanding of why we are – or are not – seeing the outcomes we expected. It can also give us early insight into the effectiveness of UC when robust outcome measures can often take time to measure. For example, we are using a range of methods including surveys and interviews with claimants, staff and other stakeholders and we are monitoring a range of management information to assess the extent to which the policy is being implemented as intended and to find out how it is affecting the attitudes and behaviours that will drive changes in outcomes.

Our evaluation efforts will focus on:

- Testing delivery and understanding of policy intent.
- Testing changes in attitudes and behaviours.
- Testing final outcomes and impacts. The focus of this paper.

This paper focuses on our plans to evaluate the labour market impacts of UC under Live Service and the expansion North West expansion.

iii) The Evaluation Question

Anyone of working age will, once UC is rolled out, be potentially eligible for UC. Introducing UC should not affect the size or composition of the working age population. Therefore, arguably we should evaluate the impact of UC on the working age population. However, the working age population is very large and diverse. Many working age people will be largely or wholly unaffected by UC. This reflects that UC is not targeting the whole working age population. Focusing on the whole working age population would dilute any impacts UC has and make them more difficult to detect. Therefore, we think it is better to evaluate the impact of UC on sub-sets of the working age population who will be significantly affected by the introduction of UC (either intentionally or unintentionally). This way we will be more likely to identify the impacts that UC has.

One sub-set of the working age population who will unambiguously be affected by the introduction of UC is those making **new claims** at a time and place when and where that type of new claim has been replaced by UC. The phased roll-out by geography and time provides scope to estimate the impact of UC on new claimants (see below). The impact on new claims is relevant to the longer-term impact of UC in steady-state (notwithstanding that the impact on new claims might evolve over time for various reasons and the impact in these early phases might differ from the impact when UC exists everywhere for all claim types).

Due to the feasibility of assessing the impact on new claims and importance of this impact our evaluation will focus primarily on estimating what impact UC has on the labour market outcomes of people who make a new claim to the full Universal Credit system that results in an award[2] in Live Service offices or under the North West expansion instead of making an equivalent new claim under the full legacy benefit system. In this case, the "treated" group are those making new claims that result in an award under UC. The "untreated" are people making new claims that lead to an award under the legacy system. Strictly the "untreated" group receive a *different* treatment – the status quo treatment – rather than no treatment.

We plan to estimate the average impact of UC on the treated (ATT). This reflects that, for a long time, we will have relatively few treated people (new UC claims) relative to the number of new legacy claimants. This makes it easier to find individuals from the very large number of new legacy claimants who are similar to the new UC claimants than it will be to find people who are similar to the new legacy claimants from the comparatively small number of new UC claims. Consequently, it is more feasible to estimate the impact on the treated rather than the impact on the non-treated.

Since Live Service and the North West expansion only replace new claims to JSA and ISLP we are only focusing on the impact of UC on these types of new claims. These are important groups both in terms of policy importance and size.

For new awards we plan to estimate the impact of UC on:

a) Time in work during different periods after the new award started, e.g. time in work during the three months after the award started.
b) Gross earnings during different periods after the new award started.
c) The proportion of people in work (i.e. the employment probability) at different points in time after the new award started.

Unit of analysis

We will evaluate the impact of UC on **individuals**. This reflects that our data typically relates to individuals. However, we will match individual data and estimates to recover impacts on couples as well. This carries risk. Data on UC claims should already be split by household type given the household nature of UC and plans to roll-out partly by household type. However, for legacy benefits we will have to rely on being able to identify household composition based on existing individual level administrative data on legacy benefits. The extent to which this is possible is unknown at this stage. We will not have information on household members outside the benefit unit making the claim. So, we might be able to estimate impacts on the benefit unit (family) level but this is not necessarily the same as the household.

[2] We only get data on the subset of new JSA claims that lead to awards or payments. This is why we can only estimate the impact on new claims which lead to new awards.

We will separately estimate impacts for different demographic groups, e.g. by age group, gender, ethnicity, etc. Depending on our ability to successfully identify these groups in existing administrative data we will also seek to estimate impacts separately for new claims by:

- Single people with no children;
- Couples with no children;
- Couples with children; and
- Lone parents.

Limitations of focusing on new claimants

The main limitation of focusing on new claimants is that we expect UC to affect take-up. Changes to eligibility and entitlement mean that UC will change who claims and the types of claims that *some* people make. Other factors such as awareness, attitudes, differences in in-work support and conditionality regimes might also affect take-up. This means UC will change the composition of new claims through entry and/or anticipation effects.

People who might have claimed the equivalent legacy benefit before and formed part of the comparison group might decide to delay their claim in anticipation of UC becoming available in their area. This could lead to a difference in the composition of new claimants between the treated and non-treated groups, which could mean that they would achieve different outcomes even in the absence of a different treatment. If we can observe these differences it does not affect the reliability of the estimates. However, it would still reduce external validity since we can only evaluate the impact on the sub-set of new UC claims who would have made a new claim under both systems in the same circumstances – common support (see below for more details). If the compositional differences are unobservable and affect outcomes then the estimates would also be internally invalid (they would be biased).

Anticipation effects may be more significant when we compare sites within the North West to evaluate the impact of UC than when we identify comparison groups from outside the North West. This reflects that UC will not be available for a reasonably long time outside the North West.

Entry effects might change the composition of the *treated* group if some people decide to claim UC who would not have claimed the equivalent legacy benefit. This would affect the internal validity of the estimates because the treatment and control or comparison groups will be different in ways that might affect the outcomes they achieve. This type of entry effect might be relatively unimportant in these early phases of the UC roll-out because people will initially be unfamiliar with the details (or even existence) of UC. Nevertheless, they are something that the evaluation will need to explore by looking at the composition of new claims over time.

Another type of entry effect is mechanical (rather than behavioural) and could arise because UC is being phased in by benefit type (as well as by geography) and once a person claims UC they will always be under the UC system from then onwards. This is sometimes referred to as the “lobster-pot”. This means that for legacy benefits that are replaced by UC later we can only estimate the impact of UC on the labour market outcomes of those new claims that have not already entered UC via another route.

A similar mechanical entry effect could arise because UC is being phased in by household type. So, UC is first replacing new JSA claims for single people without children and without housing benefit, etc. However, as these people's circumstances change they may continue to claim UC even though they would have been ineligible for UC if they not already made a UC when they were eligible. Again, this means for each type of household group we can only estimate the impact on the sub-set of new claimants within that household type who have not already entered UC by another route.

We expect that mechanical entry effects will also be minor for Live Service and the North West expansion as UC is only replacing JSA/ISlp. Moreover, offices that roll-out UC later in the NW expansion will be introducing UC at the same time for all new claims to JSA/ISlp irrespective of household type.

A third type of entry effect is analogous to anticipation effects but involves the treatment rather than the comparison group. That is people could choose to accelerate or delay a claim around the time a new roll-out phase is introduced, to affect which regime their claim falls under. These are unlikely to be problematic because we can select the treatment sample from a time well away from the introduction of UC in the comparison areas.

So, overall we expect to be able to estimate the impact of UC on most new claims to JSA and ISlp during Live Service and the North West expansion because:

a) relatively few people making new claims to UC during this time would have had the opportunity to enter UC earlier via another benefit type or eligibility route (partly because many offices UC will replace all new claims to JSA and ISlp at the same time);
b) the eligibility and entitlement criteria under both benefit systems will be similar for the types of new claims replaced by UC during this phase;
c) many people will not be aware of UC or familiar enough with it for it to produce entry effects from behavioural changes; and
d) we can select comparison samples far enough away from when UC is introduced.

The representativeness of the composition of on-flows may become more problematic when we come to evaluate the impact of UC replacing other types of new claims in these early roll-out offices during later stages of the UC roll-out. This reflects we will only be able to evaluate the impact on new claims to other legacy benefits/tax credits who have not already entered UC during the earlier phases in these offices when it replaced JSA and ISlp. There could still be scope for evaluating the impact of UC on a more representative sample of new claims to other legacy benefits in other parts of the country depending on how the national roll-out occurs.[3]

[3] However, for some new claimants (e.g. tax credits) a bigger issue is likely to be the changes in entitlement and take-up. These changes are likely to mean that any evaluation of the impact of UC replacing new tax credit claims (during the national roll-out) is likely to have to define the treatment group differently. That is it will not look at the impact on new claims because the people making new claims to UC will differ from those making new claims to tax credits. Following IFS's feasibility study we will explore the scope for evaluating aggregate and indirect effects in an effort to capture the effects on broader groups than new claimants especially for those types of claims where eligibility and entitlement will be significantly different under UC. The scope for doing so will depend on how the national roll-out occurs.

Whilst there are good reasons to expect entry and anticipation effects to be minor during these early roll-out phases we will investigate whether there are any changes in the composition of on-flows compared with what we would expect given what is happening in other areas and what has happened before.

Additional evaluation questions

If we just focus on the impact on new claims we might find that their labour market outcomes improve. However, their better outcomes could be offset or partially offset by worsening outcomes amongst other groups. Such substitution and displacement effects are likely to be more important in the short-run. In the longer-run the labour market should adjust to a bigger effective labour supply through downward pressure on wage growth. How important substitution and displacement effects are in the short-run will depend on economic conditions.

Looking at impacts on whole populations irrespective of whether they make a new claim or not (aggregate impacts) and the impacts on people not directly affected by the introduction of UC at a particular point in time (indirect impacts) are important because they encompass substitution and displacement effects. These are *general equilibrium effects*. They do not just focus on people whose incentives and behaviour changes because of UC. They also look at people who might not change their behaviour but whose outcomes might change because of the impact of UC on others.

We will explore the scope for identifying aggregate impacts of UC replacing JSA/ISlp claims under Live Service and the North West expansion versus maintaining the full legacy system elsewhere. Looking at aggregate impacts captures the impact of UC on everyone affected - not just those who would make new claims under both the UC and legacy systems. Impact estimates on sub-groups irrespective of whether or not they claim capture take-up effects.

