Community Investment and the Bottom Line

Investigating associations between community investment and housing providers’ costs using advanced data science techniques

Jim Vine with Christina Knudsen and Phil Goddard

November 2016

www.hact.org.uk
Community Investment and the Bottom Line:
Investigating associations between community investment and housing providers’ costs using advanced data science techniques

Contributions:
Report by Jim Vine with Christina Knudsen and Phil Goddard
Analysis designed by Jim Vine with Phil Goddard and Christina Knudsen
Analysis implemented by Phil Goddard and Christina Knudsen
Costs data acquired and processed by QA Housing Consulting

Published November 2016
© HACT 2016

www.hact.org.uk
@HACThousing
info@hact.org.uk

HACT
49-51 East Road
London
N1 6AH

HACT is registered as the Housing Associations’ Charitable Trust, charity number 1096829, company number 04560091.

This document may be cited as:

“Vine, Jim; Knudsen, Christina; and Goddard, Phil (2016). Community Investment and the Bottom Line: Investigating associations between community investment and housing providers’ costs using advanced data science techniques. HACT. London, UK.”
## Contents

Contents........................................................................................................................................... 4  
Table of figures.................................................................................................................................. 6  
Foreword............................................................................................................................................ 7  
Executive summary ........................................................................................................................... 9  
Introduction ...................................................................................................................................... 11  
Scope and model .............................................................................................................................. 14  
  Counting community investment.................................................................................................... 16  
Data requirements ............................................................................................................................ 18  
  Community investment activity data............................................................................................... 18  
  Community investment status data................................................................................................. 19  
  Costs data....................................................................................................................................... 20  
  Household, asset and neighbourhood data ..................................................................................... 21  
Methods .......................................................................................................................................... 23  
  Analytical approaches .................................................................................................................... 23  
  Data processing............................................................................................................................... 25  
Results – Employment status .......................................................................................................... 27  
  Responsive repairs.......................................................................................................................... 28  
  Actions pursuing rent arrears ......................................................................................................... 30  
  Anti-social behaviour (ASB) ............................................................................................................ 31  
Results – Digital inclusion status ..................................................................................................... 33  
  Responsive repairs.......................................................................................................................... 34  
  Actions pursuing rent arrears ......................................................................................................... 35  
  Anti-social behaviour (ASB) ............................................................................................................ 36  
Results: Community investment activity investigations .................................................................... 38  
  Employment and training activity .................................................................................................. 39  
  Digital inclusion activity .................................................................................................................. 40  
  Financial inclusion activity .............................................................................................................. 41  
Conclusions .................................................................................................................................... 44  
Findings............................................................................................................................................ 44
Data constraints .................................................................................................................................................. 45
Methods .............................................................................................................................................................. 47
Appendix A  Hypothesised pathways ..................................................................................................................... 49
Appendix B  Data schema ...................................................................................................................................... 51
  Person.............................................................................................................................................................. 51
  Asset ............................................................................................................................................................... 53
  Tenancy ........................................................................................................................................................... 57
  Household ......................................................................................................................................................... 60
  Housing benefit .............................................................................................................................................. 62
  Voids ................................................................................................................................................................. 63
Appendix C  Cost specification .............................................................................................................................. 65
Appendix D  Analysis ........................................................................................................................................... 68
  Propensity score methods ................................................................................................................................. 68
  Exploratory analysis ........................................................................................................................................ 73
Appendix E  Tables of results .............................................................................................................................. 74
  Employment status results ................................................................................................................................. 74
    Responsive repairs ...................................................................................................................................... 74
    Actions pursuing rent arrears .......................................................................................................................... 74
    Anti-social behaviour (ASB) ........................................................................................................................... 74
  Digital inclusion status results .......................................................................................................................... 75
    Responsive repairs ...................................................................................................................................... 75
    Actions pursuing rent arrears .......................................................................................................................... 75
    Anti-social behaviour (ASB) ........................................................................................................................... 75
Appendix F  Exploratory analysis: trend analysis of community investment activity ........................................... 76
Appendix G  Data processing ................................................................................................................................ 79
Appendix H  Chart design ..................................................................................................................................... 81
References ............................................................................................................................................................ 86
Table of figures

Figure 1 Example of possible pathway between community investment activity and lower expenditure in another part of the business ................................................................. 12

Figure 2 Example of possible route by which common causation might result in working households being associated with high repairs usage ......................................................... 15

Figure 3 Example of possible route by which common causation might result in non-working households being associated with high repairs usage ................................................. 16

Figure 4 Differences in repairs costs associated with no housing benefit and partial housing benefit, compared to full housing benefit ................................................................. 29

Figure 5 Differences in arrears collection costs associated with no housing benefit and partial housing benefit, compared to full housing benefit ......................................................... 30

Figure 6 Differences in ASB costs associated with no housing benefit and partial housing benefit, compared to full housing benefit ................................................................. 32

Figure 7 Differences in repairs costs associated with holding an email address compared to not holding an email address ................................................................................................. 34

Figure 8 Differences in arrears collection costs associated with holding an email address compared to not holding an email address ................................................................................................. 35

Figure 9 Differences in ASB costs associated with holding an email address compared to not holding an email address ................................................................................................. 36

Figure 10 Differences in costs associated with having participated in employment and training activity compared to not having participated for HP2 ......................................................................................................... 39

Figure 11 Differences in costs associated with having participated in digital inclusion activity compared to not having participated for HP6 ......................................................................................................... 40

Figure 12 Differences in repairs costs associated with having participated in financial inclusion activity compared to not having participated ......................................................................................................... 41

Figure 13 Differences in arrears collection costs associated with having participated in financial inclusion activity compared to not having participated ......................................................................................................... 42

Figure 14 Differences in ASB costs associated with having participated in financial inclusion activity compared to not having participated ......................................................................................................... 43

Figure 15 Distribution of propensity scores for digital inclusion for one housing provider ......................................................................................................................................................................... 71

Figure 16 Example of charts to check balance for digital inclusion for one housing provider, showing imbalance based on naïve comparisons (left-hand bars) and improved balance when using propensity score-weighted comparisons (right-hand bars) ......................................................................................................................................................................... 72
Foreword

Community investment and supporting tenants to live successful lives in vibrant neighbourhoods is nothing new for the affordable housing sector. A relatively hidden jewel in the crown, it has often been part of the 'heart' of the sector, something that expresses the social purpose of our organisations and recognises that the business of affordable housing is more than just providing a great place to live.

But in recent years, our investment in programmes that help tenants into work, to manage their money better or improve their health and wellbeing, has come under increasing scrutiny. We need to be clear about the value our investment delivers, for our tenants and communities, our businesses and our ability to address the housing crisis we all face.

HACT’s Social Value Bank is helping organisations to understand the outcomes they deliver and provides a way to measure and describe its social value. Housing providers already know that helping tenants to manage their money and tenancies, get them online, and into work can help mitigate the risks of welfare reform and Universal Credit, both for the individual and their business.

What has been missing is an understanding of the direct business benefits that Community Investment has on the bottom line. Asking fundamentally, do these activities not only improve the lives of residents and communities, but also provide a return on investment and save the business money?

This is the challenge we set out to answer. With our six housing association partners we used their data to model the impacts that their investments had on different parts of their business.

The project utilises data analytical methods not widely in use in our sector. It is a first step in unlocking the power of data that is collected day in, day out. There is significant potential for us to go further and use data driven insights to drive investment and improvements.

This research is a huge step forward. It demonstrates that successful community investment activities go beyond benefitting residents and improves the costs and demands on the core housing business. Given the financial pressures housing providers face, it puts community investment into the ‘must do’ rather than the ‘would like to do’ box.

It brings community investment together with other parts of the business, demonstrating how the skills and expertise of community investment professionals can benefit their colleagues in areas such as housing and asset management. It bridges the gap between different business areas, understanding the connectivity of different business areas and driving a more effective and efficient use of resources.
We are excited about what this sort of research and analysis can achieve, and the transformation it can drive in our businesses and communities.

Andrew van Doorn
Chief Executive
HACT
Executive summary

Housing providers’ community investment activity has the potential to create benefits in a range of ways. Alongside the benefits to the individual tenants and residents who participate, other impacts could include savings to local or national government, or financial savings to the housing provider.

This project analysed data from six housing associations to examine the relationships between community investment and organisations’ bottom line costs. It focused on three areas of community investment:

- employment,
- digital inclusion, and
- financial inclusion.

Within those areas, it analysed data on both community investment activities and statuses:

- Community investment activity: comparing people who have received or participated in some community investment activity with people who have not, and seeing whether the two groups are associated with different bottom line cost profiles.
- Community investment status: comparing people in different statuses that are the intended direct outcomes of community investment activity, and seeing whether the groups are associated with different bottom line cost profiles.

Differences in bottom line costs were examined across responsive repairs, rent arrears, and anti-social behaviour.

There are several routes by which these types of community investment might result in lower usage of housing provider services. However, prior to this project there had been few attempts to quantify the extent to which community investment activity is actually associated with lower costs in other parts of housing providers’ businesses.

This analysis used advanced data science methods – known as propensity score methods – to ensure that like-for-like comparisons were made between the people in different categories. Individual results were produced for each housing association, identifying the relationships between the different statuses and costs for their tenants.
Using these methods, the project identified a very consistent finding across the participating housing providers that being in **full time employment was associated with lower average levels of responsive repairs service usage** than being unemployed. The costs associated with tenants in part-time employment were typically somewhere between the two. These differences were large enough to be of practical significance, and if extrapolated across the sector would amount to a difference of some **£130 million per year** in England. **When looking at the participating housing providers individually this finding equated to an average annual savings, per tenancy, ranging from £22 to £342.**

There was also a relatively consistent finding that those in **part time (or perhaps low-paid) employment were associated with higher than average rates of activity by income officers chasing rent arrears**. In every case this was higher than arrears service usage for unemployed tenants, and in all bar one case it was also higher than for those in full time employment.

The findings of associations between other combinations of community investment outcomes and service expenditure seemed to be more variable, possibly indicating that they are dependent on the context of each housing provider.

The analysis was most powerful where it was able to draw upon **robust, relevant data, held for a large proportion of tenants**. In the case of estimating employment status, this was found from rent records of housing benefit payments, a data set that is consistently available for all tenants. In other areas the analysis was limited by constraints on the availability of data. Where data are not held, or where the intended measure is not a good indicator of the status of interest, the potential for analysis is lower. The participating housing providers did not generally hold historic records on the tenants that had participated in their community investment activities, resulting in limited sample sizes and an inability to examine the potential for cost differences to accrue over time. Similarly, the measure the project attempted to use for digital inclusion appears to be problematic in important respects.

Collecting data on these statuses and activities more consistently in the future will open up the potential for analyses to be conducted. Given the success of this project where data is consistently held, it is likely that similar future analysis would be able to identify any bottom line differences that occur when tenants have participated in community investment activity.
Introduction

Housing providers’ community investment activity has the potential to create benefits in range of ways. Alongside the benefits to the individual tenants and residents who participate, other impacts could include savings to local or national government, or financial savings to the housing provider.¹

The Community Investment and the Bottom Line project was launched with the aim of using the data held by housing providers to determine how community investment activity can be associated with different levels of bottom line costs incurred by the organisation. The findings of this project have significant implications for housing providers, and have the potential to influence business decisions around the community investment activities offered by organisations. This is especially true for housing providers that need to make decisions around the scale, scope and nature of their community investment programmes, in the light of increasing budget constraints. Drawing on its existing experience of using data science to investigate important business areas for housing providers, HACT undertook this pioneering project with Bracknell Forest Homes, Gentoo, Hyde, Helena, Knightstone, and Liverpool Mutual Homes as project partners.

