Stats NZ
Tatauranga Aotearoa

Child poverty statistics:
Technical appendix
2017/18

New Zealand Government

Citation

Stats NZ (2019). *Child poverty statistics: Technical appendix 2017/18*. Retrieved from www.stats.govt.nz.

ISBN 978-1-98-858320-4 (online)

Published in April 2019 by
Stats NZ Tatauranga Aotearoa
Wellington, New Zealand

Contact
Stats NZ Information Centre: info@stats.govt.nz
Phone toll-free 0508 525 525
Phone international +64 4 931 4600

www.stats.govt.nz

Contents

Purpose

Child poverty statistics: Technical appendix 2017/18 sets out the methodology we used to prepare estimates of New Zealand's child poverty rates for the 2017/18 year and previous years. Stats NZ has utilised different data sources and methods to produce estimates for 2017/18 that are as robust as possible. The Government will use these as the baselines for its child poverty reduction targets.

Background

Child Poverty Reduction Act 2018

The Child Poverty Reduction Act 2018 passed into law in December 2018.

This Act reflects Government's intent to achieve a significant and sustained reduction in child poverty.

The Act's stated purpose is to: encourage a focus on child poverty reduction by successive governments and society, facilitate political accountability against published targets, require transparent reporting on levels of child poverty, and create a greater commitment to action by the Government to address the well-being of all children.

While the bill does not itself define 'child poverty', it does specify four primary measures followed by six supplementary measures.

1. Low income: less than 50% median equivalised disposable household income before housing costs (BHC) for the financial year
2. Low income: less than 50% median equivalised disposable household income after housing costs (AHC) for the base financial year
3. Material hardship
4. Poverty persistence [Note: reporting not required until the financial year beginning 1 July 2025]
5. Low income: less than 60% median equivalised disposable household income before housing costs (BHC) for the financial year
6. Low income: less than 60% median equivalised disposable household income after housing costs (AHC) for the financial year
7. Low income: less than 50% median equivalised disposable household income after housing costs (AHC) for the financial year
8. Low income: less than 40% median equivalised disposable household income after housing costs (AHC) for the financial year
9. Severe material hardship
10. Low income and hardship: less than 60% median equivalised disposable household income after housing costs (AHC) for the financial year and material hardship

Using the household economic survey

Until now, Stats NZ's household economic survey (HES) has been the data source for measuring poverty statistics. It is a random sample survey of 3,000 to 5,500 households, of which around one-third are households with dependent children. It is well suited to, and delivers valuable information for, purposes such as the overall distribution of household income and material well-being, and relativities between different groups.

However, when the focus is on very short-term changes, especially year on year, or when more precision is required in a given year, HES is not able to deliver robust results due to its relatively small sample size.

