

# Memorandum

**To:** Darren Knox, Emma McCallum  
**Cc:** Chris Martin  
**From:** Linda Hutcheson  
**Date:** 19/03/2020  
**Subject:** SHS Housing Costs Imputation

## Introduction

The Scottish Household Survey collects data on how much households pay in housing costs (mortgage and rent). These have always been provided to the Scottish Government with no post-hoc processing, meaning missing values have not been imputed and unlikely outliers have not been edited<sup>1</sup>. However, this is now required so these costs can be used in the calculation of Total Net Household Income after Housing Costs.

This memo provides details of our suggested approach to implementing new processing routines to undertake this work. The analysis in this memo and the development of the imputation routines has been done using the 2018 dataset. Once these have been finalised, following feedback from the Scottish Government, they will be incorporated into the SHS data processing routines going forward, starting with the 2019 data. As with the imputation of income, the processing routines will impute amounts when missing, rather than whether the respondent pays these or not.

Appendix A1 contains an overview of the questions relevant to this work. To summarise, there is one question early in the household section that asks about tenure. This question is then used at various points in the questionnaire, including the section on mortgage and rent costs at the end of the household interview. There are two main strands of questions: one for those with mortgage costs and another for those with rent costs.

This work is split into two main stages:

- Designing the routines to clean the housing cost variables (and other key related variables).
- Designing the imputation routines for missing data.

The sections that follow deal with each in turn.

---

<sup>1</sup> From 2012 when the SHCS was incorporated into the SHS, and similarly beforehand on the SHCS.

## Data cleaning

In this section, we examine the consistency of responses within the household section relating to tenure and housing costs, before examining cases with outlier values for mortgage and rent amounts.

### Type of mortgage (HB509)

Early in the household section of the questionnaire, respondents are asked to describe their tenure (HB509): “In which of these ways do you occupy this accommodation?” There are five answer codes:

- 1) Buying with a mortgage/loan,
- 2) Own it outright,
- 3) Part rent/part mortgage
- 4) Rent (including rents paid by housing benefit)
- 5) Living here rent free.

Those who said they are in some sort of owner-occupation (codes 1, 2 or 3) are asked about mortgage payments. HJ1: “Can I just check, does your household have a mortgage or loan outstanding on this property?” It is from HJ1 that follow up questions on mortgage payments are routed, including level of monthly mortgage payments (HJ2).

There is a small discrepancy between how respondents answer at HB509 and HJ1. For example, respondents initially select buying with a mortgage (HB509) but later state they do not have an outstanding mortgage or loan on their property (HJ1). This is shown in Table 1 for 2018. This pattern is also evident in historical data (Appendix A2).

**Table 1: HB509 by HJ1 (2018)**

	Does your household have a mortgage or loan outstanding on this property?			
	Yes	No	Refused	Total
Buying with a mortgage/loan	2655	95	45	2795
Own it outright	76	3740	40	3856
Part rent/part mortgage	10	66	7	83
<i>Total</i>	2741	3901	92	6734

We suggest that in cases where there is a discrepancy, responses to whether they have a mortgage or loan outstanding take primacy and we have outlined a cleaning strategy for each group below. (We also suggest a small change to the script to minimise this discrepancy going forward – more below.)

1. In 2018, 95 respondents initially said they were buying their property with a mortgage/loan but later said they did not have an outstanding mortgage on their property. There was a similar number of cases between 2012 and 2017 (ranging from 72 to 109).
  - a. We suggest cleaning these by setting HB509 to owner occupied.

2. 76 respondents who initially said they own their property outright (HB509) later said they had a mortgage (HJ1). There were a similar number of cases between 2012 and 2017 (ranging from 85 to 138).
  - a. We suggest cleaning these based on responses to HJ2 (monthly mortgage payments). If the amount given equals £0, then set HJ1 and follow up questions to missing, otherwise change HB509 to buying with a mortgage.
3. Part rent/part mortgage. There has been a consistent high proportion (>75%) of part rent/part mortgage respondents who have later said they do not have a mortgage at HJ1. This is shown in Table 2. As well as being present in the data since 2012, the cases are spread across many interviewers and are therefore not suggestive of an interviewer effect. Given the small base size, and the consistent prevalence of missing data, it is recommended that we do not clean HB509 or HJ1 for these cases, and instead consider possible script checks going forward (see below).