There are many reasons to believe that community investment activity might have an impact on the bottom line of a housing provider. At the start of the project we drew up a map of the various intuitions and assumptions about the ways in which an impact might be delivered (see Appendix A). One example from the map of a (plausible) pathway by which community investment activity might ultimately result in lower service usage is shown in Figure 1, below:

Figure 1 Example of possible pathway between community investment activity and lower expenditure in another part of the business

The existence of these possible pathways to bottom line impact provided the rationale behind the project: they showed that it is plausible that there is a relationship, and that it was worth testing these relationships using data to see whether they occur in practice.

Understanding whether community investment activity contributes to the bottom line of a housing provider has practical implications, particularly around resourcing decisions. Housing providers are operating with constrained resources and need to identify what to spend their money on. A more complete understanding of whether parts of that investment tend to result in a bottom line impact helps to ensure those resource allocation decisions are informed by all the relevant evidence. Resolving issues like this can be out of scope of traditional business intelligence techniques, leaving gaps in how well a business understands its customers. One of the aims of this investigation is to support the emerging evidence agenda in the sector, helping it in its transition from one where business decisions are driven predominantly by ‘gut feelings’, to one where these decisions are more evidence informed.

The evidence from this project may also have the potential to support other business processes. For example, where it is found that there are differential repairs costs associated with tenants in different statuses, a housing provider may be able to use that information when negotiating a price-per-property repairs contract, to ensure it better reflects the expected service usage of the provider’s own tenant population.
The aim of this project is to investigate links specifically between community investment initiatives and the bottom line costs that housing providers incur. The project did not aim to investigate other impacts of community investment, such as its wider social value. In focusing purely on the financial implications the intention is to provide evidence that can be used to inform decisions, but not to have the final word on all potential benefits that can be achieved. HACT’s previous work² has examined some of the social value associated with community investment activity, and that remains a crucial factor in assessing its impact. This project complements the existing measures of the value of community investment, by enabling housing providers to consider the ‘blended’ return of both social and financial outcomes together.

Scope and model

With the diverse range of community investment activities offered by housing providers it was necessary to define a scope for the analysis, in particular covering the community investment areas that would be studied and the areas in which financial costs would be assessed.

The housing providers who were project partners identified three areas of community investment activity that were of particular interest for investigation:

- employment,
- digital inclusion, and
- financial inclusion.

We adopted an inclusive approach to financial data, seeking information from as wide a range of expenditure areas as possible. HACT worked with QA Housing Consulting, a partner organisation with significant experience of handling and understanding housing provider costs data, to support acquisition of the financial data. The approach adopted is described in more detail in Appendix C.

The overarching model that the project was aiming to investigate was to examine the relationship between community investment and the bottom line. In short, we wanted to create an estimate of what would happen with and without community investment. Attempts to identify what would have happened to people who received a community investment intervention, versus if they had not received it (and vice versa) is called counterfactual analysis.

To conduct counterfactual analysis robustly, we cannot just consider those that have and have not received the benefits of a community investment activity – these groups will tend to be different in other ways. For example, the groups of tenants with different working statuses (in work vs. out of work) may also differ in terms of their average household composition. It could, for example, be the case that those with more people in the home are more likely to be in work (for example if having children is a motivator for seeking work) or less likely to be in work (for example if difficulties accessing childcare make it harder to gain and retain employment). And on the costs side, perhaps having more people in the home is associated with higher repairs costs (e.g., due to higher wear and tear). In these examples, we would find that one of the statuses is associated with higher repair costs due to the association with different household composition – see Figure 2 and Figure 3. In this case the number of people in the home would be what is known as a ‘confounding factor’ (or a confounder).
Figure 2 Example of possible route by which common causation might result in working households being associated with high repairs usage
Figure 3 Example of possible route by which common causation might result in non-working households being associated with high repairs usage

Consequently, in order to gain a robust understanding of whether community investment activity leads to lower costs, we need to use more advanced methods that account for these confounding factors.

Counting community investment

Within the overarching model we identified two potential ways of counting community investment:

- **Community investment activity**: comparing people who have received or participated in some community investment activity with people who have not, and seeing whether the two groups are associated with different bottom line cost profiles.

- **Community investment status**: comparing people in different statuses, where those statuses are the intended direct outcomes of community investment activity, and seeing whether the groups are associated with different bottom line cost profiles.
Each of these approaches has its benefits and limitations. Analysis on the basis of activity looks at those tenancies where one or more person in the household has been through an intervention and aims to compare them robustly to those who have not participated in the activity, to see whether there is a lower associated bottom line cost. It aims to identify the difference associated with the specific community investment intervention(s) that the housing provider delivers, meaning it is more directly applicable to that organisation’s activities. For example, where a housing provider is delivering interventions that try to get tenants into work, this approach aims to identify whether receiving those interventions is associated with lower bottom line costs.

Analysis on the basis of statuses has the potential to be applied more generally, identifying the likely bottom line differences that could be observed for any programme of interventions that delivers that status. For example, where this approach is used to investigate tenants being in work, the aim is to see whether in-work tenants have lower bottom line costs than out-of-work tenants, without needing to know whether it was the housing provider’s activity that supported them into work. If the calculated value is sufficiently robust, it could potentially be applied to estimate the bottom line impact of any intervention that got people into work.

The two types of analysis also draw upon different data. Consequently, both analysis methods were attempted, in order to maximise the range of potential insights that the project could deliver and the options for data sources that could be used if available. ‘Activity investigations’ and ‘status investigations’ were conducted in parallel.
Data requirements

The data requirements for the project flowed from the scope and model identified above. The housing providers participating in the project provided data covering the following areas:

- **Community investment.** Two types:
  - **Community investment activity** records, where held, associated with each tenancy;
  - **Status** data providing information on whether each tenancy has a status that community investment activity is intended to achieve or not;
- **Costs.** Financial records providing an indication of the costs incurred relating to each tenancy;
- **Tenancy and asset** information. Details related to each tenancy of things that might be confounding factors.

The key factor tying each of these areas of data together is the tenancy: we effectively compare the costs of each tenancy with the community investment status or activity for that tenancy, whilst taking into account other things that we know about the tenancy that might influence the costs.

At scoping stage, the participants were asked about the data that could be produced; from this and further conversations a data schema was produced, specifying the exact format that data should be delivered in. Full details of the data schema are provided in Appendix B.

**Community investment activity data**

To investigate the specific effect of engagement in community investment activities, records of tenant participation in these activities were required.

Community investment activity data broadly included who participated in an activity, which activity they participated in and the time period in which they participated. Records were obtained, where possible, for all community investment activities for each of the interested areas – employment, digital inclusion and financial inclusion – that project partners provided.
Community investment status data

To perform status investigations, we needed data that would allow us to consider the status of each tenancy in relation to the community investment direct outcomes of interest. Employment support is intended to affect employment status, digital inclusion activity is intended to support people to be digitally included, and financial inclusion activity is intended to support people to be financially included. These statuses are not fields that housing providers will typically hold reliable data on. Housing providers may, for example, record a tenant’s employment status at the point of letting, and may seek to update those records where possible, but there will not normally be a requirement for tenants to inform their landlord of every change in employment status, so any data that is held can quickly get out of date. Consequently, we need to use alternative methods to seek to infer each tenant’s status. To do this we look to identify proxy measures – data that is held more reliably by the housing provider that gives an indication of the status in question.

For employment, housing providers do not typically hold up-to-date records of tenant’s employment status. They are, however, able to detect for each tenancy whether any of the rent due is received from housing benefit. Given the strong link between income and benefit entitlement, the proportion of rent received from housing benefit was identified as a plausible proxy for employment status amongst the age cohort whose data were included in the study (under-65s).

Digital inclusion status can be harder to define. One option that we investigated with the housing providers was whether they could provide data for each tenancy indicating whether the tenant had engaged with the landlord using any digital channels. It was established that these records would be difficult or impossible to produce. Housing providers were, however, typically able to extract an indication for each tenancy of whether they held an email address on record. It was clear that there may be some tenants who are digitally included that housing providers may not hold email addresses for, since some tenants who become digitally included may not have a particular reason to provide their email addresses to their landlords. (Conversely, some tenants may have email addresses registered but never use them in practice.) In the absence of another more plausible proxy, we decided to investigate whether this one would perform well enough anyway.

Financial inclusion was the one area where no reasonable proxy could be established. The desired outcomes of financial inclusion schemes are to improve the participant’s ability to manage their finances. One option would be to consider the rent arrears status of tenants as a proxy for being financially included. However, since the primary expected bottom line outcome of financial inclusion would be improvement in rent accounts, using the same variable as both the status flag and the outcome would inevitably result in circular reasoning. Even aside from this issue, rent accounts may not
be a good proxy for financial inclusion anyway, if housing benefit protects some of those who are unable to manage their other finances from falling into rent arrears.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Proxy</th>
<th>Assumed relationships between statuses of proxy and outcome of interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employment</td>
<td>Housing benefit as a proportion of rent due</td>
<td>Full housing benefit: unemployed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Partial housing benefit: part time employment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No housing benefit: full time employment</td>
</tr>
<tr>
<td>Digital Inclusion</td>
<td>Email address for tenancy on records</td>
<td>Email address – digitally included</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No email address – not digitally included</td>
</tr>
<tr>
<td>Financial Inclusion</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Costs data

In order to investigate whether community investment activity is associated with differences in the bottom line, it is naturally essential to examine costs. There are a range of costs that housing providers incur, and for the purposes of this project we are interested in those that are associated with individual tenancies. In order to gain a detailed understanding of the associations between community investment and costs we analysed data related to different types of expenditure separately. The participating housing providers were generally able to provide information on expenditure related to rent arrears, property repairs, and anti-social behaviour.

The ideal situation for examining the associations is to use actual costs related to a tenancy, such as the amount actually spent on repairs to the relevant property. Where
this was not available we were able to construct an approximate measure of per-tenancy costs by using records of the number (count) of actions that would have incurred costs, and the average cost per action in that area for that housing provider. For example, if a housing provider held data on the number of responsive repairs conducted in the year across its stock (say, 10,000 repairs), and the total cost of responsive repairs (for example, £2,000,000), then we could derive an average cost-per-repair (£200 in this example); if a tenancy had received one repair it would be allocated an estimated cost of £200, tenancies with two repairs would be allocated £400 estimated costs, and so on.

A project partner with considerable experience of working with financial data in the housing sector, QAHC, worked with the participating housing providers to collect and collate the costs data in a consistent format.

In practice, for each of the areas where costs data were used, the general availability was:

- **Responsive repairs** – generally available on a per-instance, per-tenancy basis, allowing yearly aggregates to be easily extracted on an individual basis.
- **Actions pursuing rent arrears** – here we are looking at the cost to the housing provider having to chase those in arrears, rather than the amount a tenancy is in debt. These costs were generally available as a total yearly aggregate cost, which was apportioned by the number of incidences per tenancy. (One constraint of this availability is that it was not possible to differentiate between tenancies receiving actions by different channels, and associated cost differences; for example, an email is cheaper than a letter, which is cheaper than a house visit.)
- **Anti-social behaviour (ASB)** – typically available as a yearly aggregate of total expenditure, which was apportioned by number of incidences per tenancy.