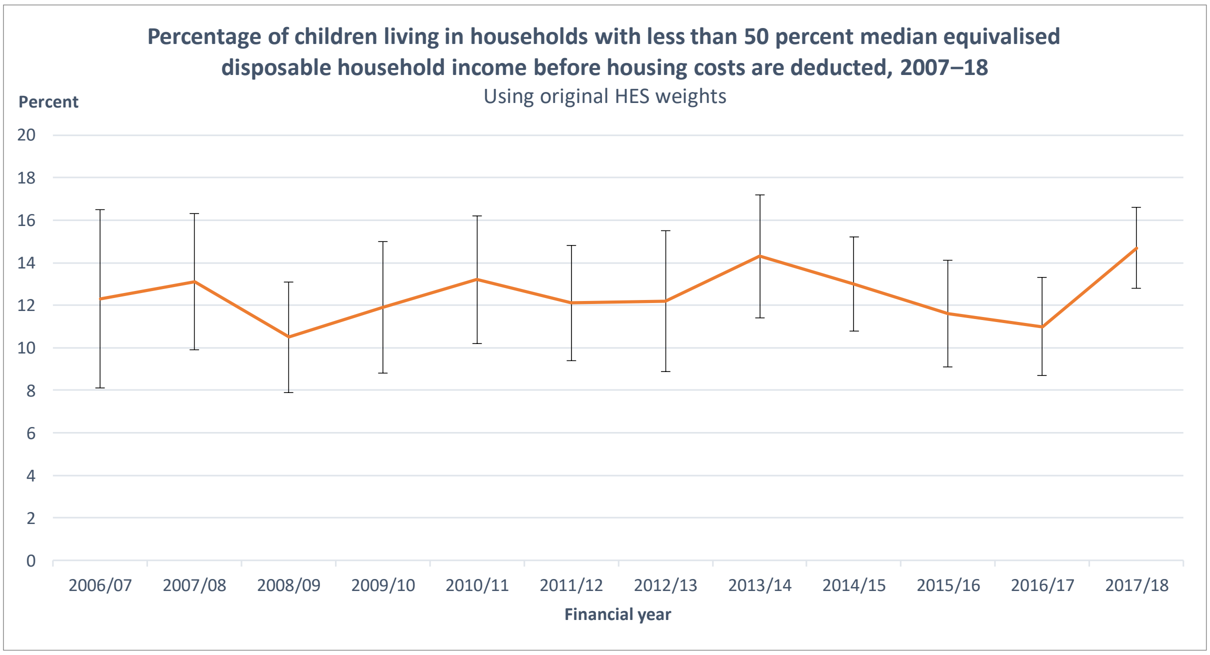
We also tend to have lower response rates from households in low socio-economic areas, which means that these households are often underrepresented in the sample. We have evidence that this sample bias is more pronounced in 2015/16 and 2016/17. Therefore, information on low-income households, such as child poverty, should be treated with extra caution for the 2015/16 and 2016/17 years.

In 2018 Stats NZ received additional funding from Government to improve the data source for measuring child poverty. This funding allowed: a substantial increase in the sample size of HES (to 20,000 households), a move to using administrative (admin) data for income rather than collecting income directly from respondents, and improvements to the survey design and operation to ensure a good representation of lower socio-economic households in the survey.

We implemented these improvements in the 2018/19 survey year, which collected data between July 2018 and June 2019. Results from this survey will be available in early 2020.

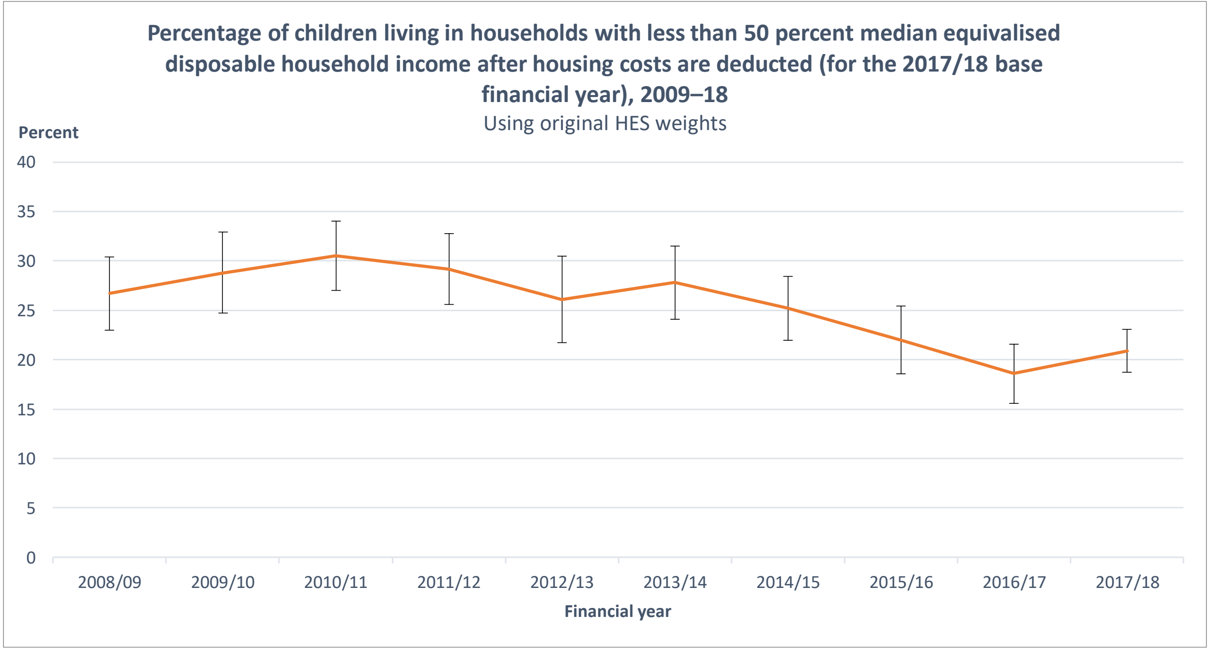
Figures 1, 2, and 3 show the estimates on three primary measures as calculated from HES before the improvements were implemented. As can be seen, these estimates are volatile year on year and have wide confidence intervals.