We will focus on groups who we think are likely to be affected and on whom the impact might be significant. This makes it more likely that we will be able to detect estimates. For example, we could look at the impact on all people who have previously made claims to out of work benefits or tax credits irrespective of whether they make new claims in future.

Since these people are not necessarily making a new claim we will need to define outcomes differently. For previous claimants we can look at the impact on the average time in work and earnings over different periods, e.g. since the start or the end of their last claim. We can also look at the impact on employment rates and earnings levels.

Aggregate impacts in these early phases of the roll-out are about estimating the impact of having a partial or more complete UC system compared with the full legacy system elsewhere. It is about differences in UC coverage. It is only likely to be towards the latter stages of the North West expansion, if at all, that it will become possible to start looking at aggregate effects, i.e. when UC coverage becomes sufficiently large relative to the population of interest. Also, because aggregate impacts consider the impact on broader groups than those directly affected by UC we expect to need significantly bigger samples than we need to evaluate the impact on new claimants. There may be more scope for looking at aggregate impacts during the national roll-out if this happens at pace and if some areas still remain outside UC for some time.

Indirect impacts of UC are the impacts it might have on those not directly affected by UC at a particular time either because they are ineligible at that time or because they were already subject to UC at an earlier time. There could be some scope under Live Service and the North West expansion to evaluate the indirect impacts of UC replacing some new benefit claims versus maintaining the full legacy benefit regime (or some hybrid regime with more limited UC coverage) for those not directly affected by a particular phase of the reform because they are not eligible for UC or were exposed to UC earlier. This reflects that within the North West significant volumes of new claims will become eligible for UC whilst many other types of new claims will remain ineligible. There may also be some scope to look at the indirect effects on the groups who were eligible for UC earlier (because they met the relatively narrow eligibility criteria) as UC becomes available to broader and larger groups of claimants.

iv) Evaluation Methods

We want to know what labour market outcomes new UC claimants under Live Service would have achieved if they had instead, at the same time and in the same place, made a new claim to the equivalent legacy benefit under the full legacy system. We cannot see both outcomes for the same individual at the same time.

We have to estimate what outcomes the 'treated' - in this case new UC claimants - would have achieved had they not been treated (i.e. if they had made a new claim under the legacy system instead). This is the counterfactual outcome for the treated. We estimate it using the outcomes of similar *untreated* people. The difference between the actual observed outcome and their estimated counterfactual outcome gives the true impact of Universal Credit **IF** the only *relevant* difference between treated and untreated groups is that one claims UC and the other claims the legacy benefit. This is the **average treatment effect on the treated (ATT)**.

A relevant difference is one that affects outcomes. So, if the non-treated group are the same as the treated group in terms of everything that affects outcomes then their untreated outcomes will be an unbiased (right on average) and efficient (close to the true value) estimate of what outcomes the treated group would have achieved had they not received the treatment.

If the impact of UC is the same for everyone (homogeneous) then the impact on the treated is the same as the impact on the non-treated (ATNT), which would also be the same as the **average treatment effect (ATE)**. If the impact of UC varies then its impact on the outcomes of the untreated had they claimed UC instead of legacy benefits can be different to UC's impact on the treated. To estimate the impact on the non-treated we have to estimate their counterfactual outcome. That is, we want to know what outcomes new claims to legacy benefits would have achieved had they instead, in the same circumstances (time and place), made a new claim to UC. We estimate this using the outcomes of *similar* new UC claimants. During Live Service and the expansion across the North West there will be *relatively* few new

UC claimants for some time from which to identify matches for the legacy claimants. That is, it will be more feasible to estimate the ATT than the ATNT if the impact of UC varies.

We cannot observe similar non-treated individuals who make similar new claims to an equivalent legacy benefit in the same place and at the same time as individuals make new claims to UC under Live Service. Nevertheless, the phased roll-out of Universal Credit does give two possibilities for constructing non-experimental comparison groups.[4] We will be able to use these comparison groups to estimate the counterfactual outcomes of new UC claims. First, there will be similar people to those making new claims to UC who at the same time are making a similar claim but are not subject to UC because of ***where*** they make their claim (i.e. they are outside the NW and Live Service offices or they are in an office in the North West that has yet to introduce UC for new JSA/ISlp claimants). Second, there will be similar people to those making new claims to UC who have made a similar new claim in the same area but were under the legacy system because of ***when*** they made their new claim. So, there will be variation in who gets exposed to Universal Credit and who gets exposed to the legacy benefit system because of both **geography and time.**

If this variation was random then it would be equivalent to an experiment conducted at the office (rather than the individual) level. This would ensure that the untreated legacy claimants (who made their claims earlier or who made their claims in other places still outside UC) are the same in all respects to those making new claims to UC and that their outcomes would consequently give an unbiased estimate of the treated group's counterfactual outcome.

However, the roll-out is not random. It has been designed, in part, to be deliverable. This is only a problem for evaluation if it means the treated group will be different from the non-treated group in ways that affect their outcomes. This is likely to be the case here. Treatment status depends on *where* and *when* a claim is made, i.e. on the office and time. However, both the office and time affect the labour market outcomes new claimants achieve. For example, different offices have different types and numbers of claimants, operate in different local labour markets, different policy environments and have different levels of performance. So, two otherwise identical new claimants could achieve very different outcomes just because they make claims in different offices (or at different times). Therefore, the office and the area the office is in can lead to selection bias because they determine treatment status and affect outcomes. Exactly the same applies to time.

A priori we think it will prove more fruitful to compare **similar offices at the same time** than to look at the **same offices in similar periods**. This reflects that we think it will be difficult to identify truly comparable periods. For example, because labour market outcomes are seasonal it implies a comparable period for the same offices might be 12 months before the new claims are made to UC. Whilst this overcomes seasonal effects it simultaneously increases the risk that the outcomes of the treated would have been different anyway (i.e. even in the absence of UC) because of the economic cycle and the potential for differential trends between areas. Indeed, we anticipate that the strength of the association between time and outcomes could create considerable selection bias. In contrast, we think that the

[4] With non-experimental methods we use 'comparison group' rather than 'control group' to reflect that the comparator group is constructed non-experimentally and so there is no control that guarantees the comparison group will be comparable to the treated group.

opportunity in the short-term to compare early and late roll-out offices within the North West and in the longer-term to compare North West offices with offices from anywhere else in the country (except for the original Live Service offices) means it is more likely we will be able to identify similar offices at the same time.

Due to the non-random roll-out we have to use a non-experimental evaluation approach. There are several different non-experimental methods. They apply different statistical approaches to individual data to estimate the counterfactual outcome of the treated group. The methods differ in the number and nature of their assumptions. Unfortunately, we cannot empirically check these assumptions. However, we can often get contextual and indicative evidence about their likely plausibility. Different non-experimental methods produce different results because they use different approaches to estimate the counterfactual, they make different assumptions and sometimes they are simply estimating different impacts.

All methods seek to eliminate selection bias due to factors that affect both treatment status and outcomes. These confounding factors may be observable or unobservable; measurable or unmeasurable. As mentioned, when we look at the impact of UC on new claims during the phased roll-out confounding factors include time and office/area. They also include the personal characteristics of individuals, which determine whether or not they claim UC.

Cross-sectional non-experimental methods such as matching focus on eliminating selection bias due to factors that we observe in the data. We can combine them with differences-in-differences (see below) to get rid of bias from unobservable confounding factors that are constant over time.

If non-experimental methods are to work, i.e. if they are to estimate the true impact of UC on the treated, they have to make sure that the untreated comparison group is similar in all relevant respects to the treated group. This does not mean that they have to be identical. It means they have to be similar in terms of anything that might affect both:

a) whether they receive the treatment, i.e. claim UC rather than the legacy benefit; **and**

b) their labour market outcomes.

Things that affect only whether people receive the treatment (UC) but not their outcomes will not bias our estimates because they do not lead to different outcomes. Analogously, things that affect only outcomes but not the likelihood of receiving the treatment will be balanced between the treated and comparison groups and so will not bias the estimates either. Problems only arise when there are differences between the control and treated groups which mean that we would expect them to achieve different outcomes even if they received the same treatment.

We focus on the two non-experimental methods that we think are most likely to generate reliable estimates of the impact that UC has on new claimants' labour market outcomes (and aggregate and indirect impacts) given the nature of the roll-out and the data available. This assessment draws on IFS's feasibility study for evaluating UC, which also discusses why some other approaches would not help evaluate UC. Whilst IFS's study was based on a hypothetical roll-out scenario, many of the principles and issues remain the same.

Matching Methods

Matching methods seek to eliminate differences between the treated and comparison group that can affect their outcomes by re-weighting the observations in the comparison group so that it is the same in all relevant and observable respects to those in the treatment group. The object is to get a comparison group that has the same characteristics and the same distribution of these characteristics across the group as there is across the treated group. The idea is that, *conditional* on their *observable* characteristics we would expect the outcomes of the comparison group to provide an unbiased estimate of the outcomes the treated group would have achieved had they not received the treatment.

The big assumptions are that:

- Conditional on observable characteristics the potential outcomes of the two groups are the same. This means we have to be able to observe and measure everything that affects both outcomes and treatment.

- There are similar people making new claims to the equivalent legacy benefit in similar circumstances – common support.

If these assumptions hold we can simply use the mean outcome of the non-treated comparison group as an estimate of the mean counterfactual outcome the treated group would have achieved had they made their new claim to the legacy system rather than UC – treatment effect on the treated.

As with other partial equilibrium estimation methods (so when estimating the impact on new claimants rather than aggregate impacts) PSM assumes that:

a) an individual's participation decision does not depend on the decisions of others. So, when we look at the impact on new claimants we assume their decision to claim does not depend on the decision of others to claim; and

b) the impact on one person does not depend on who else is, or on how many others are, in the treatment group.

Will Matching Work?