Further detail on the costs data are contained in Appendix C.

**Household, asset and neighbourhood data**

There are several factors related to households and properties that might be expected to influence the likely costs associated with a tenancy. For example, all other things being equal, we might expect larger homes to incur higher expenditure than smaller ones simply because there are more parts to go wrong. We might also expect larger families to make higher use of services than smaller ones, because some things are related to the number of people rather than being constant for any household (e.g. more people making more use of the property resulting in greater wear and tear). And
we might find that homes built in some periods are associated with higher repairs spending than those constructed at other times.

Because these factors may also be related to community investment status, we cannot assume that the differences in costs will be observed equally across each of the groups we are comparing. For example, there might be a greater number of large households in the unemployed group, or fewer large homes represented in the digitally excluded group.

In order to be able to select an analysis design that attempts to account for these differences, we needed to obtain data on households and assets. These included demographic and household composition data (such as age of tenants), as well as asset data (such as number of bedrooms and year of construction).

We also collected location data (in the form of postcodes) in order to be able to similarly include contextual neighbourhood factors that might be relevant (such as unemployment statistics).

Including this information in the analysis allows adjustments to be made to reflect the circumstances of different tenants. This extra information supports improved analysis of the association between community investment and businesses' bottom line.
Methods

The analytical approach selected was chosen to reflect the data that the housing providers were able to provide. One option that we considered, for example, was to use data on changes in tenant statuses over time in order to examine the effects of changes in status where tenants moved between statuses. This would have required, for example, information on the date that the housing provider obtained the email address from the tenant, or the date(s) on which tenants’ housing benefit statuses changed. Because the participating housing providers identified that they would not be easily able to provide reliable time series data for these statuses, we selected an investigation design that only relied on the tenant status at a point in time, and used the costs data from the housing association over the most recent financial year.

As described above, we took two different approaches to try and maximise the information we could gain from the project. The status investigations compared the difference in costs between those who had particular statuses (irrespective of how they achieved that status); the activity investigations compared the difference in costs between those who had been through community investment activity (irrespective of what status they had achieved at the end).

Aside from the process for categorising people, the approach in the two types of investigation was kept as similar as possible. The ultimate intention in each case is to produce an estimate of the likely differences in costs (service usage) associated with a particular change for a tenant:

- In the case of status investigations, the tenant experiences a change of moving from one status to another;
- In the case of activity investigations, the tenant experiences a change of moving from not having participated in community investment activity to having participated.

Consequently, the same analytical approaches can be used.

Analytical approaches

The overarching aim in selecting an analytical approach was to identify a method that would enable us to establish the association between the category the tenant is in (i.e. the status or the whether or not they have had the activity) and the costs incurred by the housing provider. In order to do this robustly, we needed to select a method that was able to account for the fact that there may be some important other differences
between the typical features of the tenancies where people are in different categories. If, for example, people living in larger homes were much more likely to be in one category than the other (say, more likely to be digitally included) then we might expect to see higher repairs costs for digitally included tenancies just because there is more to go wrong in a bigger home; this could skew the results, and show higher costs associated with digital inclusion, even if digital inclusion itself had no effect on the cost of repairs. In this example, getting someone digitally included would not tend to increase their repair costs, as it would not normally have an effect on the size of house they occupy.

To resolve this issue, any robust approach needs a method to account for differences between other factors that are associated with differences between the categories. One approach is to use exact matching: for each digitally included person in the dataset we try to find a digitally excluded person in the dataset who is living in the same sized house. Unfortunately, exact matching quickly becomes unfeasible as you increase the number of factors that you are trying to match on – once you have several factors, you are not able to find exact matches for many of the people. The ability of exact matching to provide an accurate assessment is also highly dependent on matching on the factors that turn out to be particularly important in their associations.

The approach used in this project to mitigate these challenges is a variant upon propensity score matching. Propensity score matching works by matching on a ‘propensity score’ rather than trying to match on each of the separate things that we know about a tenancy. For the investigations in this project we calculated propensity scores using the household, asset and neighbourhood data for each tenancy. The propensity score can be thought of as the probability that the tenancy would have been in a particular category, given the other characteristics we know about it (please see Appendix G for details of factors included in this analysis).

Because the propensity score aims to effectively condense all of the relevant information from the matching criteria into a single score, it is much easier to find suitable matches, as long as there is a reasonable amount of overlap in the scores for people in each of the categories. We used a particular algorithm, that is able to work with multiple categories, because in some cases (e.g. employment / housing benefit status) our analysis was comparing more than two states. This method actually uses the propensity scores to weight the outcomes of every tenancy in the dataset, rather than just looking for matches. Further details of this approach are provided in Appendix D.
The full list of questions the analysis was intended to answer was, for each cost area and for each housing provider…:

(Status investigations):

- …is there an association between different work statuses (as measured through a proxy of the housing benefit status of the tenancy) and costs incurred by the housing provider?
- …is there an association between different digital inclusion statuses (as measured through a proxy of the housing provider holding an email address for the tenant) and costs incurred by the housing provider?

(Activity investigations):

- …is there an association between having received employment support from the housing provider and costs incurred by the housing provider?
- …is there an association between having participated in digital inclusion community investment activity and costs incurred by the housing provider?
- …is there an association between having received financial inclusion support from the housing provider and costs incurred by the housing provider?

In most of these cases, the comparison was a simple comparison between two categories: for example, holding an email address vs. not holding an email address; or having participated in a particular type of community investment activity vs. not having participated. In the case of employment status, however, we were comparing between three categories: full housing benefit, partial housing benefit, and no housing benefit.

Data processing

We requested that housing providers supply data in as raw a form possible, so that any aggregations or feature extractions performed would be done consistently. Appendix G outlines the steps taken to process the data from the form requested in the data schema to that which was used in the analysis.

It was hoped initially that a pooling exercise would be feasible; this would have allowed us to gather all the data together from the different housing providers to extract insight. It soon became apparent that this would not be appropriate, as there are fundamental differences around how costs are recorded and reported between organisations. Therefore, it was decided that each housing association would be investigated in isolation, generating bespoke results for each project partner.
One additional consideration was made with respect to the employment status investigation. We chose to filter the tenants within the investigation to only include those tenancies where the first tenant was aged under 65, in order to minimise the influence on the analysis of the age group where it is more likely that tenants will have retired.
Results – Employment status

Summary: Investigations around employment status worked well, consistently finding lower costs associated with being in employment for some areas of expenditure; lower costs for other areas of expenditure were also found in some organisations. Housing benefit status appears to work well as a proxy for employment status.

As described above, the employment status investigations were designed to compare differences between three statuses: full housing benefit, partial housing benefit, and no housing benefit (as proxies for unemployment, part time employment, and full time employment respectively).

The propensity scores were calculated and the appropriate checks were run, indicating that there was acceptable balance between the groups.

All the graphs presented in the report illustrate whether a particular type of tenant is associated with an increased cost or a saving to the business compared to a baseline tenant type. The employment status results below are presented showing the associated cost differences to the housing provider for those on partial or no housing benefit, relative to those on full housing benefit. In each case, results are presented as the percentage difference from a baseline state. Results that appear above the ‘0’ line, appearing in the green section, indicate a saving while results in the red section (below the ‘0’ line) indicate an increased cost relative to the baseline, in this instance, of full benefit status.

The use of percentage changes simplifies comparisons when looking across organisations, as they may be less clear when examining absolute values due to organisations’ differing approaches to allocating costs or designing spending strategies, especially where an organisation may have some ability to reduce its costs in one area through higher spending in another. A housing provider’s baseline level of expenditure on responsive repairs, for example, may be influenced by the extent it chooses to focus on planned maintenance, or the typical number of ASB incidents it responds to may be affected by its level of investment in ASB diversionary activity.

The graphs display a central point, which is the best estimate of the most likely value of the expected percentage cost difference. The bars either side of each central estimate
show the likely range around the central point; the true value is very likely to lie within the ranges spanned by the bars. (See Appendix D for details.) Smaller ranges indicate more confidence in applying the findings, with the size of the range being down to a number of factors, mostly around the quantity and quality of the available data.

NB Where error bars cross or are close to the ‘0’ line the results should be taken with caution as it becomes likely that there is no difference in costs to the business between statuses, or the ‘true’ cost difference could be in the opposite direction.

The graphs have been designed to show the relative costs associated with different statuses symmetrically. In each case, whichever cost estimate is the lower in the comparison is plotted as the percentage it is lower than the other cost estimate. (See Appendix H for further information.)

Responsive repairs

A consistent and substantial difference in costs was found when investigating responsive repairs. Across all of the participating housing providers, tenancies with no housing benefit were found to be associated with lower responsive repair costs, with central estimates ranging from 16 to 34% less than tenancies on full housing benefit. In every case the bars around the central estimates are entirely above 0%, indicating a low likelihood of a zero (or negative) association.

Therefore, these results reasonably suggest that tenancies on no housing benefit are likely to save the business money in relation to responsive repairs relative to those on full housing benefit. In monetary terms this equated to an annual saving of £22 to £342 per tenancy depending on the individual housing provider.

---

Ranges are calculated by adding and subtracting the standard error from the central estimate.
In general, tenancies claiming partial housing benefit have lower average repair bills than those on full housing benefit but not as low as those on no housing benefit. In each housing association except one, the bars for the partial housing benefit tenants are entirely above zero, again indicating a low likelihood that the true association is zero or negative.

In one case the central estimate of the lower costs associated with partial housing is greater than that housing provider’s central estimate for lower costs of no housing benefit. The substantial overlap of the bars on the two estimates indicates that the findings are also consistent with the findings in the other housing providers (i.e., there are lots of options within the reasonable ranges that would show lower costs from no housing benefit than from partial housing benefit).

These highly consistent findings are indicative that there appears to be a reliable association across a range of contexts. Although this analysis is not able to tell us exactly the mechanisms by which savings are made, there are several plausible reasons, such as:

**Figure 4** Differences in repairs costs associated with no housing benefit and partial housing benefit, compared to full housing benefit
Those tenants who are in work may tend to be out of the house more, so may be resulting in lower wear-and-tear to the property. Tenants who are unemployed may have more spare time around the house, so may be more likely to be annoyed by minor issues or have more time to report them. Employed tenants may be more likely to have the money to afford to do minor jobs around the house on a DIY basis.

Actions pursuing rent arrears

The relationship between costs of chasing rent arrears and employment status was found to vary more between housing providers. Partial housing benefit status was consistently found to be associated with higher costs of pursuing rent arrears. The pattern for no housing benefit was more varied. The graph illustrates the differences in relative costs identified in different housing providers.

Figure 5 Differences in arrears collection costs associated with no housing benefit and partial housing benefit, compared to full housing benefit
The findings confirm the common observation that partial housing benefit recipients are associated on average with more incidents of rent arrears. In every case, the cost of pursuing rent arrears for those on partial housing benefit was greater than the equivalent costs for those on full housing benefit, although there is substantial variability in the degree of this difference (with central estimates ranging from around 5% to 55%). Whilst this analysis cannot confirm the specific mechanisms by which this occurs, it is consistent with a belief that partial housing benefit is often associated with relatively unstable employment, resulting in frequent fluctuations in benefit payments and hence opportunities for rent payment shortfalls.