Figure 1



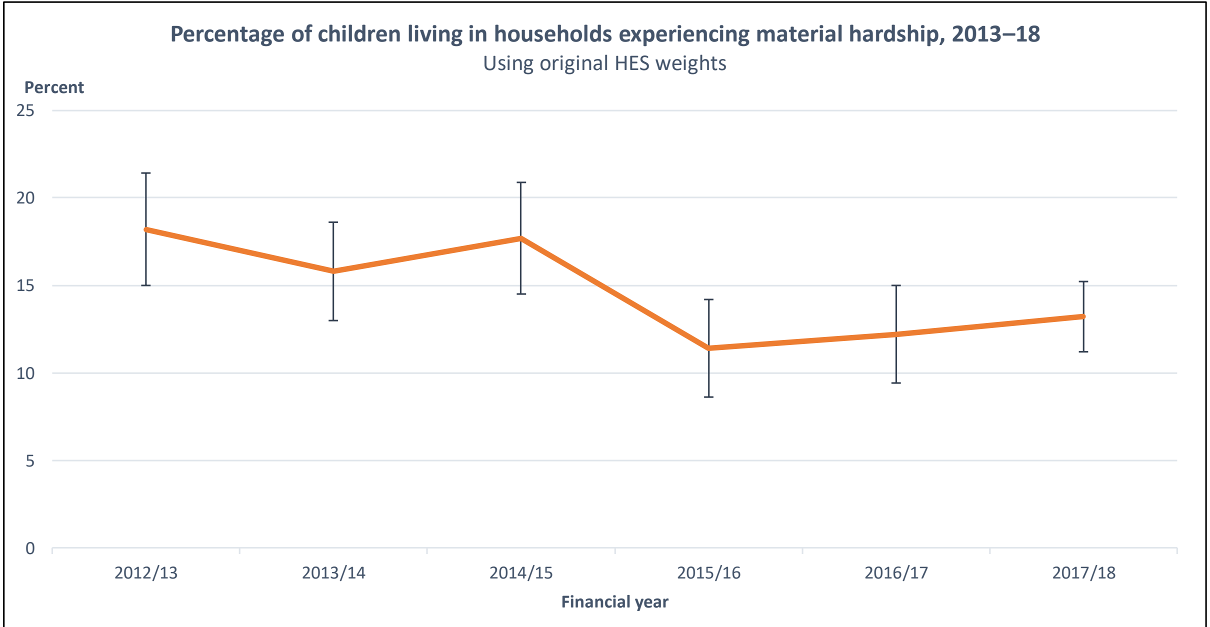
Note: Error bars in figure 1 show 95 percent confidence intervals, between which we are confident that the true rate lies.

Figure 2



Note: Error bars in figure 1 show 95 percent confidence intervals, between which we are confident that the true rate lies.

Figure 3



Note: Material hardship indicates a DEP-17 score of six or more

Note: Error bars in figure 1 show 95 percent confidence intervals, between which we are confident that the true rate lies.

The Act requires Government to set three-year and 10-year targets for the reduction of child poverty for the four primary measures. Note: targets and reporting on the poverty persistence measure (d) are not required until the financial year beginning 1 July 2025.

Cabinet decided the baseline year for targets would be the 2017/18 financial year so the first year of reporting on progress could be the 2018/19 year. This meant we needed to introduce new data sources and develop new methodologies – to provide the most robust estimate of the baselines that we can.

Approach to improvements describes the new data sources and methodologies.

Data used in this release

This section describes the data sources we used to prepare the estimates of child poverty being used for the 2017/18 baseline rates in the Child poverty statistics: Year ended June 2018 release.