We think matching will work because we will:

i) Measure outcomes in identical ways for the treatment and comparison groups primarily using Real Time Information;

ii) Sample treatment and comparison groups from similar locations or contexts. This reflects that there will be lots of offices outside UC from which we can select matched new claims to JSA and ISlp to estimate the counterfactual outcomes of the treated group;

iii) Have rich data on a range of individual and area level factors that affect both the outcome and the probability of being treated. This will ensure we get good matches and only compare observably very similar treated and un-treated individuals who in the absence of UC we would expect to achieve very similar outcomes.

Several recent evaluations of labour market interventions appear to have successfully applied matching methods using similar data sources.[5] Of course, without a robust experimental benchmark to compare against we cannot know for sure whether the estimates obtained are unbiased.

If we have a good understanding of what determines treatment and outcomes then matching is more likely to be effective. In thinking about the impact on new claims we know that (for common support) treatment depends on geography and time. We also know what sorts of things affect claimants' labour market outcomes, e.g. their labour market history, age, ethnicity, etc. So, arguably – with the data we have, our understanding of what determines treatment and outcomes – matching may be an effective method.

There are complicating factors in evaluating UC compared with other past DWP programmes and policies. UC is overhauling many systems and changing the way we collect and record data on claimants. This may make it more difficult to evaluate because we need the same data for the treated and non-treated groups. If the data is collected and recorded differently then these differences might contribute to a difference in outcomes. This comes back to wanting the treatment and non-treated groups to be the same in all relevant respects with the sole exception of their treatment status.

We do expect to have consistent data on outcomes from the RTI for JSA/ISlp and RTE for UC (see data section for more details). We should also have consistent historical information for matching. However, a concern is that we will only have a very limited ability to match on historical outcomes using the same (RTI and RTE) data that we will use to measure outcomes. Indeed, due to the build-up of RTI over time since April 2013 it may be that we do not have historical information for enough people to make good use of it and instead have to rely on other data sources (the Work and Pensions Longitudinal Study and the National Benefits Database) to build-up a picture of past employment and benefit claim history for matching purposes.

Individual characteristics are important. They affect the labour market outcomes people achieve. Moreover, in many labour market policies and programmes, especially voluntary ones, they strongly influence participation or treatment status. In focusing on the impact on new claims the participation decision is the decision to claim UC. This decision is only open

[5] Ainsworth, P. and Marlow, S. (2011) "Early Impacts of the European Social Fund 2007-13" DWP In-House Research No 3
Marlow, S. Hillmore, A. and Ainsworth, P. (2012) "Impacts and Costs and Benefits of the Future Jobs Fund" DWP
Ainsworth, P. Hillmore, A. Marlow, S. and Prince, S. (2012) "Early Impacts of Work Experience" DWP

to people who meet certain eligibility criteria and are making a new claim of a particular type, in a place and at a time where and when that claim for that type of individual would be covered by UC. If people are ineligible because of their personal characteristics, the type of claim, the place or time of their new claim then they have to claim the equivalent legacy benefit if they want to make a new claim.

We have already highlighted that UC might affect take-up. This means there may be a lack of common support. That is, there may be some people making new claims to UC for whom there are no comparable people making the equivalent new claim to the legacy system. It is likely that any differential take-up will partly reflect people's personal characteristics. Therefore, to ensure we are comparing like with like it is important to match on personal characteristics. If any differential take-up is partly due to people's unobserved personal characteristics then this would invalidate the assumption that conditional on observable characteristics the potential outcomes of the two groups are the same.

Matching individuals on detailed and extensive pre-treatment benefit, employment and earnings histories (i.e. their pre-treatment outcomes) is particularly important for welfare and active labour market policies. Some argue that these capture otherwise unobservable key characteristics such as motivation that can predict both employment outcomes and treatment status.

In evaluating the impact of UC on new claims we are already conditioning, i.e. only comparing people who are making equivalent new claims. So, at each stage of the roll-out across the North West we will only compare new claims to UC with new claims to JSA and ISlp that would have satisfied the eligibility criteria for UC had they been made at the same time and place as the UC new claims. However, this is just one part of the matching process. It would be wrong to compare everyone making new claims to UC with everyone making equivalent new claims to the legacy system. Instead, we will select a sub-sample of non-UC new claims that are likely to be more comparable to the UC treated group. We plan to match more closely on personal characteristics as we know these help determine outcomes and they could differ between the two groups.

We have already highlighted that where and when someone makes a claim is likely to affect their labour market outcomes ***and*** their treatment status. Matching will attempt to ensure that we estimate the counterfactual outcomes for the treatment group by only looking at the outcomes of similar people making the same type of claim at the same time in a *similar office* and *area*. What is important here are any area/office level factors that can affect outcomes of new claimants. This could include the past performance of the office, local labour market conditions and trends (e.g. unemployment rate, vacancy levels, etc.), the composition and size of different claimant groups, the policy environment, etc. The policy environment is very important. The evaluation needs to take into account and isolate the impact of UC from other changes happening around the same time. It might be necessary to match on additional observables related to these reforms to ensure comparability of areas.

The expansion of Live Service is occurring across the North West. If the North West is unique in ways that affect the labour market outcomes of new claims, e.g. because of the regional labour market, it will be difficult to identify suitable matched new claimants in comparable areas (i.e. because there would not be comparable areas in terms of the regional labour market). This is something we will explore carefully. We will include regional level information in the matching process as well as more local level and office information. There will also be some scope to make comparisons between early and late roll-out offices within the North West to get evidence around this issue. However, comparisons within the NW are limited to looking only at short-term impacts to avoid anticipation effects and because of the limited period in which some NW offices will remain outside UC. Moreover, the volumes might mean we can only reliably detect relative large impacts.

We will also match on other factors (aside from time and place) that affect the participation process. We will match on factors collected before treatment for things that could potentially be affected by the treatment such as health. We will also match on things that cannot be affected by the treatment (e.g. gender, age, ethnicity, etc.). Importantly we need to identify similar individuals making similar new claims who would meet the eligibility criteria for making a new claim to UC with the exception of the timing and/or location of their new claim.

When we estimate the impact of UC on couples with and without children and lone parents we will need to include couple/family level information as well as individual information. As mentioned earlier, we think this will be possible by merging individual level data to identify those in couples and those in families. However, there are risks. This is not something that we routinely do. If it proves difficult to accurately identify everyone in couples, it could reduce the internal and external validity of estimates.

Propensity Score Matching

We will use a large number of observed characteristics (at an individual and local level) to ensure we get good matches and so only compare observably very similar treated and un-treated individuals who in the absence of UC we would expect to achieve very similar outcomes.

Trying to match on many factors – because potentially there are lots of things associated with outcomes and treatment status – makes it very difficult to identify good matches on all the criteria. We will use Propensity Score Matching (PSM) to reduce this dimensionality problem. PSM estimates a propensity score for treated and untreated individuals. This score is their propensity to receive the treatment conditional on the observed variables. We then match individuals on this index. Rosenbaum and Rubin, 1993 showed that matching on a single index representing the probability of treatment given the observed variables could achieve consistent estimates in the same way as if we matched on all variables.

Propensity Score matching has two other main advantages over other regression-based non-experimental methods. Firstly, it emphasises and restricts the analysis to only estimating impacts on treated people for whom we can identify suitable matches in the non-treated sample. So, it only compares like with like. In contrast, regression methods extrapolate beyond common support.

Secondly, matching is non-parametric In particular, it does not make any restrictive assumptions about how outcomes are determined and about how the observables affect impacts. Once treated and non-treated samples are matched we can just compare mean outcomes as we would if we did an experiment. Strictly speaking, in its propensity score flavour matching becomes semi-parametric as it does involve estimating a model of participation. The main objective of this model is to ensure the treated and comparison groups are well balanced. The actual performance of the model is not regarded as important. Nevertheless, given that the model is trying to eliminate any selection bias by taking account of things that affect treatment and outcomes it intuitively means that the model should be able to explain some of the variation in the propensity to receive the treatment. If it cannot then this might reflect that there is little bias – i.e. the treatment and non-treated groups are well balanced. However, it could also mean that key factors are missing from the model, so selection bias could remain a problem.

Regression models also restrict the impact to be the same across everyone (this can however be relaxed using interaction terms). This reflects its parametric specification of the outcome equation – after adjusting for individual, household, office, area and time factors that might vary between the treated and untreated groups the model assumes the same average impact across everyone.

When using geographical variation we could perfectly predict the propensity to receive treatment by including area identifiers in the participation equation. However, this would preclude us from using matching methods to estimate impacts. Instead we will seek to eliminate any area differences between the treated and comparison group by:

a) Including area level information in the participation model to ensure that we only use matched individuals that face similar local labour markets and make claims in similar offices, etc.; and

b) Identifying matched offices and only drawing matched individuals from within these matched offices. We will identify similar offices using a range of methods. We will assess the best way of using these matched offices in terms of obtaining the best balancing of individual, office and regional characteristics between the treatment and comparison group.

With the controlled expansion of live service in the North West we have opportunities to identify similar claimants in similar offices within the North West where UC is introduced later to act as a comparison group. We will also be able to select similar claimants in similar offices from the rest of the country. This means we will have big populations of potential comparators to choose from. This increases the chances of being able to find reliable matches. It also increases the sample sizes we are likely to have which will increase the precision of our estimates. The fact that UC will not be rolled out to other parts of the country for some time also reduces the risk that anticipation effects will be problematic and allows us to look at longer-term impacts (although any comparisons drawn from within the North West will only allow us to estimate relatively short-term impacts).

The evaluation literature on PSM suggests that the matching methods used to construct a comparison group often do not have a substantive impact on the estimates obtained. Nevertheless, we will consider which methods are likely to be most appropriate for the particular context. For example, as the roll-out extends across the NW and the number of potential comparisons reduces especially in that part of the analysis that only makes within the NW we may need to match with replacement and adopt more relaxed rules about the closeness of the matches used. When we extend potential comparisons to include the rest of the country then we hope there will be a sufficiently large number of potential comparisons available to use more methods and stricter rules about the proximity of permitted matches. Throughout we will check the sensitivity of any estimates using alternative matching methods.