**Table 2: Part rent/part mortgage response at HJ1 (2012-2018)**

Response at HJ1	2012	2013	2014	2015	2016	2017	2018
Yes	8.4%	9.1%	11.5%	11.6%	15.2%	20.4%	12.0%
No	89.6%	88.6%	81.5%	85.7%	80.8%	76.3%	79.5%
Refused	1.9%	2.3%	6.9%	2.7%	4.0%	3.2%	8.4%
Base:	154	132	130	112	99	93	83

Possible script amends to minimise the discrepancies above are:

1. Introduce a soft check after HJ1<sup>2</sup> for respondents who say:
  - a. Owner occupied at HB509 but mortgage at HJ1, and
  - b. Buying with a mortgage or part rent/part mortgage at HB509 but no mortgage at HJ1.
2. For respondents who are part rent/part mortgage, introduce an instruction at HJ1 and HJ3 (rent amount) that makes clear that this question is applicable to those who have a mortgage on a share of their home, not just the entire property.

The table below shows the impact of these changes in number of respondents. In terms of unweighted percentages, a difference of ten respondents is equivalent to approximately 0.1%.

**Table 3: Impact of suggested edits on tenure**

	2012	2013	2014	2015	2016	2017	2018
Buying with mortgage / loan	62	16	16	-25	-26	-24	-23
Own it outright	-62	-16	-16	25	26	24	23

<sup>2</sup> HJ1: "Can I just check, does your household have a mortgage or loan outstanding on this property?"

## Rent before and after housing benefit

There are two points in the questionnaire where respondents are asked about receipt of housing benefit. The first is in the housing costs section when respondents are asked for rent before and after housing benefit. The second is in the income section where respondents are asked about receipt of housing benefit. Whilst this is not integral to the imputation of missing housing costs data, this project provides a good opportunity to also consider the cleaning of these variables.

The rent questions are structured as follows:

- HJ3: Rent amount before housing benefit,
- HJ4: Period rent amount covers,
- HJ5: Rent amount after any housing benefit, with separate codes for no housing benefit and full housing benefit,
- HJ6: Period amount covers.

While there are “no housing benefit” and “full housing benefit” codes at HJ5, a number of cases (381) have the same rent amount given before and after housing benefit, while a small number of respondents (11) gave a higher amount after housing benefit than before, or the amount entered at HJ5 was £0 (13). This is shown in the table below.

**Table 4: Rent before and after housing benefit (2018)**

	<b>N</b>
Full housing benefit	589
No housing benefit	1512
Amount given after housing benefit <u>less</u> than before	289
Amount given after housing benefit <u>same</u> as before	381
Amount given after housing benefit <u>greater</u> than before	11
Amount after housing benefit = £0	13
Don't know/refused	93
<i>Total</i>	2888

This data can be cleaned based on responses given to receipt of housing benefit from the income section.

**1) Same amount given at HJ3 and HJ5.**

- a) In 2018, 353 of the 381 respondents said they were not in receipt of housing benefit in the income section. We suggest setting HJ5 to no housing benefit in these instances.
- b) Of the remaining cases, 21 said they were in receipt of housing benefit in the income section. We suggest setting HJ5 to full housing benefit.
- c) In the remaining 7 cases, there was not enough information from the income section to determine whether household is in receipt of housing benefit. We suggest no action for these.

- 2) Amount given at HJ5 is greater than HJ3.
  - a) This impacts only a handful of cases. Suggest a similar strategy to the above: if the income section shows the household receives housing benefit, recode HJ5 to full housing benefit, and if it shows no housing benefit, recode to no housing benefit.
- 3) Amount of £0 given at HJ5 but coded as receiving housing benefit in the income section. Suggest recoding HJ5 as full housing benefit.

Possible script amends to minimise the discrepancies above are:

- Hard check if the amount given after housing benefit (HJ5) is greater than the amount given before (HJ3).
- Soft check if the amount given after housing benefit (HJ5) is equal to the amount given before (HJ3).

There are also a small number of cases (56) where the information given at HJ3 and HJ5 appears to suggest that Housing Benefit is received but where receipt of this benefit is not recorded in the income section. While the routines for processing income have never imputed receipt of a benefit (just the amount received) there may be a case for imputing receipt of Housing Benefit for these cases. The decision on this would not have an impact on the routines for processing housing costs but would have an impact on those for calculating household income. We would be happy to discuss this option further with you.