Household economic survey

The household economic survey (HES) is an annual survey that collects a comprehensive range of statistics relating to household income and expenditure, and demographic information on households and individuals in New Zealand. The survey runs every year, from 1 July to 30 June of the following year. It covers people aged 15 years and over (15+) who usually live in New Zealand permanent private dwellings.

Statistics on household and personal income, housing costs, household and person demographics, and material well-being are produced. Housing costs include expenditure on mortgages, rents, rates, and building-related insurance.

Current design of HES at a high level

Households selected for HES are sampled from rural and urban areas throughout New Zealand on a statistically representative basis.

Stats NZ designs HES to achieve the sample size required to meet the survey's objectives. We make adjustments at the weighting stage to account for non-response, and weights are calibrated to known population totals.

The achieved sample size of the survey varies from year to year depending on the content of the survey. It is largest (5,500 households) when net worth questions are added every three years and smallest when expenditure questions and the expenditure diary are added, also every three years.

Table 1 shows achieved response rates and sample sizes over the last few years.

Points to note: 2012/13 and 2015/16 are years when expenditure was asked as well as income; response rates are typically lower in these years. The 2014/15 and 2017/18 surveys included the net worth module and the sample size was increased in those years.

Table 1
Achieved sample size and response rate for HES, 2011/12 to 2017/18

HES year	Achieved sample size	Response rate
	Number	Percent
2011/12	3,565	83
2012/13	3,003	67
2013/14	3,391	81
2014/15	5,561	78
2015/16	3,499	78
2016/17	3,703	83
2017/18	5,482	76

We allocate the HES sample evenly across all months in the year. Households interviewed each month are asked about their income in the previous 12 months. For example, a household interviewed in May 2016 provided income for the 12 months from May 2015 to April 2016. This means that income in each dataset covers two financial years.

HES collects income information from all adults in the household aged 15+. The material hardship questionnaire is administered to one randomly selected adult aged 18+.

Administrative data

Integrated Data Infrastructure

The Integrated Data Infrastructure (IDI) is a large research database that holds microdata about people and households. The data is about life events, like education, income, benefits, migration, justice, and health. It comes from government agencies, Stats NZ surveys, and non-government organisations. The data is linked together, or integrated, to form the IDI.

The IDI contains full tax data related to individuals, including data provided by employers for each employee (the employee monthly schedule), self-employment income, and some investment income. Data from the Ministry of Social Development includes benefits paid, including working for families' tax credits, and accommodation supplement. Data on housing includes information on people in social housing and tenancy bonds data.

Appendix 2 shows more detail of the income data used in this child poverty work that we sourced from the IDI.

Admin data used to replace survey data

We use income data from the IDI to replace the income collected directly from respondents in HES and to provide income for household labour force survey (HLFS) respondents. Despite best efforts to obtain data from respondents, survey data will always be subject to some uncertainty. This is due to respondents not being able to remember or not disclosing all sources of income over the year to the interviewer.

Respondents may also provide 'rough estimates' of amounts, or amounts that are exclusive of taxes paid. In some cases, family members may not know the income of all other family members. In particular, we know that salary and wages can be overstated when compared with admin data, usually due to respondents forgetting changes in income over the year. Benefit income is often understated, due to failure to recall small periods of benefit receipt through the year.

Some income sources are not covered by the data in the IDI at present. These include investment income, some sources of irregular income, and non-taxable income. We continue to rely on data collected from respondents for these income sources.

The quality of admin data on income is good, but it has some issues with timeliness. While most salary and wage income is provided monthly and flows through into the IDI on a quarterly basis, other income (eg self-employment income) relies on individuals providing their tax return, which can be delayed before being included in the IDI.

Almost all (99 percent) salary and wages data is provided within three months of the end of the financial year. However, only 13 percent of self-employment income data is provided within this timeframe. Of all income sources salary and wages make up around 70 percent, and self-employment income is around 15 percent. Where this data is not available we have used methods to ensure we have the best estimates of income we can gain, including dealing with timeliness issues.

Approach to improvements details the methods we used.