The differences relating to partial housing benefit that are observed between housing providers could be due to different contextual factors – for example if the market for part time employment in the housing providers’ operating areas are characterised by different degrees of stability. They could also reflect housing providers’ operational differences, such as the triggers they set for taking actions that they record as pursuing rent arrears: a housing provider taking action immediately after a small discrepancy has occurred may record higher relative costs for those on unstable incomes than one that only starts pursuing rent arrears that have built up to a particular level.

In every case except one, tenancies in receipt of no housing benefit were associated with lower costs of pursuing rent arrears than those on partial housing benefit. This is consistent with expectations that partial housing benefit tenancies are associated with the highest rates of rent arrears incidents. In the remaining case there is a degree of overlap in the bars, which indicates that it is plausible that the true difference is also in the same direction in this housing provider.

The largest variation in the findings relates to whether (and the extent to which) lower costs are observed for those on no housing benefit compared to those on full housing benefit. In half of the housing providers we found that the ‘no housing benefit’ status was associated with lower costs than the ‘full housing benefit’ status; in the other half this was reversed. As with the differences observed in relation to partial housing benefit these differences may relate to various factors in the context or operational practices of the housing providers. Given the size of the differences, and particularly the variation between costs sometimes being higher and sometimes being lower, these may merit further investigation.

Anti-social behaviour (ASB)

Not all organisations were able to provide costs data around ASB. For those that did, there were consistently lower costs associated with those receiving no or partial housing benefit, compared to those on full housing benefit.
In every case, the central estimates are that those in receipt of no or partial housing benefits are associated with lower costs than those on full housing benefit. In two out of the three housing providers these differences are such that the bars are entirely above zero, indicating a low likelihood that the real difference is zero or below. In the remaining case the relatively large bars cross zero, meaning there is a chance that the real difference is close to zero (and a relatively small chance that the difference could be in the other direction).

In two out of the three cases the central estimates indicate that the costs for those in receipt of no housing benefit are lower than those on partial housing benefit. This is consistent with the theory that those who are out of the house more are around less to result in ASB costs. In the remaining case the bars overlap substantially, indicating that the findings are consistent with an outcome in the same direction as the other housing providers (or with a finding that for that housing provider the relative costs are reversed).
Results – Digital inclusion status

Summary: Investigations around digital inclusion status did not work well, which we suspect is because of our choice of proxy for digital inclusion status. In several areas we found higher costs associated with holding an email address for the tenancy. This appears to be because housing providers are more likely to obtain email addresses from tenants they engage with frequently (so being a high repairs service user, for example, causes someone to appear as digitally included by this measure), not because digital inclusion is causing higher service usage.

As described above, the digital inclusion status investigations were designed to compare differences between two statuses: tenancies for which the housing provider holds an email address and those for which it does not hold an email address (as proxies for digitally included and not digitally included, respectively). As noted above, the adopted proxy is not a perfect match for whether tenants are digitally included, but was used to facilitate some investigation of this area since housing providers were able to supply data on it. If housing providers were able to build records that more directly measure whether tenants are digitally included or not, this analysis could be re-run using a more accurate measure.

The propensity scores were calculated and the appropriate checks were run, indicating that there was excellent balance between the groups.

The results below are presented showing the associated cost differences to the housing provider for those that email addresses were held for, relative to those that email addresses were not held for. The graphs display a central point, which is the best estimate of the most likely value of the expected percentage cost difference. The bars either side of each central estimate show the likely range around the central point; the true value is very likely to lie within the ranges spanned by the bars. (See Appendix D for details.)
Responsive repairs

The relationship between holding an email address and the costs of repairs for a tenancy was found to vary considerably between housing providers.

Figure 7 Differences in repairs costs associated with holding an email address compared to not holding an email address

The chart shows that the central estimate for the difference in the costs of repairs associated with tenancies where the housing provider holds an email address varies between the costs being somewhat lower and substantially higher. In several cases the bars cross zero, indicating that there is a reasonable chance that the actual association is zero or opposite in direction to the central estimate.

When we examined the association between holding an email address and the repair costs of a tenancy, we identified a particular reason for caution over the interpretation of these results as a proxy for digital inclusion. In many housing providers, standard processes mean that when the organisation is in contact with tenants (e.g. when they call to report repairs), the tenants are requested for their email addresses.
Consequently, rather than measuring the association between digital inclusion and repair costs, it appears that this investigation is tending to measure the fact that higher contacts often cause the organisation to be more likely to hold an email address.

**Actions pursuing rent arrears**

The relationship between holding an email address and the costs of pursuing arrears for a tenancy was found to vary between housing providers.

![Figure 8 Differences in arrears collection costs associated with holding an email address compared to not holding an email address](image)

In each case, the central estimate of the difference in costs is that holding an email address is associated with higher costs for pursuing arrears, relative to tenancies where no email address is held. In most of the housing providers this association was found to be up to 10%; in one it was nearly 30%.
As described above, there are reasons to believe that holding an email address is a poor proxy for digital inclusion status. The higher costs that are associated with holding an email address could be a reflection of the fact that housing providers will tend to keep better records of contact details for tenants that they are in contact with more frequently for whatever reasons.

Anti-social behaviour (ASB)

Not all organisations were able to provide costs data around ASB. For those that did, the relationship between holding an email address and the costs of ASB for a tenancy was found to vary between housing providers.

![Figure 9 Differences in ASB costs associated with holding an email address compared to not holding an email address](image)

The central estimates of the difference in costs for holding an email address varied between slightly lower costs and substantially higher costs, relative to tenancies where no email address is held.
As described above, there are reasons to believe that holding an email address is a poor proxy for digital inclusion status. The higher costs that are associated with holding an email address could be a reflection of the fact that housing providers will tend to keep better records of contact details for tenants that they are in contact with more frequently for whatever reasons.
Results: Community investment activity investigations

Summary: Investigations around community investment activity were unable to identify associations between the activity and costs, largely due to the limited extent of data available. Constraints on historic data in particular meant that the analysis was restricted to looking at data falling within one financial year. As well as resulting in small sample sizes, which limit our ability to identify associations in any event, it is also possible that associations were not found simply because a longer period would be required to observe the activities’ impacts on the bottom line.

The community investment activity investigations were designed to compare differences between people who participated in a community investment activity compared to those who had not. Each of the three areas of community investment activity – employment and training, digital inclusion, and financial inclusion – were investigated separately.

The propensity scores were calculated for interventions taking place within the first 6 months of the 2014/15 financial year. The appropriate checks were run, indicating that there was acceptable balance between the groups.

The results below are presented showing the associated cost differences to the housing provider for those tenancies where at least one tenant participated in a community investment scheme, relative to those tenancies where no one participated. The graphs display a central point, which is the best estimate of the most likely value of the expected percentage cost difference. The bars either side of each central estimate show the likely range around the central point; the true value is very likely to lie within the ranges spanned by the bars. (See Appendix D for details.) Not all of the participating housing providers were able to provide data on all three community investment areas and in some cases the sample sizes were too small to conduct the analysis.

Due to the limitations described above, we also conducted an exploratory analysis to see whether it was possible to detect associations between community investment activities and costs using a different method. This approach is outlined in Appendix F.
NB: Please note that due to the limitations of the community investment activity investigations caused by data availability issues it is not advisable to rely on the results in this section of the report in decision-making.

Employment and training activity

The chart shows that the central estimate for the difference in the costs associated with tenancies where at least one tenant went through an employment and training activity, relative to tenancies who did not go through the activity varies. The arrears cost is somewhat lower while Repairs and ASB costs are somewhat higher. In all cases the bars cross zero, indicating that there is a reasonable chance that the actual association is zero or opposite in direction to the central estimate.
In each case, the central estimate of the difference in costs is that tenancies where at least one tenant went through a digital inclusion activity is associated with lower costs for responsive repairs and pursuing arrears, relative to tenancies that did not go through an activity. For responsive repairs this difference is such that the bar is entirely above zero, indicating a low likelihood that the real difference is zero or below. Whereas for pursuing arrears the relatively large bar crosses zero, meaning there is a chance that the real difference is close to zero (and a relatively small chance that the difference could be in the other direction).

*Figure 11* Differences in costs associated with having participated in digital inclusion activity compared to not having participated for HP6
Financial inclusion activity

![Diagram showing differences in repairs costs](image)

*Figure 12 Differences in repairs costs associated with having participated in financial inclusion activity compared to not having participated*

In each case, the central estimate of the difference in repairs cost is that tenancies where at least one tenant went through a financial inclusion activity is associated with somewhat higher costs, relative to tenancies that did not go through the activity. In all cases, the bars cross zero, indicating that there is a reasonable change that the actual association is zero or opposite in direction to the central estimate.
The chart shows that the central estimate for the difference in the costs associated with tenancies where at least one tenant went through a financial inclusion activity, relative to tenancies who did not go through the activity varies between somewhat lower cost and higher cost. In one case the bars cross zero, indicating that there is a reasonable chance that the actual association is zero or opposite in direction to the central estimate. Whereas, in the other case the bar is entirely below zero, indicating a low likelihood that the real difference is zero or above.

Figure 13 Differences in arrears collection costs associated with having participated in financial inclusion activity compared to not having participated

The chart shows that the central estimate for the difference in the costs associated with tenancies where at least one tenant went through a financial inclusion activity, relative to tenancies who did not go through the activity varies between somewhat lower cost and higher cost. In one case the bars cross zero, indicating that there is a reasonable chance that the actual association is zero or opposite in direction to the central estimate. Whereas, in the other case the bar is entirely below zero, indicating a low likelihood that the real difference is zero or above.
In each case, the central estimate of the difference in ASB cost is that tenancies where at least one tenant went through a financial inclusion activity is associated with lower costs, relative to tenancies that did not go through the activity. The bars cross or sit on zero, indicating that there is a reasonable chance that the actual association is zero or opposite in direction to the central estimate.

Figure 14 Differences in ASB costs associated with having participated in financial inclusion activity compared to not having participated

Larger numbers equal greater savings compared to not having participated
Conclusions

Overall, the project demonstrated that we can establish robust associations between community investment outcomes and bottom line ones in areas where good data were available. Where important data constraints were present, these appear to impose limits on the analyses that can be conducted. Consequently, it is likely that where good data are available the techniques used in this project could produce further valuable insights. In other areas, work should be undertaken to improve data availability in order to make these powerful techniques available across more aspects of housing providers’ activity.

Findings

This project has established that it is possible to generate robust assessments of the association between different statuses related to community investment activity and the bottom line of housing providers. Where relevant data are available, we are able to use propensity score methods to compare the different costs associated with different statuses in a way that accounts for the potential influence of a range of other factors.

The best data available in this project were related to employment status (as measured through a proxy of housing benefit receipts) and responsive repairs costs. For these we found that employment status was consistently associated with differences in repairs expenditure, with those in full time employment (receiving no housing benefit) having lower costs of repairs than those on full housing benefit, whilst those on partial housing benefit are somewhere in between. Further research would be required to investigate whether the findings from the six housing providers in this study are generalisable on a national basis, but the consistency between them is suggestive of an association that is observed across a range of contexts. The consistency of findings from these propensity score methods also supports the case for investigating whether savings of these levels can be obtained through interventions that support people into work.

These questions are of great practical significance. To put the current findings in context, routine maintenance costs the social housing sector in England in excess of
£1.8 billion per year\(^4\), with 31% of tenants unemployed / inactive and another 11% in part time employment.\(^5\) Based on these figures, if the differences in repair costs observed in this study were extrapolated across the sector, approximately £130 million per year more is being spent on repairs than would be if these households were in full time employment.