## Outliers

The final stage of data cleaning is deciding how to deal with the outliers. With data of this nature, it's often easier to identify potential errors in higher values than lower values, e.g. a rent payment of £25,000 stands out as questionable more than one of £250, and so there is the temptation to more aggressively clean higher values. There is a considerable risk that this will skew the data. We therefore suggest a light touch approach to cleaning, which is in keeping with the cleaning of the income data. This involves flagging the top and bottom 5% of mortgage and rent payments<sup>3</sup> in turn and manually inspecting the outliers alongside relevant variables (including HiH age, HiH economic status, number of people in household, household income, length of time in property etc).

## Imputation procedures

To begin, summary flags were created for mortgage and rent (before housing benefit) to indicate the status of the housing costs, namely:

- Housing costs and amount given.
- Housing costs and amount given, but it's an outlier (i.e. +/- 5%).
- No costs, i.e. owner occupied without mortgage or living rent free.
- Housing costs, but no amount given and needs to be imputed.

The tables below show a breakdown of the level of missing data for mortgage and rent payments. In keeping with the income section, we have not imputed payment of costs, just the missing data. This only impacts the 45

---

<sup>3</sup> When asked about mortgage payments, respondents are specifically asked for their monthly mortgage payments. However, with rent payments, respondents are asked for an amount and then to specify the period that it covers (see Appendix A1). Therefore, before examining the 5<sup>th</sup> and 95<sup>th</sup> percentile of rent payments, the amounts are first standardised so that they cover a period of a month.

respondents who refused to answer whether they have an outstanding mortgage at HJ1. As with the income section, the top and bottom 5% of values are excluded as donor cases in the imputation. Overall, 17% of mortgage costs need imputed and 23% of rent payments.

**Table 4: Breakdown of missing mortgage data (2018)**

	Mortgage costs		Rent costs	
	N	%	N	%
Amount given by respondent	2,031	73	2,561	70
Amount given by respondent, but an outlier (do not used in imputation routines)	222	8	262	7
Amount to be imputed	474	17	857	23
Information on mortgage refused	45	2	NA	NA
<i>Total</i>	<i>2,772</i>		<i>3,680</i>	

Methods of imputation almost always rely on modelling existing data. The more variance that these models explain, the better the resulting imputation strategies will be. To inform the imputation strategy, a regression model was computed to help identify the variables that best account for differences in mortgage and rent costs.<sup>4</sup> The adjusted R<sup>2</sup> of the model for mortgage payments was 0.39 and 0.40 for rent payments. The factors found to have the most explanatory power were:

- Mortgage: household income, whether on means tested benefits, length of time in house, SIMD, age of HiH, type of property (i.e. flat or house), and to a lesser extent, area, urban/rural classification and household size. HiH economic status was initially included in the model but found to have little explanatory power.
- Rent: household income, whether on means tested benefits, SIMD, age of HiH, household size, and to a lesser extent, type of property (house/flat), area and urban/rural classification. Similar to mortgage, HiH economic status was excluded from the model as was found to have little explanatory power.

Mortgage and rent payments were primarily imputed by hierarchical hot deck imputation, the method used for imputing many of the components of household income. In this method, respondents are sorted into imputation groups based on factors correlated to the variable of interest. This is an iterative technique, so imputation is first attempted using donor groups based on all significant factors. Where this fails to return a match and there are still empty values, the factor with the least explanatory power is collapsed or excluded to create a model with fewer donor groups. The process is repeated until a match is found for all missing values. Mortgage and rent payments were imputed separately.

When creating donor groups, the aim is to return the most specific match possible, whilst also minimising the number of cases that each value donates too, e.g. to avoid a scenario in which one case is donated to ten missing

<sup>4</sup> In the case of both rent and mortgage, the distribution of housing costs was right skewed. In other words, the distribution of values above the median was more widely spread than below it. As regression models assume that the distribution of the dependent variable is normal, the model was run on the square root of earnings to help reduce the skew.

values. Groups were created through consideration of the factors with the most explanatory power, as well as the ratio of donors to donees.