The data in the IDI is about individuals. It is not always straightforward to understand relationships between family or household members beyond assuming that respondents with the same address in the admin data form a household. Using address allows us to form households with the correct membership, when compared with census data, about 50 percent of the time. This is not good enough to create household income, which relies on having correct membership of the household. Therefore we used the household and family composition information collected in HES and HLFS to provide the household structure and relationships we need.

The potential for linked administrative data to provide household and family information describes more about the limitations of using the IDI for household information.

The IDI doesn't contain comprehensive data on housing costs. It does have information on people in social housing, and rent paid for renters with a registered tenancy bond, but does not hold information on mortgages, local body rates, or building insurance. For people who receive the accommodation supplement, some data on housing costs is also available.

Neither does the IDI have any information on the material hardship of households. For this reason, we cannot use admin data for improving the robustness of the income after housing costs measure nor the material hardship measure.

Household labour force survey

Stats NZ's quarterly household labour force survey (HLFS) is used to produce official estimates of the labour market in New Zealand, including the official unemployment rate.

The HLFS collects responses from around 15,000 households every quarter, amounting to approximately 30,000 individuals aged 15+. The HLFS interviews the same respondents over eight consecutive quarters, replacing them on a rotating basis with a new set of respondents.

The advantage of using the HLFS for the child poverty work is its bigger sample of households when compared with HES. The HLFS collects the same household structure information that HES does and can therefore provide (when linked with income data) a bigger pool of respondents. However, it doesn't collect housing costs information or material hardship measures so cannot expand the sample for these measures.

Methodology used to create estimates for child poverty measures

This section describes the methodology used to create the estimates for child poverty measures that are specified by the Act.

Measuring child poverty: Concepts and definitions explains the terms used in calculating child poverty measures in New Zealand.

Calculating disposable income

Disposable income is calculated for each household as the: sum of taxable income, non-taxable income, working for families' tax credits, and total rebates, less ACC earner's levy and tax payable.

Calculating number of children in low income

The following steps outline the process for producing estimates of the number of children in low income (before and after housing costs are deducted).

1. Calculate number of adults and children in each household to calculate equivalence factors – for modified OECD equivalence scale a child is aged under 14 and an adult is anyone aged 14 and over.
2. Calculate household disposable income by summing the disposable income of all household members 15+ years.
3. For after housing costs (AHC) measures, subtract total housing costs from household disposable income.
4. Calculate modified OECD equivalence scale factor for each household.
5. Calculate equivalised disposable household income by dividing household disposable income (BHC or AHC) by equivalence scale factor.
6. Calculate median equivalised disposable household income for each year.
7. Calculate 50 percent and 60 percent of median equivalised disposable household income to determine low income thresholds.
8. Determine if a household (and the household members) are under these thresholds (household equivalised disposable income < threshold value).
9. Calculate number (weighted) of children under the thresholds and calculate proportion of children in low income (number below threshold / total number of children).

Inflation adjustment

The low-income measure of 'less than 50% median equivalised disposable household after housing costs (AHC)' is presented as a fixed-line measure. This means the threshold is set for a reference year (in this case 2017/18) and then incomes of households are compared with this threshold.

For previous and subsequent years, the threshold is adjusted for inflation. The household living-costs price index for the low-income quintile (adjusted for housing costs) is used to adjust for inflation in the Child poverty statistics release.

Household living-costs price indexes: Background has more information on this index.

Measuring child poverty: Fixed-line measure provides further information about the fixed-line measure anchor point for measuring child poverty.

Approach to improvements

This section describes the approach we took to improve child poverty estimates for 2017/18 and previous years. This approach involved three main elements:

- using admin data
- increasing the sample size using HLFS data
- introducing new benchmarks to address known coverage issues in HES.

Not all elements of this approach can be applied to all three measures. Table 2 describes how each element was applied to the three primary measures. The before housing costs (BHC) income measure has had all three elements applied as good information on income is available in the admin data. The AHC income measure has used admin data and has been reweighted. However, the material hardship measure has only been able to be reweighted as no admin data is available for this measure.

Table 2

How the three elements of the approach to improvements were applied to each primary measure

Approach element	Before housing costs	After housing costs	Material hardship
	Whether used		
Admin data used	Yes	Yes	No
HLFS respondents used	Yes	No	No
New benchmarks used	Yes	Yes	Yes

Source: Stats NZ

Using admin data

A key part of Stats NZ's plan to improve child poverty estimates, as required by the Act, is to (as far as possible) replace personal income collected from survey respondents with data available from the IDI.