### Data constraints

The most successful investigations in this project were around variations in costs associated with employment status. A key factor in this success was the extensive availability of robust data: using housing benefit receipts we are able to make a good assessment of the likely employment status of almost every tenant.

The availability (or unavailability) of suitable data is a critical success factor for the analyses in this project more generally. The methods adopted work well where lots of data is held, is reliable, and where it relates directly to the issue of interest. In some cases, data like this will be purposely collected in order to support these analyses, but the experience of this project is that it is perhaps most likely to be available where such data are a by-product of standard processes that apply equally to all tenancies.

The available data for this project’s digital inclusion investigations fell short of the requirements to be extensive, reliable and relevant in a few ways. As we knew in advance, holding an email address for a tenancy is not a perfect measure of digital inclusion: tenants may be digitally included and simply have not provided their email addresses to the housing provider; or, in some cases, they may not really be digitally included in a meaningful sense, but the housing provider may have obtained an email address for them at some point. The misidentification of a proportion of people (potentially in both directions) would tend to reduce measured effect sizes, if there were a real effect.

Probably a bigger factor in these investigations failing to generate meaningful findings is the fact that the data generation process probably did not apply equally to all tenants. The imperfect marker of digital inclusion status may have been good enough if it were not for selection biases inherent in the processes by which it is collected; because


tenants’ email addresses are disproportionately collected for those tenants who make contact with the housing provider to use services, results related to tenants’ service usage will be skewed.

Ahead of conducting the investigations we were aware of the potential constraints of the inaccuracy of the proxy measure and felt that it was worth trying anyway as a somewhat moderated estimate of the effect could be more useful than having no estimate at all. However, the fact that the measure was not independent of the cost outcomes being assessed was unforeseen. Future investigations of this area should seek to use alternative measures that ideally should be more directly related to digital inclusion status and certainly independent of the outcomes to which associations are being examined. These might include data on whether the tenant ever engages with the housing provider online or a more direct measure if a housing provider has asked tenants about whether they are online or not in any surveys they conduct.

One further constraint in relation to the digital inclusion investigations in this project was the unavailability of data on a particularly relevant area of cost differences. One area where housing providers might expect to see lower costs for digitally included tenants is around interactions across all service areas due to ‘channel shift’: online interactions are generally thought to cost far less per interaction than face-to-face or telephone ones. If an appropriate measure of digital inclusion could be found, measuring the costs associated with the channel of interactions may be an important area for assessing associations.

Conversely, data on housing benefit status appears to be a good fit in this project. It has the benefit of being available for every tenancy: it is possible to extract from rent records the proportion of rent that was received from housing benefit, whether that is 0%, 100%, or somewhere in between, and rent records will be kept accurately for all tenancies. The remaining opportunities for the proxy to be a poor fit are relatively small:

- there may be a small number of instances where tenants are in receipt of housing benefit but it gets paid to them rather than the housing provider;
- in some cases, tenants will be eligible for housing benefit but not claiming (or, conversely, be ineligible but receiving housing benefit through fraud or error);
- in some cases, low paid full time employment may not be sufficient to lift the tenant’s income above the housing benefit threshold so some ‘partial payers’ may be in full time rather than part time employment; and
- in some cases, tenants may have moved onto Universal Credit, placing them in a different benefits system from the one this study assumes.
Overall, it appears to be a relatively strong proxy for employment status amongst working age tenants with the mis-categorisation of tenants likely to be low. For future studies, the social security system changing to Universal Credit will become more important, and it will be necessary to consider whether this proxy remains valid as more tenants move onto it.

Even in the case of employment status, some constraints remain relating to the data. For each tenancy, the measure of its housing benefits status was just recorded at a point in time, whilst the costs data were aggregated over a financial year. Further analysis could be possible if a measure of how the status changed over the year were available. One option for future investigation would be to consider use of a measure for each tenancy of the proportion of the rent coming from housing benefit over the year in question. Depending on its design, such a measure might tend to treat equally those who were receiving half of their rent from housing benefit all year and those who received no housing benefit for six months and full housing benefit for six months (which may be fine in terms of most of the hypothesised associations). Even more detailed examination might be possible if the range of costs and status data were available on a time series basis.

**Methods**

The requirement of this project was to use a method of analysis to identify differences in bottom line costs for tenants in various groups, whilst ruling out as many alternative explanations of those differences as possible. The best way to be certain that any observed variations are due to the thing you are interested in is to use a randomised controlled trial, and randomly allocate tenants into the various states; that way any other differences in tenants’ statuses will be randomly distributed between the different groups. As this project was using existing data, those methods that would enable us to be completely confident that the differences between groups were solely due to the thing we are interested in (especially such as randomised controlled trials) were not available to us. It would be impossible, for example, to go back in time to allocate tenants at random to one of the states.

Simply comparing the tenants in different groups would not give us a clear indication of the impact of being in those groups as the differences could have been due to various other factors. This project therefore deployed a propensity score method in order to be able to analyse differences between groups. The use of a propensity score enabled us to effectively create groups for comparison that were balanced across a wide range of factors.

One of the main challenges that remains after conducting data analyses like those in this project is around the extent to which they should support causal interpretation,
i.e., how likely is it that the calculated differences in bottom line between the statuses is caused by being in one group as opposed to the other. Whilst data analysis will rarely conclusively prove a causal relationship, propensity score methods aim to rule out several other possible causes.

This project provides a particular example of why data analysis without careful interpretation should not be relied upon as proof of a particular causal relationship. It is clear, in hindsight, that the differences observed around digital inclusion are most likely to be the result of high service use driving the (proxy) measure of the status, not the status driving the service usage. The propensity score method may well have identified a causal relationship, but if so it may be in the opposite direction to the one we are interested in.

For the employment status investigations, we find that the methods used give us good results and identify associations where it seems likely that an underlying causal relationship exists. The use of propensity scores means that the findings should not be due to differences in the groups on the variables that were used to calculate the propensity scores (the household, asset and neighbourhood data). However, as we observed with the digital inclusion status investigation, finding an association using propensity scores does not, on its own, prove that the status is causing the bottom line difference. It could be possible that the association could be driven in the other direction, or that some other important factor that is associated both with service usage and our measure of employment status has not been accounted for in our propensity scores. By considering together the findings of the propensity score method with our ‘domain knowledge’, we can say that we think other potential explanations are unlikely.

Because we must acknowledge that our attempts to think about other possible drivers of the results are fallible and not data-driven, we have to accept that other explanations could remain. Associations identified through data analysis of this type should therefore be treated as indicative of plausible causal relationships, where there are not alternative more likely explanations. Where feasible and appropriate it may make sense to further study these using methods like randomised controlled trials.
Appendix A  Hypothesised pathways

Ahead of undertaking any data analysis we produced a map of potential pathways by which community investment activity might have an impact on the bottom line of housing providers. This map illustrates that there are many reasons for intuitively believing that community investment could have an impact on the bottom line, and provides a motivation for wanting to test whether these can be detected in practice.

One example from the overall map of a (plausible) pathway by which community investment activity might ultimately result in lower service usage is shown below:

The full map is shown below, illustrating the wide variety of ways in which various community investment activities might contribute to differences in service use. The map outlines plausible links between community investment and service usage. It also gives an idea of how a potential association might break down – if any of the assumptions encoded as points within a given chain do not hold in practice there will not be an impact through that route. Consequently, whilst it neither proves nor disproves the existence of a causal relationship, it does support the justification for investigating whether any differential costs are observed in practice.
## Appendix B  Data schema

### Person

<table>
<thead>
<tr>
<th>Field</th>
<th>Field description</th>
<th>Requested format for data supply</th>
<th>Notes/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>housing_provider_ref</td>
<td>Reference code allocated to the housing provider.</td>
<td>Categorical field: select one from: {LMH, BFH, GEN, KNI, HYD, HEL}</td>
<td>Each housing provider will be provided with a reference code to use by HACT</td>
</tr>
<tr>
<td>person_id</td>
<td>The unique id that references a particular person</td>
<td>Character field with no special characters or white space</td>
<td>Your standard person reference numbers. This will probably be from your housing management system.</td>
</tr>
<tr>
<td>date_of_birth</td>
<td>The date of birth of the person.</td>
<td>Date field with format &quot;yyyy-mm-dd&quot;.</td>
<td>In cases where this is unknown/incorrect leave this field unpopulated</td>
</tr>
<tr>
<td>Field</td>
<td>Field description</td>
<td>Requested format for data supply</td>
<td>Notes/Comments</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------------------------------------------------------------------------------</td>
<td>----------------------------------</td>
<td>-------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>gender</td>
<td>The gender of the person. Categories as follows: M (male) F (female) NK (not known)</td>
<td>Categorical field: select one from {M, F, NK}</td>
<td>Any unknown or tenants who do not identify as male or female should be classed as NK.</td>
</tr>
<tr>
<td>email_address</td>
<td>Is an email address held in records for the person? Categories as follows: Y yes N no</td>
<td>Categorical field: {Y, N}</td>
<td>If information not held, leave unpopulated.</td>
</tr>
<tr>
<td>first_online_interaction</td>
<td>Date of person's first online interaction with HA.</td>
<td>Date field with format &quot;yyyy-mm-dd&quot;</td>
<td>Date of first online interaction with HA (email, online form, survey etc.). If information not held, leave unpopulated.</td>
</tr>
</tbody>
</table>
## Asset

<table>
<thead>
<tr>
<th>Field</th>
<th>Field description</th>
<th>Requested format for data supply</th>
<th>Notes/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>property_id</td>
<td>The unique id that references a particular property.</td>
<td>Character field with no special characters or white space</td>
<td>Your standard property reference numbers. Only include dwellings, not garages, storage units, etc. where considered separate to dwellings.</td>
</tr>
<tr>
<td>property_type</td>
<td>The dwelling type of the property. Categories described as follows: <strong>Bungalow</strong> - A dwelling which occupies only one floor at ground level and has no other dwelling or non-domestic unit either above or below it. <strong>House</strong> - A dwelling other than a bungalow, which occupies more than one floor and has no other dwelling or non-domestic unit either above or below it. <strong>Flat</strong> - A dwelling which occupies only one floor and is located in a structure which contains other dwelling or non-domestic units on floors above and/or below. <strong>Maisonette</strong> - A dwelling which occupies more than one floor.</td>
<td>Categorical field select one from: {Bungalow, House, Flat, Maisonette, Other}</td>
<td>You may have more specialised names for your properties but define them under the broader definitions. For example, a mid-terrace house will come under House. Where not known leave unpopulated</td>
</tr>
<tr>
<td>Field</td>
<td>Field description</td>
<td>Requested format for data supply</td>
<td>Notes/Comments</td>
</tr>
<tr>
<td>---------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>----------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>construction_date</td>
<td>Date of construction. Please give to the best available accuracy (e.g. year or decade if exact date not available).</td>
<td>Date field with format &quot;yyyy-mm-dd&quot;.</td>
<td>The date should relate to when the oldest part of the dwelling, or the structure in which the dwelling is contained, was originally completed. If unknown please leave field unpopulated. If only partially know complete as first day of year or decade e.g. 1960's building as 1960-01-01.</td>
</tr>
</tbody>
</table>

and is located in a structure which contains other dwelling or non-domestic units on floors either above or below **Other**- any other dwelling type.
<table>
<thead>
<tr>
<th>Field</th>
<th>Field description</th>
<th>Requested format for data supply</th>
<th>Notes/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>number_of_bedrooms</td>
<td>Number of rooms intended for sleeping.</td>
<td>Integer field with no special characters. If not applicable complete as -1</td>
<td>Where a bedroom has been converted to another use such as an office, it should still be counted as a bedroom. Conversely a living room now used as a bedroom should not be counted as a bedroom. Where not known leave unpopulated</td>
</tr>
<tr>
<td>postcode</td>
<td>The postcode of the property</td>
<td>Character field without spaces</td>
<td></td>
</tr>
</tbody>
</table>
## Tenancy