Matching must reflect the timing of claims. So, when exploiting geographical variation we only want to compare people who make similar claims at the same time in similar areas. Again, this should be feasible given the large number of potential comparators we will have.

We will routinely check how well matching balances the treatment and comparison groups. Recent evaluations of DWP programmes and policies have found PSM performs well in terms of balancing observed characteristics.[6] However, they acknowledge there could still be unobservable differences between the groups. Also, UC is different and a priori it is difficult to know if matching will work as well in balancing observed characteristics. Should any residual differences remain between the treatment and comparison groups after matching we will adjust for these using regression analysis.

Matched Difference-in-Differences

Matching methods only generate reliable (unbiased) estimates if they accurately capture everything that affects both treatment and outcomes. Therefore, to be confident that matching works, we need to be able to observe and accurately measure everything that affects both treatment and outcomes. This is a strong and untestable assumption.

We will relax this assumption by combining matching with difference-in-differences – conditional DiD. This enables us to control for observable differences <u>and</u> unobservable differences that are constant over time between the treated and comparison group. It also controls for changes in outcomes due to common trends. We can use DiD with matched treated and non-treated groups parametrically or non-parametrically. The latter allows for more flexibility but we will try both.

DiD involves comparing the *change* in the mean outcomes of:

a) New claimants in UC areas before and after UC is introduced under Live Service and the North West offices; and

[6] Ainsworth, P. and Marlow, S. (2011) "Early Impacts of the European Social Fund 2007-13" DWP In-House Research No 3
Marlow, S. Hillmore, A. and Ainsworth, P. (2012) "Impacts and Costs and Benefits of the Future Jobs Fund" DWP
Ainsworth, P. Hillmore, A. Marlow, S. and Prince, S. (2012) "Early Impacts of Work Experience" DWP

b) Matched new claimants in similar non-UC areas before and after UC is introduced into the treated areas under a).

Figure (i) illustrates the difference-in-differences approach. The DiD estimator takes the 'normal' difference between the treatment and comparison group as CB and estimates the treatment effect as the distance AC. The key assumption is that the trend in outcomes is the same between the treatment and comparison groups. We can assess the plausibility of this assumption by looking at past trends in outcomes. DiD also assumes that there are no time-varying factors that lead to individuals claiming JSA/UC differently in the treated and non-treated areas – Ashenfelter's dip is the classic example. This goes back to the point about changes in entitlement and eligibility affecting take-up of UC vis-à-vis the equivalent legacy benefit. We can only evaluate the impact on those who would make the same type of claim under both systems.

In most cases we anticipate using cross-sectional differences-in-differences because we will not observe the same outcome for the same people at the required intervals. Cross-sectional DiD involves looking at the change in outcomes between cross-sections of new claimants drawn from the same populations in both treated and comparison areas before and after UC is introduced in the treated areas. To obtain reliable estimates there has to be no differential change in the composition of new claims in terms of unobserved characteristics between the UC and non-UC areas (since we are matching on observables).

The DiD estimate will be a clean comparison of the full UC system versus the full legacy system as long as we measure the outcomes since the claim start long enough before UC gets introduced into the comparison areas (i.e. the areas that will be introducing UC later) to avoid anticipation effects. As we will not be introducing UC into other areas outside the North West and original Live Service offices for longer we will be able to cleanly evaluate the impact of UC for longer.

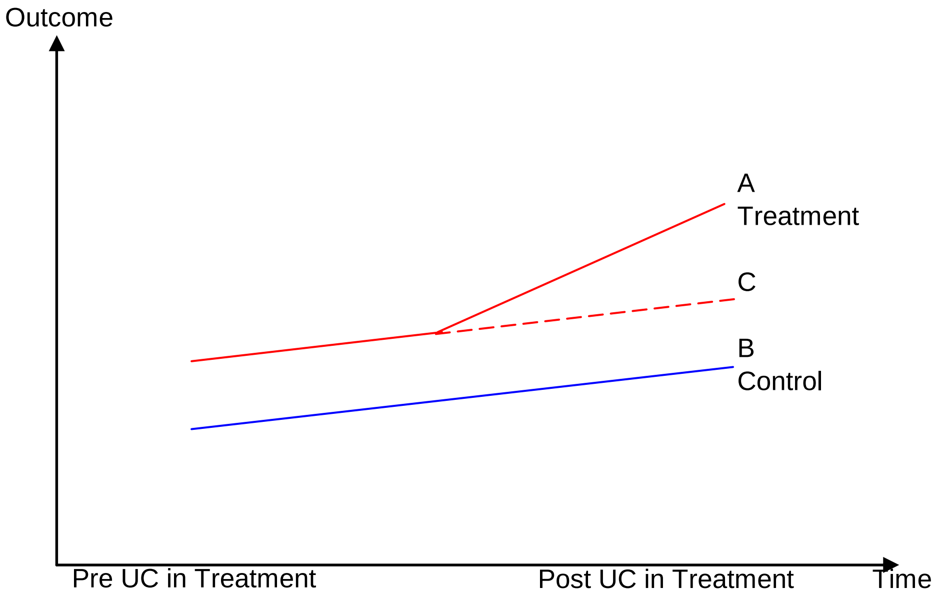
If matching works well (i.e. balances the samples) then conditional on observable factors (including the outcomes of the corresponding cohorts from the before period in treatment and comparison offices) the outcomes of the treatment and comparison group will coincide in the pre-treatment period. The gap between them in figure (i) would not exist and there would be no advantage to be gained from using differences-in-differences. For this to be true there must be common trends (conditional on the observables) and no differential composition in terms of outcome-relevant unobserved characteristics of matched treated and comparison groups pre- and post-UC. Also, repeated cross section DiD requires matching the inflow in the treated area not only to the inflow in the non-treated areas, but also to the inflow in the treated area pre-UC and to the inflow in the non-treated area pre-UC, imposing common support throughout. This is why matched cross-sectional DiD can be difficult to implement. For this reason (and several others discussed elsewhere in this report) the evaluation will focus mainly on ensuring high quality matching to get as well balanced samples as possible on as many key factors as possible.

We will also seek to use triple differences. This allows for differential trends in outcomes between the treated and comparison group as long as these differential trends are constant over time. Triple differences exploits both the geographical and time variation in treatment.

In particular, we would estimate the change in outcomes amongst the treated group using time variation and then do the same for the comparison group and take the difference between the two changes. In contrast to the case when we just use time variation it is not necessary to constrain the time periods to be comparable – they can be different as long as the differences between them would be the same for both the treated and non-treated groups in the absence of UC. IFS's feasibility report summarises it neatly – "Instead of having to select a comparable period prior to the reform or a comparable group of areas, one can implement triple differences against each untreated earlier period or each sub-group of untreated areas and construct a set of comparable estimates in the absence of a reform".

The disadvantages of triple differences include that it reduces efficiency/precision of estimates and it can lead to more bias if constrained to a single comparison as this requires the offices/areas and time periods to be comparable.

Figure (i) Differences-in-Differences



We will explore the value of combining matched difference-in-differences with regression analysis to control for any residual observable differences that remain and to increase efficiency (i.e. parametric matched DiD).

The case for using difference-in-differences depends on there still being some differences in unobserved characteristics between the treatment and comparison groups post-matching, which are constant over time (or in the case of triple differences, where any difference in trends is constant over time). Similarly, regression analysis can reduce any bias due to residual differences in observed characteristics post-matching. Regression analysis can improve the efficiency of DiD estimator if there are observed characteristics which affect the outcomes (whether or not the samples are balanced).

We will estimate impacts using PSM separately for different treatment groups (primarily different household types). For each matched sample we will consider several different outcomes. We will also replicate the same analysis on the same groups and outcomes but for different cohorts of new claims. This will be important because impacts may vary between different cohorts due to:

a) cyclical and seasonal factors;

b) economic trends;

c) the evolution of the policy;

d) anticipation and entry effects; and

e) changes in data sources over time.

For example, we plan to begin by defining the treatment group as all single childless simple new claims to UC in Pathfinder offices and identifying a matched comparison group from all similar new JSA claims in similar offices in similar areas at the same time. We plan to look at the proportion of people in work at certain times post claim. e.g. time in work during first 3 months after the claim started and earnings during the first 3 months weeks after the start of the new claim. We will look at outcomes over different periods.

Over time we will estimate the impact of UC on the same outcomes for the same Pathfinder eligible group but we will have data on more treated individuals (new claims to UC) and we can split the cohorts of new claims in different ways to try and explore a)-e) above. Over time we will be able to estimate impacts over longer periods. As UC gets rolled out to more offices we can also estimate impacts for different groups of offices to see if there are any differences between different groups of offices as well as continuing to explore whether impacts for particular offices are changing over time both in terms of over later inflow cohorts and over longer time horizons. The feasibility of this depends on the size of the offices rolling out UC at different times in terms of the volume of UC on-flows we will be able to look at.

We will assess the quality of the matches we get using PSM before deciding on the merits of combining with differences-in-differences and regression analysis. As a precursor to the PSM estimates we will also explore basic trends in outcomes between different groups and raw differences-in-differences (i.e. just looking at changes in outcomes and not controlling for any other things).

As UC gets rolled out to more new claim types in more offices we will estimate impacts on these new claims as well when volumes build up sufficiently.

Regression Analysis

For the reasons outlined earlier we will use PSM as our primary non-experimental method. However, we will also conduct a series of analysis using simple regression analysis to estimate impacts (i.e. not combined with matching as mentioned above). This strand will help to test the sensitivity of the estimates obtained through matching to alternative assumptions and methods. Also, if it is possible to correctly specify the regression models, it can generate:

a) more efficient estimates than PSM; and

b) estimates we can easily disaggregate according to any of the variables we include (e.g. individual characteristics, types of area/office, etc.).

The regression approach models outcomes and includes a treatment dummy as an explanatory variable. It seeks to ensure comparability by including in the model all the background and area level information that we think might affect outcomes. So, we adjust the expected outcomes of the treated and untreated groups to reflect any differences between them. The coefficient on the treatment variable gives an estimate of the policy impact conditional on (i.e. controlling for) all the other variables in the model that might affect outcomes and which may or may not differ between the treated and comparison group.