Administrative data discusses the reasons for doing this and the quality of admin data.

Linking HES to the IDI

Using admin data requires linking individuals in HES to the IDI spine. A high link rate is needed to ensure the best quality data.

The link rate of the HES sample to the IDI is 94 percent; for adults it is 95 percent.

The link to the IDI uses address, address history, name, and date of birth.

The link rate for children is lower than for adults because date of birth is not collected for children (although age is). This is not expected to affect the estimation of poverty rates using the IDI information as we rely on the HES data to tell us about the presence of children in households. All income (including benefits) is allocated to the adults in the household.

A false positive is where a link is made between individuals in the two sources who are not actually the same person. The false positive rate is estimated through a manual clerical review – a sample of links is drawn, and subject matter experts make judgements about whether links are correct or not. The estimated false positive rate for the HES-IDI link is 1.49 percent, with a sampling error of 0.63 percent. This rate is considered acceptable for this work.

Imputing income for missing links

While the link rate between HES and IDI is high (94 percent), we must ensure that any bias in the unlinked records was adjusted for. We know that people who are linked are more likely to have higher incomes, are more likely to be male, are more likely to be of European ethnicity and have lower reported benefit receipt than those who are not linked. We therefore imputed income for all the unlinked individuals.

For this imputation, we used the nearest neighbour donor imputation method implemented in the Statistics Canada software, Canceis. The method replaces missing values of a non-linked HES respondent with values from a linked respondent called a donor. A donor is selected by finding a respondent 'nearest' to the recipient (the non-linked respondent), in terms of other known characteristics that are correlated to the missing value being replaced. Characteristics include labour force status, age, and highest educational qualification. Tests of the resulting income distribution showed similar distributions when compared with HES.

Extracting income data

Once linked to the IDI, income data from the various sources described above is extracted for each individual in the final dataset. The admin data sources used are the same for HES and HLFS respondents, except for investment income for HLFS respondents – which is either sourced from admin data or imputed from HES data. Table 2 shows where data is sourced for HES and HLFS respondents.

Appendix 2 provides a fuller description of the sources of the income data extracted, rules relating to the extraction, and which high-level component of disposable income the data feeds into.

As HES interviews are conducted over a year the relevant income for the period before the interview is extracted from the admin data. For example, if a household was interviewed for HES in May 2016 we would retrieve income for May 2015 to April 2016.

For income data with timeliness issues, such as self-employed income and working for families' income, we take the most-recent income from returns within three years of the survey interview date.

Table 3

Data source for child poverty measures: For HES and HLFS households

Information needed	HES	HLFS
	Data source	
Salary and wages	Admin data	Admin data
Government benefits	Admin data	Admin data
Self-employment income	Admin data	Admin data
Investment income	Survey data	Imputed
Other regular income	Survey data	No data
Other irregular income	Survey data	No data
Housing costs	Survey data	No data
Material hardship	Survey data	No data

Source: Stats NZ

Sample size increased using HLFS

To overcome the problem of the relatively small sample size in HES we decided to use the HLFS information on household composition linked to the income data in the IDI. This data, combined with the HES sample, provides a larger pool of respondents. Using admin data on income provides a larger sample of households with household income.

The combined sample of HES and HLFS respondents gives an overall sample size of 19,991 households in the 2017/18 financial year.

To use the HLFS in this way requires the same process of linking to the IDI and extracting admin data as described above for HES.

Linking HLFS to the IDI

We wanted to ensure maximum linking rates between the IDI's spine and HLFS. An address history-based linking methodology was applied to the HLFS using the same linking variables as HES.

The link rate of the HLFS sample to the IDI is 91 percent; for adults it is 92 percent.

The estimated false positive rate is 1.18 percent, with a sampling error of 0.26 percent.

Imputation for non-linked individuals

As with HES, we imputed total admin income for the whole unlinked sample in HLFS using the same methodology.