As noted above, we have not historically imputed receipt of income or, in this case, payment of housing costs. This means that we have not imputed mortgage payments for cases of shared ownership where the respondent has said they do not have an outstanding mortgage on their property at HJ1, nor have we imputed costs where a respondent has said they live rent free (in 2018, there were 118 such cases).

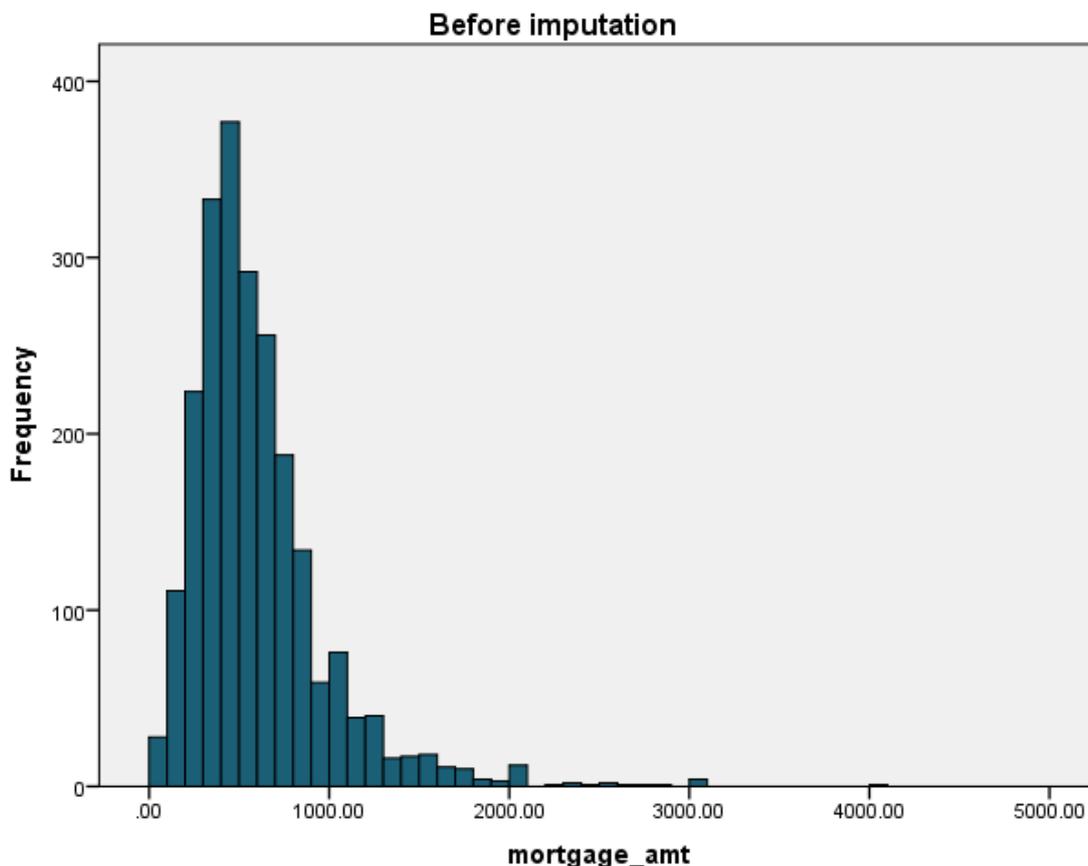
In 2018, there was only one shared ownership respondent that said they have a mortgage at HJ1 but is missing a value for a mortgage payment at HJ2. Given the small pool of donors, hierarchical hot deck imputation is not possible, and instead the median mortgage payment was imputed for this case. Likewise, there are 18 missing rent values for this group, which were again imputed using the median value.