We can estimate different regression models to recover estimates under different assumptions and which exploit different sources of variation. Cross-sectional models exploit only the geographical variation in treatment. They ignore the variation in treatment status due to time. The results are likely to be sensitive to the chosen period as this will affect which offices have introduced UC and which have not and how long the different treatment offices have been operating UC (and how long before the comparison offices will introduce UC themselves). Longitudinal models only exploit the time variation in treatment.

We propose to focus on a **difference-in-difference model** as this exploits both sources of variation that are available - geographical and time. It compares the *change in outcomes* before and after UC has been introduced for a particular type of new claim between offices where UC has:

a) been introduced in the 'post-treatment' period for the particular type of new claim; and

b) not been introduced in the 'post-treatment' period for the particular type of new claim.

So, it takes the change in the average outcome pre- and post-UC in the treated offices and takes this away from the change in the average outcome pre- and post-UC in the non-treated offices.

This average estimate takes into account any differences between individuals, offices, areas, households and time periods that can help explain the variation in outcomes over time and between offices.

The treatment variable depends both on *office and time*. The difference-in-difference model has the advantage that it controls for all time invariant differences between the treatment and comparison groups whether they are observable or not.

Our intention is to estimate this model for different cohorts, i.e. different before and after periods and for different groups of individuals and different groups of offices.

Whilst for the reasons outlined earlier our analysis will focus primarily on Propensity Score Matching our regression analysis will focus on estimating DiD models that exploit both the time and geographical variation to identify the treatment effect. Such a model can be written as:

$$Y_{ijt} = \alpha + \beta_1 X_i + \beta_2 X_i Z_j + \tau UC_i + \mu_j + \pi_t + \varphi_{jt} + \varepsilon_i$$

where:

Y_{ijt}: outcome for individual *i* starting a new claim at time *t* in office *j*.

X_i: individual *i* observed characteristics.

Z_j: office *j* observed characteristics. These should be measured prior to the local implementation of UC to ensure they are not affected by the treatment.

UC: treatment indicator for individual *i*: equals 1 if individual *i* starts a new claim in a pilot office after the roll out of UC and 0 otherwise.

μ_j: office *j* fixed effects.

π_t: aggregate time trends (period dummies), common across areas/offices.

φ_{jt}: time-varying unobserved office characteristics. For DiD to consistently estimate the impact of UC on treated offices, these must be uncorrelated with the observables: $E(\varphi_{jt}|X_i, Z_j, UC_i) = E(\varphi_{jt})$

ε_i: individual unobserved characteristics. Again, consistency requires that $E(\varepsilon_i|X_i, Z_j, UC_i) = E(\varepsilon_i)$. Amongst other things this means assuming no entry or anticipation effects on unobserved characteristics.

This is a panel model at the office-level. We have an observation for each office in each period. This implies that, even with many individuals, the precision of the estimates will depend on the number of offices in the treated and comparison groups. This also means that there will be significant limits on the number of regressors that can be included in the model.

We think there will be little scope for estimating individual level panel models to evaluate the impact of UC on new claims. This reflects that such models will only be feasible for those individuals who make a new claim of a particular type before UC replaces those types of claims in their area and who subsequently make another new claim of the same type after UC has replaced that type of claim. So, the panel model will only be available for a very small and unrepresentative sub-set of new claimants who make multiple claims and where those claims occur at either side of the time that UC is introduced for that claim type in some offices.

For reasons outlined earlier we plan to focus mainly on PSM. However, as well as separately estimating impacts using regression models we will also combine regression models with PSM to adjust mean outcomes of the treatment and comparison groups to eliminate any residual observable differences between them (should they exist) and/or to increase efficiency.

We will also use differences-in-differences both to test how successful PSM matching appears to be and potentially to control for any pre-existing differences that are constant over time that matching may fail to eliminate. However, we are hopeful, given other applications of PSM to evaluate labour market policies and programmes that we will have sufficiently rich information at the individual and office level to identify good matches.

As part of the regression analysis we will consider different dimensions of heterogeneity in treatment effects. For a *particular claimant type* we will look:

Across areas - differences in the composition of new claimants of the same type and/or differences in their local market conditions

Over calendar time - differences in the composition of new claimants of the same type and/or differences in aggregate economic conditions

By time since UC locally introduced – which may lead to differences in the composition of new claimants of the same type and/or differences in the treatment provided to new claimants as offices adjust to the new regime and gain experience in running it.

We will also see how the impact varies between different sub-groups. We will explore the scope of estimating regression models that interact treatment status with demographic, area and time information to explore the extent to which the impact varies with these things. Since the number of offices is key to the precision of the estimates it is important to be selective in choosing which interactions and explanatory variables to include.

v) Data Sources

Any evaluation is only as good as the data available to measure actual outcomes and to estimate counterfactual outcomes. We have already noted that matching methods are more likely to provide estimates of the true impact if we have rich data on all the factors that affect treatment status and outcomes (although we also noted that it is difficult to know in practice whether we have all the data we need).

It is vital that the data is collected and recorded in the same way for both the treated and comparison groups. This goes back to ensuring that the new claimants in the treated and non-treated groups are as similar as possible so that the only difference between them is that one group is making new claims to UC and the other is making new claims to the equivalent legacy benefit. So we want the same data on matched individuals in matched areas making equivalent legacy benefit claims at the same time as the treated UC group to estimate the counterfactual outcome of the treated group.

The evaluation will rely on a combination of administrative data sources. The main data sources we anticipate using include;

- **Real Time Information** – detailed and comprehensive timely data on earnings but only building up from April 2013.
- **Work and Pensions Longitudinal Study** – information on employment (from P45/P46 forms) and some earnings information (total annual earnings) from P14

forms. 5-6 month lag. The WPLS also encompasses the National Benefits Database and data from the Single Housing Benefit Extract (SHBE).

- **National Benefits Database** – historical comprehensive data on past benefit claim history. 2-3 month lag.

- **NOMIS** –local area labour market information.

- **Local Deprivation Indices** – local labour market and socio-economic indicators.

- **BOXI** - Office level data on past performance, size, claimant composition, etc. from MI systems.

- **Atomic Benefits Database (ADS)** – very timely data on new JSA and ESA claims collected by JCP management information systems.

- **Manual Capture Tool and Evidence Manager –** data on UC new awards including the type of award, start and end dates and some demographic information on claimants.

- **Master Index** – information about people's participation in other programmes that might affect their outcomes.

These data sources have several advantages. They cover the whole population of interest – new claims – and so provide the biggest possible samples. This is important as it affects:

- Representativeness of the treated group and the external validity of estimates;

- Precision of estimates

- Scope for obtaining impacts for sub-groups; and

- Our ability to find good matches amongst the comparison group.

These data also provide:

- comprehensive data on pre-participation benefit claim histories,

- some albeit less complete information on past employment and income histories;

- fixed demographic information (gender, age, etc.);

- time-coded treatment information; and

- time coded data on post-treatment outcomes. This includes more complete information on earnings and employment from the Real Time Information System than we have about their pre-treatment employment and earnings from pre-existing data sources (from the Atomic Data Store and the Work and Pensions Longitudinal study).

Ideally we would carry out the matching using the same data sources that we will use to measure outcomes. Sometimes we will do this, e.g. when looking at impacts on employment as recorded by the Work and Pensions Longitudinal Study. However, in other cases we will want to use the more complete and accurate information on earnings from RTI to measure outcomes than we will have to match individuals on pre-treatment. RTI excludes self-employment and has missing start and end dates for lots of employment spells. However, it is timely (though the data does get updated retrospectively) and comprehensive and it is more likely than WPLS to capture earnings below the national insurance contributions threshold. We may be able to explore how well matches based on pre-existing data sources compare with those from the RTI but only for a relatively limited recent period when the different data sources co-exist.

The RTI data is important.[7] This reflects that it will be the most reliable source of information we have on the outcomes we are interested in – employment and earnings. But it also reflects that it is one of the main sources that we expect to be consistent between the treatment and non-treated groups. Its main disadvantage is that it lacks much historical data for matching and so in many cases we expect to use other sources (National Benefits Database and HMRC's employment data) to match individuals and couples on their past employment and benefit claim histories. The lack of historical data might also prevent DiD analysis where the 'before' cohort predates RTI.

The RTI data is new. Consequently it will be important to carry out exploratory analysis and quality assurance of the data to assess its strengths and limitations. We already know that it excludes self-employment. This could be important. If UC claimants are more (or less) likely to move into self-employment than legacy claimants we will not be able to pick this up using RTI (or WPLS).

Early indications suggest that employment start and end dates are missing for lots of employment spells. This could be an issue although we have no reason to expect this problem to differ between the UC and legacy claimants. We are exploring how we can derive employment outcome variables using the fields available. Vital throughout is treating both the JSA/ISlp and UC samples the same.

We also hope to look at the impact of UC on the time new UC claimants spend claiming benefits (UC) whilst out of work.

The data sources we will use tend to be individual level. However, we also want to evaluate impacts on households. Data on UC new claims should enable us to identify which claimants are in couples and which have children and which do not. However, we also need to identify which individuals making new claims to legacy benefits are in couples with and without children by merging individual level data. There are risks here. This is not something we typically do. It is uncertain how accurately and comprehensively we will be able to identify

[7] We use the term "RTI" to cover both the RTE data we get for UC claimants which is a sub-set of RTI data and the RTI data we get for legacy benefit claimants.

which individuals are single without children, in couples without children, lone parents and couples with children.

The data we will use to estimate impacts using conditional DiDs will vary according to the different time periods we consider and the different outcomes we look at. Again, it would be best to use the same data sources to measure the changes in outcomes before and after the introduction of UC in some areas. Especially during the early phases of the roll-out this will mean relying on the Atomic Data Store, National Benefits Database and the Work and Pensions Longitudinal Survey. However, there may be issues over the consistency of these data between people making new claims to UC and those making new claims to legacy benefits.