Imputation for income variables not in admin data or HLFS

Investment income is currently not available in the IDI or in the HLFS. Investment income is likely to be a significant part of income, particularly for older individuals, and therefore if we did not include it this would have a significant effect on overall median incomes. Investment income has been imputed for HLFS respondents.

However, since other sources of income not available in the IDI or for HLFS respondents, such as non-taxable income or directors' fees, have a minor impact on median incomes, they were not imputed.

An HLFS respondent who did not report income from these Inland Revenue income sources is imputed – using investment income available from HES respondents who also do not have Inland Revenue-sourced investment income. Multiple imputation using the R-package MICE (multiple imputation using chained equations) is used.

Tests show the HLFS median investment income as imputed does not significantly differ from that reported in HES across all years.

Combining HES and HLFS samples (pooling)

Before pooling HES and HLFS unit record datasets, we need to satisfy some assumptions required for data pooling. The assumptions that the characteristics of the population covered by each survey and that each of these surveys is an independent non-overlapping sample of the same population are assured by the sample design of these surveys, although there's a slight difference in the definition of their target populations.

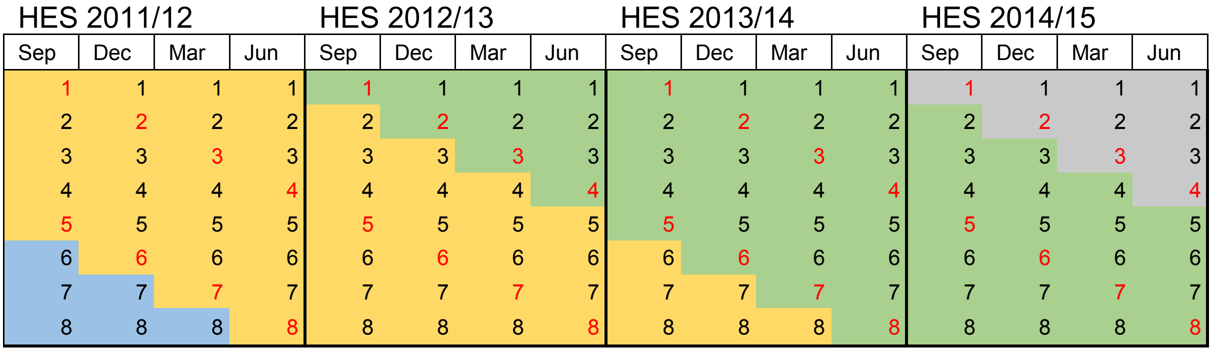
The target population for both surveys includes all civilian, non-institutionalised, usual residents of New Zealand aged 15 years and over. HES includes only usual residents from private dwellings while HLFS extends this to include usual residents from non-private dwellings. Both surveys use the same sampling frame, the household sample frame. This is a database of primary sampling units (PSUs) formed from data collected from the census. PSUs are small geographic areas containing an average of 70–100 dwellings; they are designed to be geographically contiguous in most cases. The PSUs are updated after each census. Each survey uses a different set of PSUs, which ensures they are non-overlapping samples of the same population.

An additional challenge to including HLFS is its design. The HLFS uses eight rotation groups; a group stays in the survey for eight consecutive quarters. For each quarter, one of the eight rotational groups moves out of the sample and a new group moves in. Each quarterly sample retains seven of the eight respondents interviewed in the previous quarter.

To match the HES interview pattern we decided to use two rotation groups each quarter from the HLFS. Income information would be retrieved from admin data for the 12 months before the quarter the interview was held in. We also decided to select rotation groups that had completed their first and fifth interviews.

Table 4 illustrates the rotation groups we chose for the different quarters, to align to specific HES years. The solid block of colour is the portion of the HLFS sample assigned to a rotation group. Respondents remain in this rotation group for the eight quarters. The red numbers show which rotation groups we selected for each quarter. Over the year all eight rotation groups are selected. The selection can be extended in the same way to other HES years.