## Conclusions

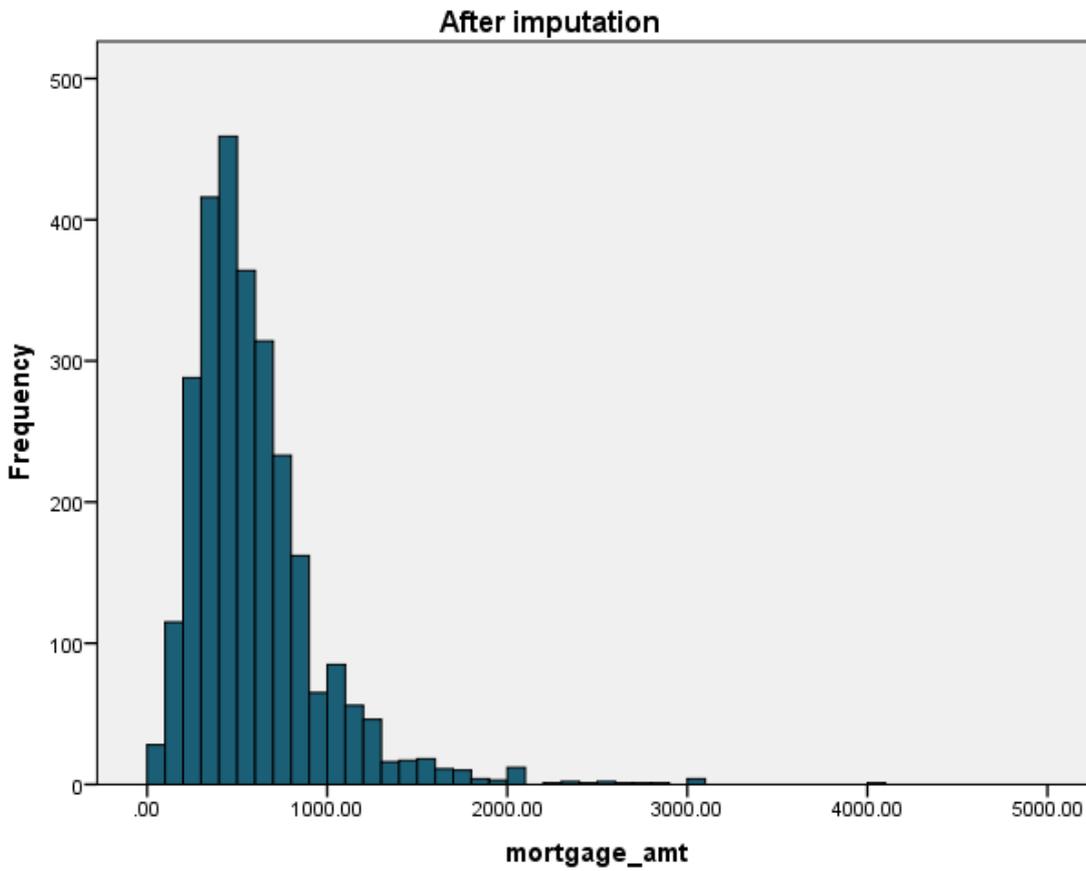
In conclusion, we foresee no problem in imputing the missing data for housing costs. There is a relatively low proportion of missing data (17% of mortgage costs and 23% of rent payments), and a good proportion of the variation is explained in the modelling. Going forward, implementation of the proposed script changes will help to tighten up the data collection and minimise the amount of cleaning and imputation needed.

An overview of the distribution of mortgage and rent costs before and after imputation is given below. As can be seen, in both instances, imputation has minimal effect on the mean and median payments.

