# Memorandum

21 October 2016

From: Tom Love
Re: Social Housing Test Case: Technical Report

I would like to set out my peer review of the draft *Social Housing Test Case Technical Report*. In my discussion I will set out my views on the strengths and weaknesses of the technical analysis, and make some remarks upon the interpretation of the analysis. I will conclude with some suggestions for refining this, and future, analyses.

In carrying out this peer review I met with and interviewed the analytical team, and observed work in progress, in August 2016. I have been provided with a draft report which I have considered in detail.

I understand that this work is, as the name suggests, a test case. This applies in the context of the SIU's explicit approach of working iteratively, and sharing material at an early stage. As such, the methodology and analytical approach are being newly developed, and the technical report is not intended to be the last word – but rather an incremental step forward in applying social investment analysis.

## Goals

The stated goals of the test case are to:

- Understand whether it is possible to calculate a fiscal return on investment for a given investment within the social sector;
- If so, develop the methodology in a reusable manner that allows the analysis and associated data set to be used in subsequent analyses;
- Understand the limitations associated with the particular methodology;
- Generate insight into the impact of social housing on broader social sector outcomes, in time for use in Budget 17

The test case is therefore intended to be a learning exercise in developing a methodology, rather than a definitive statement of the return on investment for social housing.

## Strengths and weaknesses

Overall there are three fundamental elements to the test case: the data manipulation involved in defining a cohort and extracting and coding the relevant variables; the statistical analysis involved in conducting a propensity matching analysis; and the application and analysis of patterns of investment and cost.

### Data manipulation

The test case describes much of the process of data manipulation in detail, particularly the development of the Social Investment Analytical Layer, as an attempt to impose a systematic framework over a diverse range of datasets with different sources and structures. This is an important contribution to the use of

these datasets, and indicates a systematic approach to good analytical practice in terms of trying to assess and describe datasets, and make them tractable, before jumping to more complex analyses. The description of standard table structures, and standardised naming conventions is unglamorous but important underlying work which is necessary to support good analysis. This systematic approach is also reflected in the decision log, reported in the appendices.

While the consistent manipulation of the data is well described, the completeness of the datasets is less explicit in the report. For example, section 2.2.1 lists events which have been rolled up during the profile window, but the completeness of these datasets, or how they apply to the population, is not clearly described, nor are the datasets defined in detail at this point. To a reader who is not directly familiar with the data, it is not clear what a student intervention is, and whether there are issues in the data for that variable. Even the table in section 4.2.1 doesn't elaborate upon what a student intervention actually means. Similarly, I understand that there are serious issues of incompleteness in the national laboratory dataset, which don't appear to be recorded.

While HNZ application variables, applicant variables and household variables are described in the appendices, and while there are some variable transformation rules reported in a further appendix, some of the other variables are covered. Nor is there any detail about data completeness or other issues in the variable descriptions. This is useful information which would help the reader to assess the robustness of the underlying data.

The other issue of completeness is the effectiveness of the datamatch within the IDI – the proportion of people who have been accurately matched from the IDI spine to the other datasets, particularly the core housing dataset. The completeness of this is unclear, and it would be helpful to be more explicit about this. There are hints, for example section 7.6.5 notes that 4.2% of records lacked an address UID or TA variable, but it would be good for this issue to be addressed directly. At a level of 4.2% this is not likely to be a serious problem for the overall conclusion of the analysis, unless there is a specifically identified bias, but it is important to be clear on whether this is a significant issue or not, and whether it has the potential to bias results, or to undermine generalisability.

The cohort definition is well described, and the diagram in section 2.2.1 sets out the overall approach clearly. The methodological issue which arises from this is the absence of information used about events between applying for, and receiving, a house. While section 1.5 suggests that most are housed within one year, and presumably that most are housed relatively quickly and with a long tail of duration, it would again be good to be explicit on this issue. It seems likely that households will still be seeking alternative accommodation while waiting , and there may be selection effects which remove particular kinds of households, or make them wait longer. Again, this issue may not be material to the overall conclusions of the analysis, but it is helpful to explore whether there is an affect or not.

A broader issue in the data manipulation is the complex relationship between household and individual perspectives. The analysis explicitly takes the perspective of households, particularly in the use of household characteristics for the purposes of the propensity scoring. But some elements of the household definition are aggregate individual characteristics, and in future analyses should probably be adjusted to reflect comparable household characteristics. For example, equivalised income and bedroom size would probably be more useful than raw values, and would deal with the (expected) correlations between variables noted in section 4.2.2. Equivalised income and bedroom size is a relatively standard method of dealing with these issues as used for, example, in the New Zealand Deprivation Index. This issue could be noted as a limitation in the current analysis, and considered in more detail for future work.