<table>
<thead>
<tr>
<th>Field</th>
<th>Field description</th>
<th>Requested format for data supply</th>
<th>Notes/Comments</th>
<th>Purpose of obtaining</th>
</tr>
</thead>
<tbody>
<tr>
<td>tenancy_ref</td>
<td>The id which relates a person to a specific tenancy (ie one tenancy may have multiple persons residing).</td>
<td>Character field with no special characters or white space</td>
<td>Your standard tenancy reference numbers.</td>
<td>Will allow data sets to be merged</td>
</tr>
<tr>
<td>property_id</td>
<td>The unique id that references a particular property.</td>
<td>Character field with no special characters or white space</td>
<td>Your standard property reference numbers.</td>
<td>Will allow data sets to be merged</td>
</tr>
<tr>
<td>tenancy_start_date</td>
<td>The date the tenancy started residing at the property.</td>
<td>Date field with format &quot;yyyy-mm-dd&quot;</td>
<td></td>
<td>Will allow tenancies to be followed over multiple residences</td>
</tr>
<tr>
<td>Field</td>
<td>Field description</td>
<td>Requested format for data supply</td>
<td>Notes/Comments</td>
<td>Purpose of obtaining</td>
</tr>
<tr>
<td>------------------</td>
<td>------------------------------------</td>
<td>------------------------------------</td>
<td>-------------------------------------------------------------------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>tenancy_end_date</td>
<td>The date the tenancy ended residence at the property.</td>
<td>Date field with format &quot;yyyy-mm-dd&quot;</td>
<td>Leave field unpopulated where the tenancy is current / ongoing.</td>
<td>Will allow tenancies to be followed over multiple residences</td>
</tr>
<tr>
<td>Field</td>
<td>Field description</td>
<td>Requested format for data supply</td>
<td>Notes/Comments</td>
<td>Purpose of obtaining</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>----------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>weekly_rent</td>
<td>The most recent weekly rent level of the property/tenancy. Worked out as follows: If rent period is Fortnightly - divide the rent due by 2. If rent period is Monthly - multiply rent due by 12, then divide by 52. If rent period is Quarterly - multiply rent due by 4, then divide by 52. If rent period is Annually - divide rent period by 52.</td>
<td>Numerical field with no special characters</td>
<td>This should not include any arrears accrued to date, and should not be inclusive of any fixed service charges.</td>
<td>Explanatory factor to be included in modelling / factor for group matching</td>
</tr>
</tbody>
</table>
### Community Investment and the Bottom Line

<table>
<thead>
<tr>
<th>Field</th>
<th>Field description</th>
<th>Requested format for data supply</th>
<th>Notes/Comments</th>
<th>Purpose of obtaining</th>
</tr>
</thead>
</table>
| has_paid_online | Has the tenancy ever made an online rent payment to the HA?  
Y yes  
N no | Categorical field: \{Y, N\} | If information not held, leave unpopulated. | Will be used as a proxy for a tenancy being digitally included, allowing status to be tracked over time |

### Household

<table>
<thead>
<tr>
<th>Field</th>
<th>Field description</th>
<th>Requested format for data supply</th>
<th>Notes/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>tenancy_ref</td>
<td>The id which relates a person to a specific tenancy (i.e., one tenancy may have multiple persons residing).</td>
<td>Character field with no special characters or white space</td>
<td>Your standard tenancy reference numbers.</td>
</tr>
<tr>
<td>Field</td>
<td>Field description</td>
<td>Requested format for data supply</td>
<td>Notes/Comments</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>person_id</td>
<td>The unique id that references a particular person</td>
<td>Character field with no special characters or white space</td>
<td>Your standard person reference numbers. This will probably be from your housing management system.</td>
</tr>
<tr>
<td>person_type</td>
<td>The tenant type of the person. Categories described as follows: <strong>Tenant</strong> - member of the household who are in the contract on the property <strong>NonTenant</strong> - member of the household (which would include children etc.) who are not in the contract on the property</td>
<td>Categorical field: select one from {Tenant, NonTenant}</td>
<td>If distinction not held (because only tenants are recorded) please complete as 'Tenant'</td>
</tr>
</tbody>
</table>
### Housing benefit

<table>
<thead>
<tr>
<th>Field</th>
<th>Field description</th>
<th>Requested format for data supply</th>
<th>Notes/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>tenancy_ref</td>
<td>The id which relates a person to a specific tenancy (i.e., one tenancy may have multiple persons residing).</td>
<td>Character field with no special characters or white space</td>
<td>Your standard tenancy reference numbers.</td>
</tr>
<tr>
<td>housing_benefit_level</td>
<td>Level of housing benefit. Categories as follows: <strong>Full</strong> <strong>Partial</strong> <strong>None</strong></td>
<td>Categorical field: select one from: {Full, Partial, None}</td>
<td>If only the current level of housing benefit it held, enter the current level</td>
</tr>
<tr>
<td>benefit_level_start_date</td>
<td>Date of which the benefit level starts.</td>
<td>Date field with format &quot;yyyy-mm-dd&quot;</td>
<td>If time series is unavailable, leave unpopulated</td>
</tr>
<tr>
<td>benefit_level_end_date</td>
<td>Date of which the benefit level ends</td>
<td>Date field with format &quot;yyyy-mm-dd&quot;</td>
<td>If benefit status is current, or time series unavailable, leave field unpopulated</td>
</tr>
</tbody>
</table>
## Voids

<table>
<thead>
<tr>
<th>Field</th>
<th>Field description</th>
<th>Requested format for data supply</th>
<th>Notes/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>property_id</td>
<td>The unique id that references a particular property.</td>
<td>Character field with no special characters or white space</td>
<td>Your standard property reference numbers.</td>
</tr>
<tr>
<td>void_start_date</td>
<td>The date the void period started.</td>
<td>Date field with format &quot;yyyy-mm-dd&quot;</td>
<td></td>
</tr>
<tr>
<td>void_end_date</td>
<td>The date the void period ended.</td>
<td>Date field with format &quot;yyyy-mm-dd&quot;</td>
<td>Leave unpopulated where the property is still void.</td>
</tr>
<tr>
<td>reason_for_void</td>
<td>The reason for the property void. Categorical: <strong>Eviction</strong> the previous tenancy was evicted <strong>Other</strong> the previous tenancy left for a reason other than eviction</td>
<td>Categorical field: select one from:{Eviction, Other}</td>
<td></td>
</tr>
<tr>
<td>Field</td>
<td>Field description</td>
<td>Requested format for data supply</td>
<td>Notes/Comments</td>
</tr>
<tr>
<td>------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>last_tenancy_ref</td>
<td>The reference number of the tenancy immediately before the void period</td>
<td>Character field with no special characters or white space</td>
<td>Your standard tenancy reference numbers.</td>
</tr>
<tr>
<td>next_tenancy_ref</td>
<td>The reference number of the tenancy immediately after the void period</td>
<td>Character field with no special characters or white space</td>
<td>Your standard tenancy reference numbers. Leave unpopulated if the property is currently uninhabited.</td>
</tr>
</tbody>
</table>
Appendix C  Cost specification

QA Housing Consulting (QAHC) was responsible for acquiring and processing costs data from the organisations involved in this project. This appendix outlines the approach QAHC used.

**Non-pay costs** – Organisations provided cost codes as per their accounting systems with a description and actual cost for up to 5 years’ financial year data. These data were then mapped according to the QAHC efficiency tool benchmarking categories. These categories map the data into the following key areas of delivery:

- Major repairs
- Planned repairs
- Routine repairs
- Rent arrears
- Resident involvement
- ASB
- Tenancy management
- Lettings
- Care and support
- Estate services
- Home ownership
- Overheads - office
- Overheads - finance
- Overheads - IT
- Overheads - central

**Employee costs** – These data were provided by the organisations based upon information they had submitted for annual benchmarking for up to 5 years. The data were converted into the categories above.

From the above data, QAHC identified the areas in which it was possible to split the data to produce a per-tenancy assessment of costs:

- **Responsive repairs** – Costs were provided on a per-tenancy basis for responsive repairs. This data was provided directly from each organisation’s housing management IT system. For some organisations this was not possible; in these cases, we generated an average cost per repair from the number of responsive repairs divided by total responsive repairs cost. This was then multiplied per tenancy depending on the number of repairs.
Rent arrears – Organisations were not able to provide a specific cost of rent arrears collection and recovery per tenancy. However, from the cost data above for rent arrears pay and non-pay costs it was possible to establish the total for the financial year. Organisations were also able to provide the total number of arrears incidents, and consequently we were able to establish an average cost per incident. For each tenancy this was multiplied by the number of rent arrears actions, to create an estimate of the per-tenancy expenditure on arrears collection and recovery.

ASB – Costs per tenancy were estimated in a similar way to the arrears costs, by dividing the total cost of ASB by the number of cases. For ASB we requested that, if possible, we would be interested in data on both incidences where the tenant was recorded as the perpetrator of ASB and where the tenant was the person reporting an incident. In most cases, the organisations were able to provide details on incidents attributed to suspected perpetrators only.

The accuracy of the data obtained was supported by QAHC’s mapping processes and by using data that had previously been validated for benchmarking purposes.

Below we summarise the data collected per organisation.

Housing Provider 1:

- 3 years’ employee and non-pay data from 2012/13, 2013/14 and 2014/15.
- 2 years’ repairs per tenancy from the housing management system. The organisation provided the information by property reference and a separate extract was used to match property IDs to tenancies.
- Arrears actions and cost for 3 years from 2012/13 to 2014/15.
- 2 years’ ASB data from 2013/14 to 2014/15.

Housing Provider 2:

- 3 years’ employee and non-pay data. 2012/13, 2013/14 and 2014/15 financial years.
- The organisation was not able to provide a direct cost per tenancy for responsive repairs. It was, however, able to provide the total number of repairs per tenancy for 2012/13, 2013/14 and 2014/15. The average cost per repair, and the cost per tenancy were calculated as described above.
- 2 years’ cost per rent arrears data. 2013/14 to 2014/15
- 3 years’ data relating to incidents in which tenants were recorded as being the perpetrator of ASB for 2012/13, 2013/14 to 2014/15

Housing Provider 3:

5 years’ repairs cost per tenancy. Data from housing management system directly for the same years as above.

5 years rent arrears actions per tenancy based on the above years.

2 Years cost per tenancy for ASB reported by the tenant and the tenant as the perpetrator. The years used were 2014 and 2015 calendar years.

**Housing Provider 5:**

4 years’ employee and non-pay data for years 2011/12, 2012/13, 2013/14 and 2014/15.

5 years’ responsive repairs cost per tenancy from the housing management system, from 2010/11 to 2014/15.

2 years’ rent arrears data per tenancy from 2013/14 to 2014/15. This was provided via tenancy reference numbers.