Figure 1: Buying with a mortgage – distribution of monthly mortgage payments

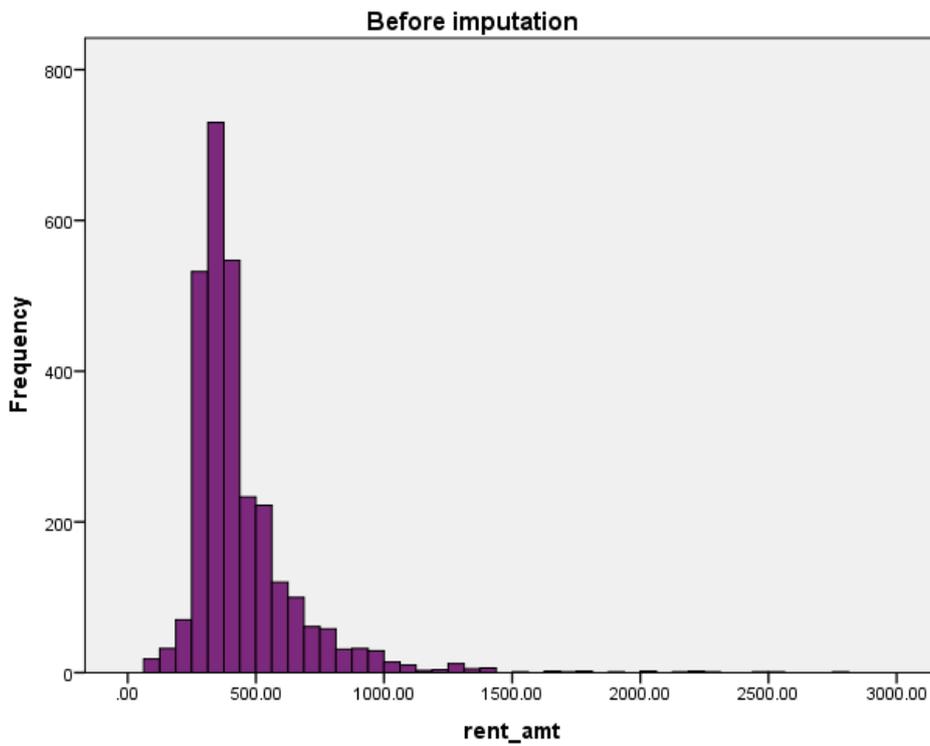


Mean = £585; Median = £500

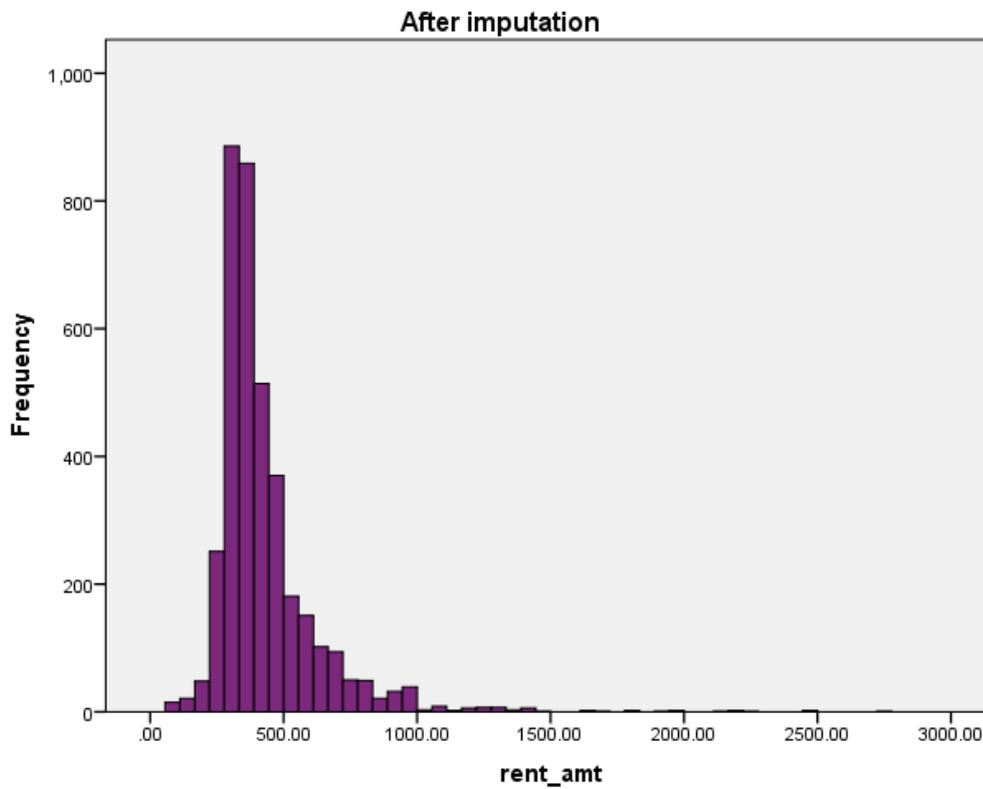


Mean = £576; Median = £500

Figure 2: Renting – distribution of monthly rent payments



Mean = £441; Median = £380



Mean = £433; Median = £373

An overview of the new variables that will be included in the dataset going forward is given in the table below. In addition to these new variables, once finalised, the cleaning routines will be used as standard for the delivery of the SHS data from 2019. These small changes will improve the quality of the data, but if used on historic data, will generate a minor discrepancy with previous datasets (as illustrated in this memo in relation to the 2018 data). Whilst supplying historic data is beyond the scope of this piece of work, should this be required in the future, two potential routes are:

1. Running the imputation routines but omitting the data cleaning. In effect, this would mean that the discrepancies between hb509 and hj1 would remain, as would those for the before/after housing benefit, but the new housing costs variables can simply be appended to existing datasets.
2. Running all the cleaning routines on historical datasets, and resupplying all the variables changed (e.g. HB509, HJ1, HJ2, HJ3, HJ4, HJ5, HJ6). This will mean changes to select variables in a small number of cases, particularly tenure (HB509) and outstanding mortgage (HJ1).