We will conduct preliminary analysis using a sub-set of these data. This reflects that there is significant work required to quality assure, transform and combine the data from different sources. Initially we plan to merge RTI/RTE, ADS and MCT/EM data. These data are all timely, which will allow us to analyse larger volumes earlier. The RTI/RTE contain the consistent information on outcomes whereas ADS and MCT/EM include **demographics and information on the claim.** The ADS will also provide some benefit history for both the treated and non-treated groups we can use for matching and we can aggregate ADS data to provide historical office-level information for matching.

Over time we will merge more data from the above sources primarily to improve the quality of the matching. We will also explore the scope for obtaining data from other sources including data on criminal records.

We have focused on the data we will use to estimate the impact of UC replacing new claims to JSA and ISlp under Live Service and the North West expansion. This focus reflects that our evaluation strategy will primarily focus on estimating this impact for reasons outlined earlier. To estimate the impact on aggregate and indirect effects will require different data sources since these impacts are not confined to new claimants. However, we will still be able to use some of the same datasets when considering the impact on sub-populations of past claimants. The other main data set we will use to evaluate aggregate and indirect impacts is the Labour Force Survey (LFS). This is a large representative survey containing information on labour market outcomes and background characteristics. The roll-out across the North West could provide sufficient volumes for aggregate and indirect effects to be detectable.

vi) Other Issues

Short and Long Run Impacts

We can distinguish between short-run and long-run impacts in two senses. First, the impact on individuals' labour market outcomes might change over time. For example, UC might increase the days spent in employment during the first 6 months after an individual's claim but the impact on days spent in work 6-12 months after the claim start might be bigger or smaller or insignificant.

We will evaluate impacts over different time horizons. What is possible will vary according to the claim type and roll-out schedule. The latest roll-out schedule does allow us to look at the clean impact on longer-term impacts than the hypothetical roll-out schedule IFS considered as there will be a much longer time when some people are subject to UC whilst other similar people in similar areas are still under the legacy system.

Secondly, the impact of the policy might evolve over time for several reasons. Indeed, the greater emphasis on test and learn means that we would expect the delivery and effectiveness of UC to improve with time. Also, the initial UC policy under Pathfinder is not what we expect UC to look like longer-term. In particular:

- there will be much more automation in steady-state;
- in the longer-term UC will be available to all new claims and all existing claimants will be on UC;
- transitional protection will end;
- people (claimants, potential claimants and staff) will become more familiar with UC over time, which may mean that people's experience of UC changes over time and/or their responses to it changes;
- the phased roll-out, as we have already noted, may affect the composition of new claims so that they are different to what they will be in the longer-term.

We might expect that the impact of UC will vary over time because of all these factors and also because the effect of UC may vary with local and macro-economic conditions. Again, our intention is to estimate impacts separately for different time periods, covering different groups of UC offices and different claim types. This should help to build-up a picture of how and why the labour market impacts of UC might change over time.

As well as disaggregating estimates by time, claim type and different groups of treated and non-treated offices we will also estimate how the impact varies between different sub-groups of the population, e.g. by age, sex, ethnicity, employment/benefit history, etc. Clearly, any estimates will only be as reliable as the data available. Many demographic variables are recorded in the administrative data we will use. However, there may still be issues around its quality and completeness in some cases and sample sizes can become problematic once we start attempting to break down impacts on several dimensions. For things that might be affected by UC (e.g. health status) we would want to estimate impacts based on health status prior to exposure to UC and be mindful of possible anticipation effects.

Under the current roll-out plans we do not anticipate being able to estimate what impact different aspects of the UC policy have on labour market outcomes. This reflects that all aspects of UC are introduced at the same time. There may be some scope during the test and learn phase to test different variants of UC to help evaluate their relative effectiveness and cost-effectiveness. We will also continue to explore the scope for introducing different variants of UC in different areas during the national roll-out alongside plans to explore the scope for randomisation by office (see below).

Timescales

The quantitative evaluation of the impact UC has on the labour market under Live Service and during the North West expansion will be iterative and on-going. We will be able to derive more reliable estimates and use more methods as the volume of UC claimants builds up and after we allow time to capture their labour market outcomes. The roll-out schedule means that we will be able to get initial impacts on a sub-set of single claimants without children making new claims to UC first.

Since UC is only being introduced in a relatively small area for a while the number of non-treated new claims from which we can draw matches will build up quickly. We will typically have to wait for the volume of new UC claims to build up. The volumes we need to derive statistically reliable estimates will vary depending on the outcome we look at and the sub-group and context (e.g. timing of the new claims) as well as the size of any impacts.

We will be able to start estimating a range of impacts as soon as the volume of new UC claims of a particular group is big enough and enough time has elapsed to obtain reliable outcome data. The scope for obtaining more reliable estimates that will have greater external validity will increase as the volume of UC claims builds up. This reflects that the range of UC claimants will increase allowing us to capture impacts on a more representative set of claims.

As already mentioned we envisage that we will need volumes to build up more before we can start to estimate aggregate and indirect impacts. When looking at the indirect impacts on people who are already subject to UC when a further expansion occurs the timescales are clearly greater. This reflects that we need time under UC for this treated group and then time to allow the volumes affected by subsequent changes to build-up sufficiently. It may be that this will be feasible when the national roll-out starts and more claim types are replaced in the North West offices. It depends on how quickly large volumes of new claims of different types get replaced by UC – if this happens quickly as we hope it will then we might be able to pick up indirect effects on those who were already subject to UC under the expansion of Live Service.

vii) Longer-term Evaluation Plans

We are considering the scope for some randomisation in the national roll-out and as part of the test and learn approach because of the value it could add to the evaluation. In particular, we are exploring the scope and value of selecting 2-3 offices in each District at random to remain outside UC for as long as possible because:

i) This would generate robust impact estimates that would be subject to the least possible amount of doubt and criticism; and

ii) It would help validate the likely reliability of other impact estimates we obtain from non-experimental methods. To do this the design would need to be set up to recover the impact of the full UC system versus the full legacy system for new claims to JSA/ISlp.

A partial randomisation by office, whereby we randomly choose 2-3 offices per District outside of UC for as long as possible, still faces methodological challenges. For example, whilst it could still produce unbiased estimates for the same reasons that randomisation by individual does, it tends to be much less efficient than individual randomisation. This means the estimates tend to be further away from the true impact more of the time. What is most important here is

a) how many offices we randomly assign out of UC (there will be lots who are included in the UC treated sample); and

b) how much of the variation in outcomes is at the office rather than the individual level.

We can get an idea of both of these using historical data. This means we can work out how much uncertainty there might be over the impact estimates and how this would vary depending on the design (e.g. how many offices we randomly assign out of UC).

Randomisation can, by chance, mean that the randomly assigned 'control' offices look different to the treated UC offices in some Districts on the basis of salient observable characteristics. For example, offices in the control sample might be smaller on average or contain a bigger proportion of new claims with better employment records, etc. We may use pre-randomisation blocking or use regression analysis to adjust results post-randomisation to enhance face-validity.

We are also exploring the scope for randomisation by individual during the testing of the end-state digital UC service with 100 claimants in 2014, moving to 1,000 and 10,000 claimant tests in 2015. In particular we are looking at the feasibility and value of randomisation as volumes build-up in 2015 and beyond.

viii) Conclusions

We will estimate the impact that UC has on the labour market outcomes of many new claimants during the roll-out of UC under Live Service and the North West expansion. We will estimate impacts using a combination of administrative data sources (primarily the RTI and WPLS/NBD) and by combining matching, difference-in-difference and regression non-experimental evaluation methods. We will rely primarily on the geographical variation in exposure to UC due to the phased roll-out but also conduct sensitivity analysis by exploiting the time variation in treatment.

The evaluation will estimate the impact of UC on new claimants to UC compared with new claimants to the legacy system on a range of outcomes including time in employment; earnings and the likelihood of being in work at different points in time after the claim start.

The evaluation will focus mainly on recovering estimates of the average impact on the treated.

The expansion of live service means that there will be more time when UC is in place in one area whilst the legacy system remains in place, in its entirety, elsewhere. The fact the legacy system will co-exist in most of the country increases the chances that we will be able to find reliable matches for the vast majority of people who make new claims to UC during the North West expansion. The longer timescales also allows us to explore longer-term outcomes. We expect anticipation effects to be less problematic due to the length of time that most offices will remain outside UC.

The main limitation is that the evaluation of the North West expansion will only be able to evaluate the impact on those new claim types that UC replaces (JSA and ISlp) during these early roll-out phases. This also limits the scope for looking at aggregate and indirect impacts as the volumes will be lower than would have been the case had UC replaced all new claim types in the North West during the expansion. The nature of the North West expansion also limits the scope for evaluating the impact on other new claim types in that region during the later roll-out because more people will have entered UC via another route already.

The nature of the national roll-out is to be confirmed. There may be scope for keeping some areas outside UC for longer. This might provide opportunities to evaluate the impact of UC on other new claim types and it could enhance the opportunities for looking at indirect and aggregate impacts if the roll-out in treated areas occurs relatively quickly.

Annex A

Review of "Evaluating the Impact of Universal Credit on the Labour Market in Pathfinder and the North West Expansion of Live Service"

Stuart Adam, Monica Costa Dias and Barbara Sianesi, Institute for Fiscal Studies

11 June 2014

1. Introduction

This report outlines how the Department for Work and Pensions intends to evaluate the impact of Pathfinder and of the North West expansion on a number of labour market outcomes experienced by new UC claimants.

The 'treatment' being examined is a gradual replacement of new claims to out-of-work legacy benefits (JSA and IS for lone parents) with entry to the full UC regime – gradual in the sense that the policy is being rolled out over time to more offices and more claimant types. The report discusses in detail the evaluation opportunities offered by this gradual roll-out, with the main focus being on the labour market outcomes of people who successfully claim UC.