**Housing Provider 6:**

3 years’ non-pay and employee mapping from 2012/13 to 2014/15

4 years’ repairs per tenancy directly from the housing management system from 2010/11 to 2014/15.

Arrears actions per tenancy for 3 years from 2012/13 to 2014/15.

ASB costs per tenancy for incidences where the tenant was recorded as the perpetrator for 3 years from 2012/13 to 2014/15.
Appendix D  Analysis

Propensity score methods

The primary analysis method used within this project was a propensity score method. Propensity scores are used to enable an analysis to match comparison groups on a like-for-like basis; in this case the analysis is intended to compare those in a particular community investment status with those who have a different status.

Naïve comparisons of two groups can result in identifying spurious associations, in particular where both the status (e.g. in work or out of work) and the outcome measure (costs) are independently related to some third factor. The aim of our analysis is to rule out as many of those sorts of associations as possible, to enable us to make like-for-like comparisons between tenancies in different categories. One approach to attempt this is exact matching: for each tenant in one of the categories a tenant is identified in the other category who shares all of the attributes that might be relevant. This approach breaks down where there are several potentially relevant attributes that you wish to match upon, as it quickly becomes impossible to identify sufficient matches.

Propensity score matching improves upon exact matching by matching upon a single value (the propensity score) that summarises all of the potentially relevant attributes. This means that, as long as there is a reasonable degree of overlap in the propensity scores of the two groups (known as ‘common support’) it will be possible to match all of the tenants.

A propensity score is a conditional probability of assignment to a treatment condition given a set of observed characteristics. As an example, the propensity score is the probability of a tenant being employed. The propensity score for each tenant is estimated by fitting a model on a selection of predetermined characteristics, such as age and gender. A generalised boosted model was used for this purpose. Using iterative algorithms, multiple simple decision tree models are combined to create a complex model. The advantage of using this approach over others is that it will

---

automatically include nonlinear interactions to create a better model. By running the data through the final model we get a propensity score for each tenancy.

The method adopted in this project was a variant that uses the propensity scores as weighting factors rather than for matching. This effectively means that each tenancy is used within the assessment of the relevant costs for its category, weighted on how similar it is to the ‘typical’ tenancy in the category of interest. By using weights, we are able to include the data for all tenants’ in the analysis of the outcome. The calculation for weight estimates are as follows:

- For the treatment group (e.g. Digitally included):
  \[ w = 1 / \text{propensity score} \]

- For the comparison groups (e.g. not digitally included):
  \[ w = 1 / (1 - \text{propensity score}) \]

The propensity scoring was conducted using the ‘twang’ R package,\(^7\) developed by RAND Corporation. The methodology and applications are published in peer-reviewed academic journals, and the package is openly available on the Comprehensive R Archive Network under a GPL-3 license. The package works within the R statistical programming language.\(^8\)

One of the advantages of the twang implementation of propensity scoring is that it works with multiple categories for comparison. Many variants of propensity scoring only work with two categories (e.g. treated vs. untreated). As this project included circumstances where we needed to compare three categories (full housing benefit, partial housing benefit, no housing benefit) one of our requirements in selecting a propensity scoring approach was that it had to be able to cope with more than two states.

Having calculated propensity scores for each tenancy, the quality of the propensity scores and weights was assessed. This assessment was conducted to confirm that there was balance between the groups (i.e., that they were sufficiently similar across

---


the measured factors) and that there was common support between the range of propensity scores.

**Figure 15** give an example of an assessment of common support, showing the distribution of propensity scores for digital inclusion for one housing provider. The figure shows that there is reasonable overlap between the digitally included and not digitally included groups on their propensity scores.

**Figure 16** gives an example of a set of charts examining the degree of balance between two categories, before (left-hand bars) and after (right-hand bars) the use of propensity scores. The check for balance compares the means for each of the characteristics for each of the groups of tenancies. If perfect balance were achieved, the means for each characteristic would be the same for each group. This is unlikely to occur exactly; consequently what we want to see is a reasonable degree of improvement in the closeness of the means from the original data and for the means to not be significantly different. As the graphs show, after the use of propensity scores the distance between the mean values of the treatment and comparison groups for every characteristic have substantially been reduced. Additionally, all characteristics have means that are not significantly different after the use of propensity scores, establishing that the use of propensity scores has created reasonable balance across groups.
Figure 15 Distribution of propensity scores for digital inclusion for one housing provider
Figure 16 Example of charts to check balance for digital inclusion for one housing provider, showing imbalance based on naïve comparisons (left-hand bars) and improved balance when using propensity score-weighted comparisons (right-hand bars)

Once the propensity scoring was completed and evaluated, and weights established, the ‘survey’ R package\(^9\) was used to perform weighted regression calculations to obtain the estimates of the costs in each category. This calculation takes into account the propensity scores determined in the previous step. As well as producing the central estimate for each cost in each group, these calculations also produce confidence

---

intervals. The confidence interval is the range of values either side of the central estimate in which the actual value is relatively likely to sit.

Exploratory analysis

As well as the propensity score methods we also conducted some exploratory analyses to see whether further insights could be generated from the data using different types of analysis. The explorations focused on attempting to generate a tenancy-level predictive model that could estimate an expected level of a given cost for each tenancy.

These explorations used a class of analyses called regression models. We started from the simplest form of regression model (linear regression), and tried increasingly sophisticated analytical approaches to see whether it was possible to create an accurate predictive model. These attempted enhancements included:

- Using logarithmic transformations on some of the data to enable the linear model to better cope with exponential relationships;
- Adopting a ‘hurdle’ approach, splitting the data between those with zero costs in an area and those with some costs, since zero-bounded data can cause problems with regressions;
- Using a complex tree-based ensemble model (random forest), since these typically outperform linear models;
- Using non-linear (random forest and neural network) models for identifying classification boundaries in the data.

In spite of these increasingly sophisticated methods we found that we were not able to produce a model with sufficient predictive power to create estimates that were much more accurate than chance. The most likely explanation for this is that the data did not contain sufficient factors to be able to predict the expected level of expenditure for a given tenancy at an individual level. This contrasts with the propensity score method, which works on an aggregate basis, and hence is able to produce useful results with somewhat less information.
Appendix E  Tables of results

The tables below provide the numerical results, as presented in the charts within the body of the report.

### Employment status results

#### Responsive repairs

<table>
<thead>
<tr>
<th></th>
<th>Baseline (full HB)</th>
<th>std error</th>
<th>No HB</th>
<th>std error</th>
<th>Partial HB</th>
<th>std error</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP1</td>
<td>66.343</td>
<td>6.41</td>
<td>-22.335</td>
<td>7.223</td>
<td>-5.896</td>
<td>7.911</td>
</tr>
<tr>
<td>HP2</td>
<td>1049.26</td>
<td>28.75</td>
<td>-342.17</td>
<td>30.44</td>
<td>-132.31</td>
<td>33.12</td>
</tr>
<tr>
<td>HP3</td>
<td>332.73</td>
<td>14.14</td>
<td>-53.34</td>
<td>16.22</td>
<td>-33.5</td>
<td>18.63</td>
</tr>
<tr>
<td>HP4</td>
<td>180.592</td>
<td>6.992</td>
<td>-49.765</td>
<td>8.155</td>
<td>-36.207</td>
<td>8.933</td>
</tr>
<tr>
<td>HP5</td>
<td>112.16</td>
<td>11.17</td>
<td>-29.06</td>
<td>13.48</td>
<td>-33.58</td>
<td>14.73</td>
</tr>
<tr>
<td>HP6</td>
<td>332.96</td>
<td>24.41</td>
<td>-102.66</td>
<td>33.23</td>
<td>-65.04</td>
<td>27.88</td>
</tr>
</tbody>
</table>

#### Actions pursuing rent arrears

<table>
<thead>
<tr>
<th></th>
<th>Baseline (full HB)</th>
<th>std error</th>
<th>No HB</th>
<th>std error</th>
<th>Partial HB</th>
<th>std error</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP1</td>
<td>90.264</td>
<td>2.775</td>
<td>-36.795</td>
<td>3.284</td>
<td>45.881</td>
<td>4.022</td>
</tr>
<tr>
<td>HP2</td>
<td>49.648</td>
<td>1.14</td>
<td>-2.086</td>
<td>1.245</td>
<td>2.57</td>
<td>1.338</td>
</tr>
<tr>
<td>HP3</td>
<td>52.271</td>
<td>7.336</td>
<td>28.418</td>
<td>8.701</td>
<td>64.425</td>
<td>10.938</td>
</tr>
<tr>
<td>HP4</td>
<td>49.131</td>
<td>2.083</td>
<td>48.699</td>
<td>2.983</td>
<td>43.019</td>
<td>3.239</td>
</tr>
<tr>
<td>HP5</td>
<td>84.884</td>
<td>3.35</td>
<td>11.817</td>
<td>4.577</td>
<td>51.739</td>
<td>6.071</td>
</tr>
<tr>
<td>HP6</td>
<td>91.809</td>
<td>2.165</td>
<td>-21.642</td>
<td>3.015</td>
<td>50.899</td>
<td>3.054</td>
</tr>
</tbody>
</table>

#### Anti-social behaviour (ASB)

<table>
<thead>
<tr>
<th></th>
<th>Baseline (full HB)</th>
<th>std error</th>
<th>No HB</th>
<th>std error</th>
<th>Partial HB</th>
<th>std error</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP1</td>
<td>15.096</td>
<td>1.097</td>
<td>-7.119</td>
<td>1.19</td>
<td>-3.956</td>
<td>1.435</td>
</tr>
<tr>
<td>HP2</td>
<td>72.635</td>
<td>6.831</td>
<td>-32.078</td>
<td>-4.47</td>
<td>-10.973</td>
<td>7.966</td>
</tr>
<tr>
<td>HP3</td>
<td>No data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HP4</td>
<td>No data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HP5</td>
<td>No data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Community Investment and the Bottom Line

74
# Digital inclusion status results

## Responsive repairs

<table>
<thead>
<tr>
<th></th>
<th>Baseline (no email address held)</th>
<th>std error</th>
<th>email (holds email address)</th>
<th>std error</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP1</td>
<td>53.441</td>
<td>2.772</td>
<td>5.55</td>
<td>4.597</td>
</tr>
<tr>
<td>HP2</td>
<td>739.899</td>
<td>8.058</td>
<td>176.343</td>
<td>15.592</td>
</tr>
<tr>
<td>HP3</td>
<td>297.305</td>
<td>7.469</td>
<td>-10.19</td>
<td>12.696</td>
</tr>
<tr>
<td>HP4</td>
<td>132.115</td>
<td>3.456</td>
<td>14.848</td>
<td>5.895</td>
</tr>
<tr>
<td>HP5</td>
<td>92.681</td>
<td>6.278</td>
<td>-9.186</td>
<td>11.719</td>
</tr>
<tr>
<td>HP6</td>
<td>279.781</td>
<td>8.219</td>
<td>14.47</td>
<td>27.672</td>
</tr>
</tbody>
</table>

## Actions pursuing rent arrears

<table>
<thead>
<tr>
<th></th>
<th>Baseline (no email address held)</th>
<th>std error</th>
<th>email (holds email address)</th>
<th>std error</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP1</td>
<td>63.621</td>
<td>1.12</td>
<td>4.725</td>
<td>2.13</td>
</tr>
<tr>
<td>HP2</td>
<td>47.6768</td>
<td>0.4234</td>
<td>0.8767</td>
<td>0.71</td>
</tr>
<tr>
<td>HP3</td>
<td>84.373</td>
<td>4.138</td>
<td>1.8</td>
<td>7.472</td>
</tr>
<tr>
<td>HP4</td>
<td>57.596</td>
<td>1.419</td>
<td>22.569</td>
<td>2.5</td>
</tr>
<tr>
<td>HP5</td>
<td>95.829</td>
<td>2.307</td>
<td>7.883</td>
<td>4.502</td>
</tr>
<tr>
<td>HP6</td>
<td>78.152</td>
<td>1.017</td>
<td>2.4</td>
<td>2.839</td>
</tr>
</tbody>
</table>

## Anti-social behaviour (ASB)

<table>
<thead>
<tr>
<th></th>
<th>Baseline (no email address held)</th>
<th>std error</th>
<th>email (holds email address)</th>
<th>std error</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP1</td>
<td>8.7291</td>
<td>0.4407</td>
<td>-0.3952</td>
<td>0.9317</td>
</tr>
<tr>
<td>HP2</td>
<td>38.07</td>
<td>1.673</td>
<td>17.931</td>
<td>3.09</td>
</tr>
<tr>
<td>HP3</td>
<td>No data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HP4</td>
<td>No data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HP5</td>
<td>No data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HP6</td>
<td>31.768</td>
<td>2.006</td>
<td>19.413</td>
<td>11.143</td>
</tr>
</tbody>
</table>
Appendix F  Exploratory analysis: trend analysis of community investment activity

For some of the community investment activity data, we were provided with details of the dates on which each tenant participated in the activity. This allowed us to try an exploratory analysis to attempt to see an effect on service usage of participating in an activity by comparing those who participated earlier in the year rather than later.