We would be happy to discuss further with the Scottish Government if needed.

**Table 5: Overview of new housing costs variables**

variable	Value labels
<b>mortgage_sum</b>  Summary variable outlining the status of mortgage costs. Note 9 is an interim status and cases will not have this value when supplied to Scottish Government.	-9: Owner occupied without mortgage
	-8: Renting HA/LA/PRS
	-7: Rent free
	-6: Part rent part mortgage
	-5: Information on mortgage refused
	1: Buying with mortgage - amount given by respondent
	2: Buying with mortgage - amount imputed
	3: Buying with mortgage - amount given by respondent, but not used in imputation routines
9: Buying with mortgage - amount still to be imputed	
<b>mortgage_amt</b>  Monthly mortgage amount for those buying with a mortgage.	N/A
<b>rent_sum</b>  Summary variable outlining the status of monthly rent costs. Note 9 is an interim status and cases will not have this value when supplied to Scottish Government.	-9: Owner occupied without mortgage
	-8: Buying with a mortgage
	-7: Rent free
	-6: Part rent part mortgage
	1: Renting - amount given by respondent
	2: Renting – amount imputed
	3: Renting - amount given by respondent, but not used in imputation routines
9: Renting - amount still to be imputed	

rent_amt Monthly rent amount (before housing benefit) for those renting (includes HA/LA/PRS but excludes those living rent free).	N/A
shared_ownership_sum Variable outlining the status of monthly housing costs (mortgage and rent) for part mortgage/part rent respondents.	-9: Owner occupied without mortgage
	-8: Buying with a mortgage
	-7: Renting HA/LA/PRS
	-6: Rent free
	1: Part rent part mortgage - both mortgage and rent given by respondent
	2: Part rent part mortgage - either mortgage or rent given by respondent, other imputed
	3: Part rent part mortgage - both rent and mortgage imputed
shared_ownership_amt Monthly housing costs – combined mortgage and rent where relevant.	4: Part rent part mortgage - rent given, but REF/No at mortgage question
	5: Part rent part mortgage - rent imputed, and REF/No at mortgage question"
	N/A
hcost_sum Variable outlining the status of monthly housing costs across those buying with a mortgage, renting and part mortgage/part rent.	-9: Owner occupied without mortgage
	-8: Refused question on whether had mortgage
	-7: Rent free
	1: With mortgage - amount given by respondent
	2: With mortgage - amount imputed
	3: With mortgage - amount given by respondent, but not used in imputation routines
	11: Renting HA/LA/PRS - amount given by respondent
	12: Renting HA/LA/PRS - amount imputed
	13: Renting HA/LA/PRS - amount given by respondent, but not used in imputation routines
	21: Part rent part mortgage - both mortgage and rent given by respondent
	22: Part rent part mortgage - either mortgage or rent given by respondent, other imputed
23: Part rent part mortgage - both rent and mortgage imputed	
24: Part rent part mortgage - rent given, but REF/No at mortgage question	

	25 Part rent part mortgage - rent imputed, and REF/No at mortgage question
hcost_amt Monthly housing costs.	N/A

## Appendices

### Appendix 1: Overview of relevant questions

An overview of the questions discussed in this note is given below. Please note, these questions are shown chronologically, but there are other questions asked in between. The full questionnaire is available at:

<https://www2.gov.scot/Topics/Statistics/16002/PublicationQuestionnaire/SHS2018Questionnaire>