For the most part the report is excellent: carefully written and very thorough. It has incorporated the major points and caveats raised in our feasibility study. In particular it clearly:

- defines the evaluation questions it aims to answer, while being aware of their limitations;
- highlights the different sources of variation that can be used to isolate the impact of UC, makes a case for the preferred methodology to be employed and discusses potential threats and robustness analyses; and
- recognises the need to ensure that the definition and measurement of anything of relevance (start and end dates of new awards, outcomes etc) are carried out in a comparable and consistent way across UC and legacy benefit claimants.

We have incorporated most of our detailed and specific comments directly in the report as MS Word comments or tracked changes. Those are numerous, but many – not all – are minor. This document sets out our main comments that apply across the paper, starting with issues of substance. Our most serious concern is about the section on regression analysis, which we discuss fully in a separate section. We then turn to presentational issues, and conclude with a brief summary of our views.

2. Strengths and limitations of the proposed approach

In terms of understanding the labour market impacts of UC, the biggest limitation of this evaluation is not a weakness of the proposed evaluation strategy but its inherently narrow scope. Since it looks only at replacing new claims to JSA and ISlp, not other benefits (where the change to UC might be more important), and it is restricted to the impacts in Pathfinder and Live Service areas in the North-West, it can only ever provide a very partial picture of the impacts of UC as a whole. This is of course implicit in the very title of the report but it is nevertheless worth highlighting.

Likewise, the report rightly chooses to focus primarily on the impact on new claimants, since aggregate and indirect impacts will be harder to detect. But that does mean that the evaluation will probably fail to capture one (perhaps small) part of the overall impact of UC on employment and earnings during Pathfinder and Live Service: the impact on non-claimants via entry effects and indirect effects. That is not a weakness of the proposed evaluation strategy given the available options. However, on this subject it is worth noting that the report says that, despite the difficulties, "we will explore the scope for evaluating aggregate and indirect effects" (p.6); yet in the data requirements section of the report there is no discussion of the data requirements to recover aggregate and indirect effects. Currently, the data section (unlike the rest of the report) addresses only the data that might be used to look at the impact of claimants. If aggregate and indirect effects are to be explored at all then the data requirements section needs to consider what available data could be used for that purpose.

Within the limited scope of looking at the impact on (only) claimants of the replacement of (only) new JSA and ISlp claims with UC in (only) the Pathfinder and Live Service areas, the proposed evaluation strategy is a very good one. The baseline choices of evaluation question, sources of variation, methodologies and data, and the alternatives to be explored as variants, all seem entirely appropriate.

The early phases of the roll-out of UC were singled out in our feasibility study as offering the cleanest design for evaluating the full UC regime *versus* the full legacy benefit regime on new claimants, albeit only when either regime is entered via a JSA or equivalent UC claim. Indeed, relative to the roll-out plan examined in the feasibility study, the current roll-out plan is more conducive to evaluation given that:

- the period before UC replaces new claims to other benefits will be longer, and
- the rest of the country will now remain entirely outside UC for quite some time.

As a source of variation, the plan to rely primarily on geographical variation in exposure to UC due to the phased roll-out, i.e. compare similar offices at the same time, but to also carry out sensitivity analysis by exploiting the time variation in treatment, looks like a sound choice for the reasons explained in the report.

Similarly, the choice of matching as the principal evaluation method seems sensible and is based on a thorough discussion and assessment of the reliability of matching estimates based on the available information (p.11-15), while the proposed variants to explore are also sensible ones.

In the following two subsections we discuss potential threats to the validity of the estimates and suggestions for where to focus attention in terms of methodology.

Potential threats to the validity of the estimates

One potential threat to the validity of the estimates arises from effects of the policy on the composition of the treated group and on the composition and outcomes of the comparison group: entry and anticipation effects. Anticipation effects, the report rightly notes, can be safely ignored if comparison areas are chosen from outside the NW or from NW offices well before UC is introduced in those offices. Entry effects are more complicated, and the report would benefit from distinguishing more clearly between three different types that may arise:

- People choosing to accelerate or delay a claim around the time a new roll-out phase is introduced, in order to affect which regime their claim falls under. These are analogous to anticipation effects (but for the treated group rather than the comparison group) and are likely to be unproblematic for similar reasons: the sample can simply be chosen from a time well away from the introduction of UC.
- Mechanical entry effects – arising not from individuals' behavioural responses to UC but simply from the fact that new claimants must be people who are not already within the UC regime – are expected to be minor at this stage as UC is only replacing JSA/ISlp, and will be even less of a problem if offices in the NW expansion introduce UC at the same time for all household types.
- The people eligible for and (more importantly) choosing to claim UC might be different from the people choosing to claim legacy benefits, for a variety of reasons including different entitlement levels, conditionality, awareness, attitudes, having access to a different in-work support regime if moving into work later, etc. While this possibility is mentioned under 'limitations of focusing on new claimants' at the top of p.5, its likely extent in practice is not discussed and it is not mentioned at all under 'entry and anticipation effects' in section (vi). This type of entry effect might also be relatively unimportant in this phase of the UC roll-out – not least because of initial unfamiliarity with the details (or even existence) of UC – but in this case it is less obvious, and , especially given that one aim for UC is to increase take-up, it ought to be assessed.
- For example, are such entry effects likely to be less of an issue in the Pathfinders than in the NW expansion if information takes time to spread? Might it be more of an issue for families with children than for single people without children, since it is more likely to change the entitlements of those with children (see Section 3.4.2 of the feasibility study).

There is also a second threat to the validity of the estimates which is not mentioned at all in the report: the challenge of isolating the effects of UC from the effects of other changes happening around the same time, particularly changes that might have differential effects across areas. As discussed in Section 5.2 of our feasibility study, careful knowledge of other reforms in the treated areas and in the potential comparison areas is required. It might be necessary to match (or condition) on additional observables related to these reforms to ensure comparability of areas. This is a potentially serious issue and ought to be discussed in the report.

Further methodological suggestions

The main focus of attention should be on the quality of matching. Entry effects aside, both treated and comparison groups are starting equivalent out-of-work benefit claims, so it would probably pay to focus on carefully adjusting for differences especially at the area/office level. In particular, matching exactly on local outcome history of past benefit inflow cohorts would go a long way towards summarising past performance of the office and local labour market conditions.

In contrast, if resources are constrained, it seems to us that the two lowest priority variants to explore are:

- Cross-sectional matched DiD. This is difficult to implement, and given the discussion of matching vs DiD below the extra work might not be warranted.
- Combining propensity score matching with multi-level models. We are not very familiar with the literature; however, straightforward clustering of the standard errors is typically used in PSM models to reflect the hierarchical structure of the data. While interesting, exploration of this methodological area might be seen as lower priority.

Regression analysis: Our advice to increase precision is to run more parsimonious regression models that rely on clean variation allowing the identification of the impacts. This is discussed further in Section 3 below. Regression models might be particularly useful for subgroup analysis where sample sizes are a concern; in this case, matching might be used as the robustness check to provide reassurance about the extent of overall comparability and hopefully that the conclusions remain the same.

Matching vs DiD: The report is right that (repeated cross-sectional) DiD will only be useful in cases where matching does not succeed in balancing the outcome histories well. If matching includes <u>equivalent</u> <u>outcomes of past inflow cohorts in that area</u> and manages to balance these well, then the DiD correction term to subtract would be zero. If matching does not balance these particular (i.e. the underlined) past outcomes well, one might indeed want to subtract the bias (i.e. do DiD). Note, however, that for this to be accurate one would need to have convinced oneself that there are common trends (conditional on the observables) and that there is no differential composition in terms of outcome-relevant unobserved characteristics of matched treated and comparison groups pre- and post-UC. Notice also that two extra matching steps are required, as repeated cross section DiD requires matching the inflow in the treated area not only to the inflow in the non-treated areas, but also to the inflow in the treated area pre-UC and to the inflow in the non-treated area pre-UC, imposing common support throughout.

To reinforce the earlier message: given that past performance of the area/office in terms of outcomes is critically important, when doing the (cross-sectional) matching one might wish to give these variables extra weight, e.g. by matching on them exactly in a first step. As per above, with such balanced (in this case exact) matching on the relevant outcome variable the DiD correction term to subtract would be zero anyway.

Combining matching/DiD with regression: Regression adjustment on matched samples (i.e. combining matching with regression) improves efficiency if the X's affect the outcomes. It will also parametrically correct for any residual imbalance in observed characteristics that are thought might affect the outcome. A couple of considerations:

- Rubin (2007) suggested specific criteria for regression adjustment to be reliable: the absolute standardized differences of means of the propensity score should be less than 0.25 and the ratio of the variances of the propensity score in the treated and control groups should be between 0.5 and 2.[8] Outside these ranges the matched samples are considered still too far apart to rely on linear regression adjustment. An open issue is how one would then proceed to calculate standard errors that take into account both the PSM (on an estimated propensity score) and the regression adjustment. Our intuition is that if kernel matching is chosen, then bootstrapping the entire estimation sequences would be appropriate.
- The relatively new literature on doubly robust estimators has a different take on the issue. The idea is to combine weighting estimators with regression adjustment to construct estimators which are consistent when either the outcome model or the treatment model is correctly specified (such estimators include the augmented inverse-probability-weighted estimator or the inverse-probability-weighted regression-adjustment (aka Wooldridge's double-robust) estimator). This is not PSM but inverse-probability weighting based on the estimated propensity score, and appropriate standard errors have been derived.

Note that in combining matched DiD with regression (what the report refers to as 'parametric matched DiD'), the regression adjustment improves efficiency if the X's affect the *trend* in outcomes. It will also parametrically correct for any residual imbalance in observed characteristics that are thought might affect the trend.

3. Regression analysis

We struggled for a long time to understand the model specifications because the notation is not always clear and there isn't always an explanation of why certain terms have been included. Lack of clarity notwithstanding, there are numerous elements that are simply wrong, most importantly terms that cannot be separately identified. We mention some of the problems here, but not all, and then discuss what an appropriately specified regression model would look like.