This analysis was designed to take advantage of any potential difference in service usage given the fact that the costs data we held was on an annual basis. The hypothesis that we were testing was that tenants who participated in a community investment activity tended to subsequently make lower use of a service. We were trying to detect whether those who participated in the activity earlier in the year made detectably lower use of the service over the year than those who participated in the activity later. This possibility is based on the fact that they spend more of the year in the ‘post activity’ state, which is assumed to be a lower service usage state.

This analysis also takes advantage of the fact that people participating in the activity throughout the year may be assumed to be roughly similar in terms of their characteristics. If this assumption holds, the analysis will be comparing like-with-like: the people who received the activity at the end of the year are a suitable comparison group for those who received it at the start of the year.

The figure illustrates the proportion of the year that a tenant would spend in the pre-intervention state, dependent on the date he/she received the intervention. If the pre-intervention state were associated with higher service usage, then the increasing proportion of the year in this state would be associated with higher service usage over the course of the year on average.
It should be noted that in some cases the effect of an intervention may be delayed until some time after the intervention itself. For example, for an employment intervention that is designed to equip unemployed people with the skills needed to find work, there could be a delay before there is an effect on employment status as people have to go through processes like job applications and interviews before they actually start work. The figures below illustrate what would happen in scenarios where there are delays in the effect of certain time periods. As can be seen, with longer delays there is less of a difference available to be observed over the course of the year.
These exploratory analyses did not identify any clear trends. This could be caused by the effect size being too small to be detected using this method, or because the effect lags after the intervention by a significant proportion of a year. Further investigation on the basis of a longer time period and more tenants would potentially identify smaller effects or those with a longer lag.

If a further study were using this method as a primary investigation tool, rather than for exploration, it would be advisable to specifically test the assumption that the makeup of the participant group is consistent throughout the period. This could include examining the baseline balance of various attributes of tenants entering the programme at different times.
Appendix G  Data processing

The data received from housing providers were processed to clean invalid values, filtering missing values, and make other conversions to include relevant variables within the analysis. The primary processing details are described below:

- **Person data:**
  - Where a date of birth was provided this was processed to calculate an age value.

- **Asset data:**
  - Where the construction date was provided this was processed to calculate an asset age value;
  - Postcode values were used to identify the Lower Super Output Area (LSOA) that the property is located in, and this in turn was used to to neighbourhood data statistics.

- **Tenancy data:**
  - The tenancy duration was calculated from records of the tenancy start date.

- **Household data:**
  - A value for household size was calculated from the number of person IDs that were associated with each tenancy
  - Tenancies were categorised as Joint, SingleF, or SingleM, depending on the composition of the household identified in the data;
  - For joint tenancies, the ‘age’ value associated with the tenancy was treated as the age of the eldest of the joint tenants;
  - For joint tenancies, the digital inclusion status was determine to be “yes” if an email address is held for at least one of the joint tenants.

- **Filtering:**
  - For employment investigations, only those with an age value under 65 were included.
  - When conducting analysis on the basis of costs data for a particular year, only those tenancies that were ongoing throughout the whole of the year were included.

The processed dataset used contained the following fields:

- `property_id`
- `tenancy_ref`
- `lead_tenant_type` (joint, singleM, singleF, singleNK)
Community Investment and the Bottom Line

The following factors were used in generating propensity scores:

- lead_tenant_type (joint, singleM, singleF, singleNK)
- age
- household_size
- tenancy_duration (years)
- property_type (house, flat, bungalow, maisonette, other)
- asset_age (years)
- number_of_bedrooms
- weekly_rent
- email_address (Y/N)
- housing_benefit_level (full, partial, none)
- is_current (Y/N)
- tenancy_start_date
- tenancy_end_date
- postcode
- pct_unempl (from neighbourhood statistics)

The scripts used to process the data (which do not contain any data in themselves) will be retained, and are available upon request.
Appendix H  Chart design

In many of the cases the results included findings where some of the estimates were in one direction and some of them in the opposite direction (i.e. the group of interest being higher than the baseline group in some cases and lower in others). We wanted to create a chart design that would enable these to be presented in a manner that was easy to understand and that did not create an artificially skewed impression of the findings.

One simple form of chart to show the relative differences from baseline would be to show the percentage that the costs for the group of interest were lower than the comparison group, for example, the percentage that the costs for those on no housing benefit are lower than for those on full housing benefit. This can work well where all of the cost estimates are in the same direction, but produces results that could be considered skewed where they are in opposite directions.

The percentage that costs were lower, in this case, would be calculated using the following formula:

\[
\text{percentage that no HB costs are lower} = \left( \frac{\text{costs}_{\text{full HB}} - \text{costs}_{\text{no HB}}}{\text{costs}_{\text{full HB}}} \right) \times 100
\]

Where the 'no HB' costs are lower, this produces consistent results:

<table>
<thead>
<tr>
<th>costs_{full HB}</th>
<th>costs_{no HB}</th>
<th>percentage that no HB costs are lower</th>
</tr>
</thead>
<tbody>
<tr>
<td>£1000</td>
<td>£900</td>
<td>10%</td>
</tr>
<tr>
<td>£1000</td>
<td>£800</td>
<td>20%</td>
</tr>
<tr>
<td>£500</td>
<td>£400</td>
<td>20%</td>
</tr>
<tr>
<td>£500</td>
<td>£250</td>
<td>50%</td>
</tr>
</tbody>
</table>

However, where the results are sometimes higher and sometimes lower, the results could be perceived as being skewed:

<table>
<thead>
<tr>
<th>costs_{full HB}</th>
<th>costs_{no HB}</th>
<th>percentage that no HB costs are lower</th>
</tr>
</thead>
<tbody>
<tr>
<td>£1000</td>
<td>£900</td>
<td>10%</td>
</tr>
<tr>
<td>£900</td>
<td>£1000</td>
<td>-11.11%</td>
</tr>
<tr>
<td>£1000</td>
<td>£800</td>
<td>20%</td>
</tr>
<tr>
<td>£800</td>
<td>£1000</td>
<td>-25%</td>
</tr>
<tr>
<td>£1000</td>
<td>£500</td>
<td>50%</td>
</tr>
<tr>
<td>£500</td>
<td>£1000</td>
<td>-100%</td>
</tr>
</tbody>
</table>
Community Investment and the Bottom Line

As the table above shows, differences that are equal in size but in the opposite direction result in different percentage differences when calculated in this fashion. At the extreme, if one group had some costs (any number above £0) and the other costs approaching £0, the maximum percentage lower in one direction would tend towards 100%, whilst on the other it would tend towards minus infinity.

Even at more likely levels, the chart below illustrates a circumstance where the values below zero are much further from zero than the equivalent differences above zero.

An alternative presentation that we considered was the use of a logarithmic scale (known as a log-plot). One example of this is shown in the version below, which presents each point as a proportion of the baseline value. In this case, points above the axis represent higher costs (whereas in the previous chart we chose to show lower costs (which would equate to savings if these were to be achieved in practice) above the axis. Each value is calculated as follows:

\[
\text{costs as a proportion of full HB costs} = \frac{\text{costs}_{\text{no HB}}}{\text{costs}_{\text{full HB}}}
\]

These calculated proportions still appear skewed when viewed as values:

<table>
<thead>
<tr>
<th>costs_{full HB}</th>
<th>costs_{no HB}</th>
<th>costs as a proportion of full HB costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>£1000</td>
<td>£900</td>
<td>0.9</td>
</tr>
<tr>
<td>£900</td>
<td>£1000</td>
<td>1.111111</td>
</tr>
<tr>
<td>£1000</td>
<td>£800</td>
<td>0.8</td>
</tr>
<tr>
<td>£800</td>
<td>£1000</td>
<td>1.25</td>
</tr>
</tbody>
</table>
However, the logarithmic scale means that, for example, 0.5 and 2.0 will sit equal distances either side of a baseline at 1.0.

The use of a log plot results in a visual scaling that avoids the problems of results appearing skewed on one side. However, outside of certain technical audiences logarithmic plots are unfamiliar to many people, and the irregular spacing of the values on the y axis may be disconcerting. The other major difference in this presentation is that the “no difference” point is represented by 1 rather than 0. For future analyses it would be worth testing with audiences whether these features of log plot charts make them unusably difficult to interpret or whether audience are able to easily acclimatise to their way of presenting the results.
An alternative solution to the potential asymmetry created by expressing all values as savings relative to a baseline is to present the extent to which the lower value is lower than the higher value, either way round; values of above or below the x axis can still be used to indicate which way the relationship is.

\[
\text{percentage by which lower costs are lower} = \left(\frac{\text{costs}_{\text{full HB}} - \text{costs}_{\text{no HB}}}{\max\{\text{costs}_{\text{full HB}}, \text{costs}_{\text{no HB}}\}}\right) \times 100
\]

<table>
<thead>
<tr>
<th>costs_{\text{full HB}}</th>
<th>costs_{\text{no HB}}</th>
<th>percentage by which lower costs are lower</th>
</tr>
</thead>
<tbody>
<tr>
<td>£1000</td>
<td>£900</td>
<td>10%</td>
</tr>
<tr>
<td>£900</td>
<td>£1000</td>
<td>-10%</td>
</tr>
<tr>
<td>£1000</td>
<td>£800</td>
<td>20%</td>
</tr>
<tr>
<td>£800</td>
<td>£1000</td>
<td>-20%</td>
</tr>
<tr>
<td>£1000</td>
<td>£500</td>
<td>50%</td>
</tr>
<tr>
<td>£500</td>
<td>£1000</td>
<td>-50%</td>
</tr>
</tbody>
</table>

This results in values that are symmetrical irrespective of which costs are higher.
In order to ensure the charts are unambiguous, the visualisation was further developed to make it easier to identify the implications of values above and below the x axis.
References