**Propensity scoring**

The description of methods for the propensity scoring is very complete. This excellent description gives a good level of detail for a reader to assess specifically what choices were made and why. The description of the pros and cons of the different approaches for the scoring is well made, and the choice of a gradient boosted tree seem to be clearly explained. The benefit of avoiding the complexity of interactions in a regression model seems a worthwhile rationale in a model where there are so many variables, and the

consistency of the results presented in the table in section 5.2 gives confidence in the stability of the model.

Overall, the important question about the propensity scoring is whether the matching is good enough, or whether material biases remain in the information. There are several factors which influence this, including:

- The variables used to calculate the scores, and how well these address underlying bias;
- Whether the inverse probability treatment weighting has effectively reduced any remaining bias;

A range of variables have been used to reduce bias. The important issue is whether they are likely to influence simultaneously the treatment decision (being housed) and the outcomes of interest (in this case the various social costs in the ROI). The variables chosen for this analysis appear to have been selected opportunistically, on the basis of information which may have relatively quickly accessible within the IDI and which reflect various socioeconomic factors of the households. Subject matter experts and front line staff were consulted to identify areas in which lack of information could introduce bias, which considered that the key drivers were of the housing decision were:

1. The total scores from the housing application process;
2. The presence of rheumatic fever;
3. The relatively opaque process of matching clients to houses on the part of individual housing providers;
4. The available social housing stock.

Of these, the third seems to be an important limitation, in that unobserved factors such as smoking status or gang affiliation could affect likelihood of housing. But the report correctly notes that the direction of this bias is likely to be to reduce the observed benefit of housing in future cost of corrections and justice services. However there may be quite a number of idiosyncratic effects at work here, which present a limitation to the analysis.

The fourth point is also an important limitation. The test case suggests that it is unclear whether this will introduce bias, but it seems to me that a little more might be said about this. I think it is likely that biases arising from this issue would show up as regional effects, since housing stock is essentially a regional variable. It should also be noted that housing markets could also have an impact, and that these are also likely to show up as biases in the regional variable. In light of this it is interesting to note that the later segment results in section 8.3 show consistent effects for both Christchurch and Wellington, and this suggests that there may be some consistent bias in the regional results. At one level this might be an interesting finding in itself (which isn't discussed as such in the current draft), but I think this bias is worth more unpacking and discussion. In general I think that contextual market effects could be important in deciding who gets housed, and are worth more attention in the variable selection, analysis and discussion.

Clearly there are opportunities to try to match on other variables within the IDI (although time constraints may have limited that in this particular test case), but equally, propensity score matching does not gain any value from including extraneous variables, and these have the potential to increase the variances in the final model (although extraneous variables will not introduce fresh bias). In general, for future analyses, a more systematic approach to variable selection in the context of the treatment and outcomes of interest would help to ensure that models are as robust as can be managed, without being overspecified.

The inverse probability treatment weighting is, again, well described. It might be worth plotting the pre and post IPTW differences on a single graph, rather than using the series of the box plots presented (for example figure two in Austin and Stuart[1]).

## Costs and benefits

The return on investment calculation relies upon a range of direct costs of government services. The cost values range for specific costs of particular services which are observed directly at the level of the individual, whereas some costs, such as hospital outpatients or education interventions are derived (section7.3.5). The report notes that with additional time it might have been possible to improve accurately attributed costs for the ROI calculation. This is a limitation (which is noted in the caveats at the front of the report), but it is an important area for further development in this kind of analysis. Cost data in general appear to require considerable further systematic data cleaning and manipulation in order to arrive at improved ROI estimates.

Section 8.1 notes the important point that some fiscal spend can improve fiscal, economic and social outcomes. By contrast, other forms of expenditure are signs of poor outcomes. The obvious areas where expenditure is likely to represent a positive investment in itself are education, and some elements (but not all) of healthcare. By contrast, expenditure on corrections, CYF services, or on avoidable health care (such as ambulatory sensitive hospitalisations) are likely to represent government money spent upon adverse social outcomes. It is noted in section 8.2 that increased expenditure on social housing appears to increase the cost of education, because children stay in education longer when socially housed.