It is crucially important whether μ_j are unobserved error terms (random effects), assumed to be uncorrelated with treatment status and other observed characteristics, or office dummies (fixed effects) which may be correlated with other regressors. Equally important is to consider whether π_t are time dummies or random shocks accounting for some variation around a parametric (possibly linear) time trend.

In the cross-sectional model, we think μ_j must be a random error term (otherwise the office fixed effects, the effects of office characteristics and the treatment effect are not separately identified as, for example, the effect of being a treated office cannot be disentangled from the variation in μ_j). If μ_j is a random error term, the cross-sectional model looks OK as

[8] Rubin, D.B. (2007), "The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials", Statistics in Medicine, 26(1):20-36.

written – the description of μ_j as an “average unobserved office/area effect” just needs to be clear that this is a random error term, uncorrelated with the observed regressors.

When it comes to the longitudinal and DiD regression models, however, the text discusses μ_j and π_t as dummies and talks about interacting them. So μ_j has switched from being a random effect to a fixed effect without warning.

In the longitudinal model, if π_t is a dummy for each period then the treatment effect is only identified from variation (if any) between offices in the calendar date at which UC is introduced (since if UC is introduced at the same time everywhere then its effect is indistinguishable from the change in π_t at that point). Since identification is coming entirely from some offices being treated at a certain point while others are not, it is not a purely longitudinal model at all. Incidentally, there are also typos here: the treatment effect has changed from τ to β_2 and there are two β_4s.

In the DiD model, the effect of *PostUC* cannot be separately identified from π_t, the effect of being in a treated office *UCoff* cannot be separately identified from μ_j, and β_4 is subsumed within β_5, so in each of these three cases the first term should be dropped.

In both longitudinal and DiD models:

- there is no rationale for including a time trend on top of the period effects.
- office characteristics (*office*$_{jt}$) should not have a time subscript: the regressors should include only characteristics which are historical or time-invariant to avoid the risk of their being affected by the treatment.
- the interaction term between individual characteristics, office characteristics and period dummies can in principle be included but it may not be sensible to do so as (particularly if there are more than a couple of individual and office characteristics) it adds a large number of regressors when, as discussed below, the fact that treatment status varies only by office and time puts a premium on parsimony in the specification.
- the interaction term between office dummies and period dummies ($\mu_j\pi_t$) should not be included since the treatment effect cannot be separately identified (the effect of treatment cannot be disentangled from the effect of particularly high or low office-period interaction effects).
- what must be included, however, is an error term reflecting time-varying unobserved office characteristics (assumed uncorrelated with treatment status and other regressors), which could be called φ_{jt}.

Note that if μ_j and π_t were unobserved error terms (random effects) in the longitudinal and DiD models, a different but still important set of problems would arise!

When using *all time and geographical variation* to identify the treatment effect, a parametric regression model which is fully identified can be written as follows:

$$Y_{ijt} = \alpha + \beta_1 X_i + \beta_2 X_i Z_j + \tau UC_i + \mu_j + \pi_t + \varphi_{jt} + \varepsilon_i$$

where:

Y_{ijt}: outcome for individual *i* starting a new claim at time *t* in office *j*.

X_i: individual *i* observed characteristics.

Z_j: office *j* observed characteristics. Note again that these should be measured prior to the local implementation of UC to ensure they are not affected by the treatment.

UC: treatment indicator for individual *i*: equals 1 if individual *i* starts a new claim in a pilot office after the roll out of UC and 0 otherwise.

μ_j: office *j* fixed effects.

π_t: aggregate time trends (period dummies), common across areas/offices.

φ_{jt}: time-varying unobserved office characteristics. For DiD to consistently estimate the impact of UC on treated offices, these must be uncorrelated with the observables: $E(\varphi_{jt}|X_i, Z_j, UC_i) = E(\varphi_{jt})$

ε_i: individual unobserved characteristics. Again, consistency requires that $E(\varepsilon_i|X_i, Z_j, UC_i) = E(\varepsilon_i)$. Note that this means assuming, amongst other things, no entry or anticipation effects on unobserved characteristics.

In this model, the identification of τ relies on office and time variation in treatment status: there is no variation in individual's treatment status within office at a point in time. In the presence of office fixed effects (μ_j), in effect we have one observation per office (not per individual) in each period (since identification relies on changes over time in mean outcomes at the office level, as explained below). This implies that, even with many individuals, the precision of the estimates will depend crucially on the number of offices in the treated and comparison groups. Moreover, the number of offices also imposes important restrictions on the number of regressors that can be included in the model.

To make these two points clear, notice that τ is identified from the difference (between treated and untreated offices) in the differences (between pre- and post-treatment periods) of mean office outcomes. Formally:

$$\Delta\bar{Y}_j = \beta_1\Delta\bar{X}_j + \beta_2 Z_j\Delta\bar{X}_j + \tau UC_j + \Delta\pi_t + \Delta\varphi_j + \Delta\bar{\varepsilon}_j$$

where Δ represents the before-after change and the bar stands for the average within office and time period.

This regression model is estimated at the office level, so the number of coefficients that can be estimated is limited by the number of observations (i.e. the number of offices in each time period). Furthermore, β_1 and β_2 can only be identified when the office-level average characteristics of new claimants (X_j) changes over time.

The report suggests running separate regressions for different subgroups in order to identify how the effect of UC varies between them. But rather than running separate regressions, one can run a single regression with the treatment dummy interacted with some of the observed characteristics. We could write the model as:

$$Y_{ijt} = \alpha + \beta_1 X_i + \beta_2 X_i Z_j + \tau UC_i + \tau_1 UC_i X_i + \tau_2 UC_i X_i Z_j + \tau_3 UC_i Z_j + \tau_3 UC_i t + \tau_2 UC_i T_j + \mu_j + \pi_t + \varphi_{jt} + \varepsilon_t$$

where

- T_j is time since UC is first introduced in office *j* and can be interacted with the treatment dummy if there is variation in the time offices are brought within UC;
- *t* is calendar time and can be interacted with the treatment dummy if the data span more than one post-treatment period.

Again, we advise parsimony in selecting the interaction terms as identification is limited by the number of treated offices. Running separate regressions would be equivalent to interacting the treatment dummy with all the other regressors: very flexible, but correspondingly losing precision.

The model discussed above uses both geographical and time variation to identify treatment effects. Pared-down specifications can be used to exploit only cross-sectional, or only longitudinal, variation.

4. Presentational comments

We understand that the report provided to us was in draft form and would presumably be 'tidied up' before publication in any case. But in the process of revision, you should bear the following in mind:

- The report would benefit from an **Executive Summary** – a page or two, say – setting out the basic plan. The report does say that its scope is just the replacement of new out-of-work benefit claims in certain parts of the country; and that the intention is to focus primarily on the impact on new claimants, using primarily geographical variation and propensity score matching, and initially using a combination of RTI/RTE, ADS and MCT/EM data (but will also look at aggregate and indirect effects, explore time variation, experiment with other methodologies such as matched DiD, triple differences, regression on its own or combined with matching, etc, and bring in other datasets over time). But each of these details is buried in a separate place in the middle of (good) discussions of alternative options – it would help to bring it together into a concise description of the proposed evaluation.
- The **structure** could be improved in some respects. Broad suggestions to this end have been included at various points in the text. An overall comment is that there seems to be repetition of the material at times, especially in the methodological part, which could gain from a tightening and reorganising (possibly adding further subheadings and turning some of the discussion into bullet/numbered points).
- Use of **language and terminology**. There are some terms that are potentially ambiguous or used inconsistently. In particular, 'family' and 'household': in some places (e.g. bottom of p.7) 'family' is used to mean 'benefit unit' as distinct from 'household' (and that is how we tend to use the terms at IFS); but in most places 'household' seems to be used to mean 'benefit unit', 'family' is used specifically to mean 'benefit unit with dependent children', while 'couple' and 'single person' apparently refer only to those without children. For example, the paper refers repeatedly to 'couples and families': but couples with children *are* families; for many readers couples without children are families too; and if you use 'family' to mean 'benefit unit' then single people without children are also families! 'Couples without children and families with children' would

therefore be clearer. It would help to have definitions, consistent usage and avoidance of ambiguity.

There are also many cases where technical terms are used but only explained later or not at all. These range widely, from 'ISLP' meaning income support for lone parents, to referring to matching and difference-in-differences before having explained what they are, to the description of different datasets appearing only after the discussion of how each might be used. We have highlighted some examples in the text, but far from all. You need to decide how comprehensible you want the report to be to readers with different levels of prior knowledge (about institutional details or econometrics), but in any case you should think about whether the material is always presented in the right order, and the first time you use terms or ideas that you don't want to assume your readership already understands, you should explain them, refer forward to an explanation later in the report, or refer to an explanation in a different publication. This leads on to the next two points...

- There are no **references** in the report. Bibliographic references need to be included when referring both to 'other/previous work' and to technical concepts and methodologies that aren't fully explained in the text.
- The report discusses a lot of complex and technical ideas. Given that the paper will be available to a wider readership (albeit perhaps not intended for the proverbial man on the Clapham omnibus), our advice is to ask someone not intimately familiar with the issues (e.g. who hasn't read our feasibility study) to read the paper through to check for **clarity and comprehensibility**. We are obviously not best placed to judge how clear the whole paper will be to someone less familiar with this topic.

5. Conclusion

In most respects the report is excellent, the proposed evaluation strategy is wholly appropriate (subject to its inevitably limited scope).

In terms of substance, the main areas where the report could be improved are:

- reconsider the proposed regression specifications;
- assess the potential risk of confounding the effects of UC with the effects of other changes happening around the same time, and strategies for dealing with it;
- assess the likely risk of entry effects arising from UC affecting take-up decisions, for different stages of the initial roll-out and for different subgroups;
- consider the data requirements for any exploration of aggregate and indirect effects.

The way in which the material in the report is presented could also be improved.

In implementing the evaluation strategy, particular attention should be paid to doing the matching as well as possible – especially matching offices/areas as exactly as possible on the outcomes of previous cohorts of claimants – and checking the match quality achieved.