Question name	Question	Stream
HB509	<p>SHOWCARD A6</p> <p>IN WHICH OF THESE WAYS DO YOU OCCUPY THIS ACCOMMODATION?</p> <p>(1) Buying with mortgage/loan [1]            (2) Own it outright [2]            (3) Part rent/part mortgage [3]            (4) Rents (including rents paid by housing benefit) [4]            (5) Living here rent free [5]</p>	CORE ANNUAL ALL
HJ1	<p><b>Ask if has a mortgage (HB509=1,2,3). Others go to HJ3.</b></p> <p>CAN I JUST CHECK, DOES YOUR HOUSEHOLD HAVE A MORTGAGE OR LOAN OUTSTANDING ON THIS PROPERTY?</p> <p>Yes [1]            No [2]  <b>Refused [3]</b></p>	FULL NOW ANNUAL ALL
HJ2	<p><b>Ask if has a mortgage</b></p> <p>AT THE MOMENT, HOW MUCH DOES YOUR HOUSEHOLD PAY <b>EACH MONTH</b> IN MORTGAGE PAYMENTS, EXCLUDING ANY CONTENTS OR BUILDINGS INSURANCE OR MORTGAGE PROTECTION?</p> <p><b>Code to the nearest £. Ask to estimate if unsure. Record total for all mortgages/loans.</b></p> <p><input type="text"/></p> <p>Don't know [9998]  <b>Refused [9997]</b></p>	FULL NOW ANNUAL ALL
HJ3	<p><b>Ask if HB509 is coded 3 or 4 (rented or shared ownership)</b></p> <p>I'D NOW LIKE TO ASK YOU ABOUT THE COST OF RENTING YOUR HOME.</p> <p>HOW MUCH IS THE TOTAL RENT FOR YOUR HOUSEHOLD'S ACCOMMODATION, THAT IS BEFORE ANY HOUSING BENEFIT IS DEDUCTED?</p> <p><b>If crofter, or two or more rents paid, enter total rent payable. If unsure code best estimate.</b></p> <p><input type="text"/></p> <p>Don't know [1]  <b>Refused [2]</b></p>	FULL NOW ANNUAL ALL
HJ4	<p><b>Ask if amount entered at HJ3 else go to HK1</b></p> <p>HOW LONG A PERIOD DOES THAT RENT COVER?</p> <p>A day [1]            A week [2]            Fortnight [3]            4 weeks [4]            Month [5]            Year [6]  <b>Other [7]</b></p>	FULL NOW ANNUAL ALL

HJ5	<p><b>Ask if amount entered at HJ3 else go to HK1</b></p> <p>AND HOW MUCH DOES YOUR HOUSEHOLD ACTUALLY PAY FOR THIS ACCOMMODATION AFTER ANY HOUSING BENEFIT HAS BEEN DEDUCTED?  <b>If no housing benefit received, leave amount blank and code 'no housing benefit'.</b></p> <div style="border: 1px solid black; width: 150px; height: 20px; margin: 10px auto;"></div> <p>Don't know [1]  Refused [2]  No housing benefit [3]</p> <p><b>Full housing benefit [4]</b></p>	FULL NOW ANNUAL ALL
HJ6	<p><b>Ask if amount entered at HJ5</b></p> <p>HOW LONG A PERIOD DOES THAT RENT COVER?</p> <p>A day [1]  A week [2]  Fortnight [3]  4 weeks [4]  Month [5]  Year [6]  Other [7]</p> <p><b>Not Applicable [8]</b></p>	FULL NOW ANNUAL ALL

Appendix 2: Historical trends in HB509 by HJ1 (2012-2017)

Buying with a mortgage/loan

	Does your household have a mortgage or loan outstanding on this property?			
	Yes	No	Refused	Total
2012	2986	72	55	3113
2013	3011	79	33	3123
2014	2826	87	27	2940
2015	2693	106	53	2852
2016	2675	109	41	2825
2017	2645	108	58	2811
2018	2655	95	45	2795

Owner occupied

	Does your household have a mortgage or loan outstanding on this property?			
	Yes	No	Refused	Total
2012	138	3377	49	3564
2013	95	3330	50	3475
2014	106	3389	31	3526
2015	85	3407	53	3545
2016	88	3515	42	3645
2017	85	3781	49	3915
2018	76	3740	40	3856

## Part mortgage/part rent

	Does your household have a mortgage or loan outstanding on this property?			
	Yes	No	Refused	Total
2012	13	138	3	154
2013	12	117	3	132
2014	15	106	9	130
2015	13	96	3	112
2016	15	80	4	99
2017	19	71	3	93
2018	10	66	7	83