These cost results make sense in terms of face validity, but suggest that the ROI calculation requires further theorising and refinement. If social investment in one sector (housing) actually results in increasing positive social investment in another sector (education), then there are multiplier effects in play. This has the potential to increase the true ROI from the first intervention (although there are clearly complex causal interrelationships between the different services). But ultimately, these effects have the potential to increase the estimated positive ROI, if well applied fiscal expenditure has the impact of improving the effectiveness of other kinds of fiscal expenditure. This means that reporting a single overall ROI of -$0.10 is problematic. I suggest that, while the components of the ROI appear to have reasonable face validity on their own, the costing approach cannot yet support a single, overall ROI figure.

These issues are complex, and are not likely to be resolved satisfactorily in the context of a single analytical test case. But they should, in my view, be the subject of more discussion than the brief mention they receive in the draft, since they are fundamental to the overall conceptualisation and calculation of social investment. This is an important issue to explore for future test cases, and is likely to become increasingly important as this kind of analysis is applied to different sectors. Complex causal questions arise: for example, in the converse case to the current example, does housing expenditure increase the effectiveness of education expenditure?

## Presentation and discussion

In general the draft supplied requires a number of proofing and minor drafting corrections, which I have not commented on in this review. More broadly, it would be helpful to have some list or reference of the meaning of the variables at an earlier stage, since the descriptive statistics reported in section 1.5.2 cover a number of variables which the reader may not be familiar with (such as the sustainability score), and leave the reader searching to find their meaning. So more discussion and contextualisation at an earlier stage would make some of the data presentation more accessible to the reader.

---

[1] Austin, Peter C., and Elizabeth A. Stuart. "Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies." *Statistics in medicine* 34, no. 28 (2015): 3661-3679.

The caveats and assumptions supplied at the beginning of the document are a very good thing to present so directly. The existing list has elements of caveat and assumption, but also more indirect explanation and statements (such as "findings are preliminary"). This section could benefit from a more structured approach, and perhaps does not need to be at the very front of the document, but at the end of the initial background section, or as an appendix.

The presentation of segment results in section 8.3 is nicely templated, but could do with some form of summary table, setting out the results for the different segments in a way which can more easily be compared. The issue of what the segment results mean, and whether they show real effects or may the consequence of artefacts could do with further discussion. This is important, since the relevance of the previously identified caveats for any particular finding is important in order to guide proportionate policy responses. Further, it seems to me that there are clearly important regional effects in play, given the consistent Wellington and Canterbury results, and that some discussion of regional markets and housing circumstances is important for the interpretation of these results.

In general, the discussion tends to focus upon the analysis itself rather than the results. While this is a test case, and it is explicitly stated that a large part of the aim is to establish a method and learn lessons for future analysis, it seems unfortunate note to discuss the substantive results and their implications in more depth. For example, only one paragraph in the executive summary appears to apply to the actual result of the analysis, which seems disproportionate.

More generally, while the report does well in trying to set out caveats and limitations in a technical sense, the real import of limitations is what they imply for what you can interpret and generalise from the analysis. In my view, it is therefore important that a technical report should not just document limitations, but give some consideration to the what those limitations mean for the data interpretation. This is done to some extent as the report progresses (eg noting the direction of potential bias in section 4.3), but it should be an important part of the final findings of the report.

## Overall comments

The goals of the social housing test case were:

1. Understand whether it is possible to calculate a fiscal return on investment for a given investment within the social sector;
2. If so, develop the methodology in a reusable manner that allows the analysis and associated data set to be used in subsequent analyses;
3. Understand the limitations associated with the particular methodology;
4. Generate insight into the impact of social housing on broader social sector outcomes, in time for use in Budget 17

Using these as a measure, I suggest that goal one has been partly achieved. The results are interesting and suggest that it is possible to calculate a fiscal return on investment but require, in my view, further thinking about the nature of investment and returns and how these will be calculated in complex social services environments.

Goal two has, in my view, been largely achieved. While this is an early test case, as its name implies, substantial work has been completed on data manipulation and methodology in way which can be reused and further developed in subsequent analyses. The methodology requires refinement, but will serve as a good basis for further work.

Goal three has been largely achieved. I think the limitations of the statistical components of the methodology have been well explored and that, while the implementation may necessarily be imperfect given the time limitations of the project and the exploratory engagement with many of the datasets, the technical issues appear to be understood.

Policy experts in housing will be able to provide better assessment of goal four than I can.