Table 4
Household labour force survey rotation groups



HES 2011/12				HES 2012/13				HES 2013/14				HES 2014/15			
Sep	Dec	Mar	Jun	Sep	Dec	Mar	Jun	Sep	Dec	Mar	Jun	Sep	Dec	Mar	Jun
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8

Source: Stats NZ

Weighting the combined samples together

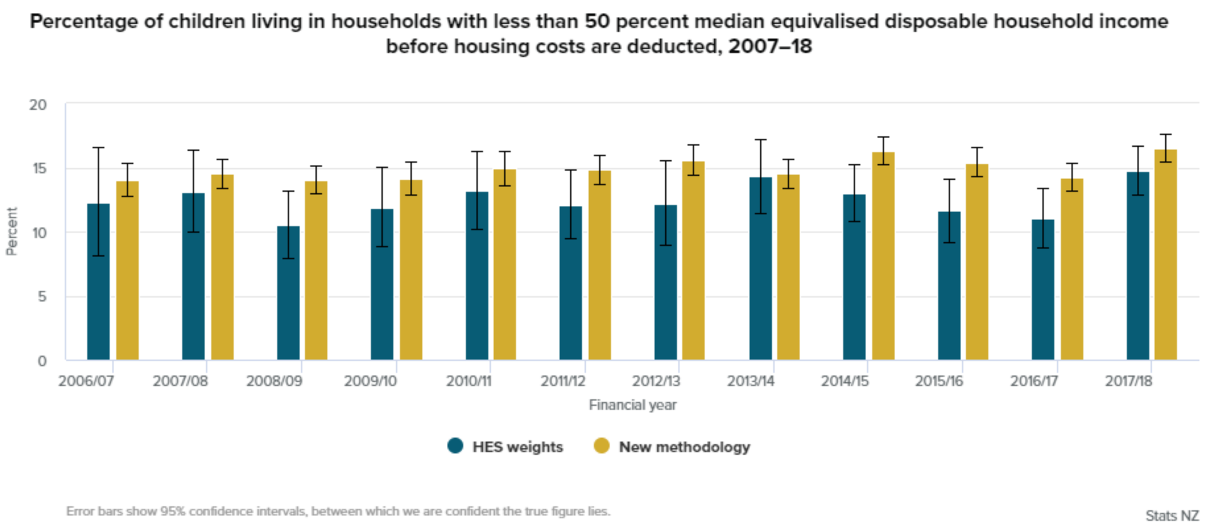
We rescaled the selection weights of the HES and HLFS respondents before estimation was carried out to represent the increase in sample size. The selection weight of a HES household was multiplied by the ratio of the number of households selected in a HES year to the sum of the HES and HLFS households selected in a HES year.

Similarly, the selection weight of a HLFS household was multiplied by the ratio of the number of HLFS households selected in a HES year to the sum of the HES and HLFS households selected in a HES year. We then calibrated to HES benchmarks and applied integrated weighting to the rescaled selection weights. Again, we used integrated weighting to ensure all individuals in a household received equal weights.

Outcome of the pooling for BHC

Figure 4 shows the resulting measure for BHC income. We have reduced the variability year on year, and sample errors are now in the 1- 1.3 percent range (where previously they were in the 1.9 -4.2 percent range). Due to replacing income with administrative data we have corrected some of the reporting error we have seen – as a result low-income rates are higher than those calculated from HES previously.

Figure 4



Percentage of children living in households with less than 50 percent median equivalised disposable household income before housing costs are deducted, 2007–18

Error bars show 95% confidence intervals, between which we are confident the true figure lies.

Stats NZ

Introducing new benchmarks

For the BHC income measure we can combine the HLFS and HES samples because we have admin data that we can use to supply the income data required for this measure.

We do not have similar high-quality admin data sources for either housing costs or material well-being information. Therefore, we cannot use the approach that we have taken for the BHC income measure, using the HLFS sample linked to income from admin data to create a larger sample. For the AHC income measure and the material hardship measure we investigated ways of reweighting the sample to account for sampling variability and the observed non-response bias. This investigation has had mixed success as discussed below.

We use population benchmarks to adjust for possible under-coverage of certain population groups in the sample. This ensures our final weighted estimates reflect the actual distribution of the population. In most cases, using population benchmarks controls volatility in the proportions that is due to sample variations; it can also correct somewhat for sample bias caused by non-response.

For example, HES may obtain responses from more women than men – we use population benchmarks to adjust the proportions to what is seen in the total population. Population benchmarks are usually derived from census data or other sources where we are confident we have known population distributions.

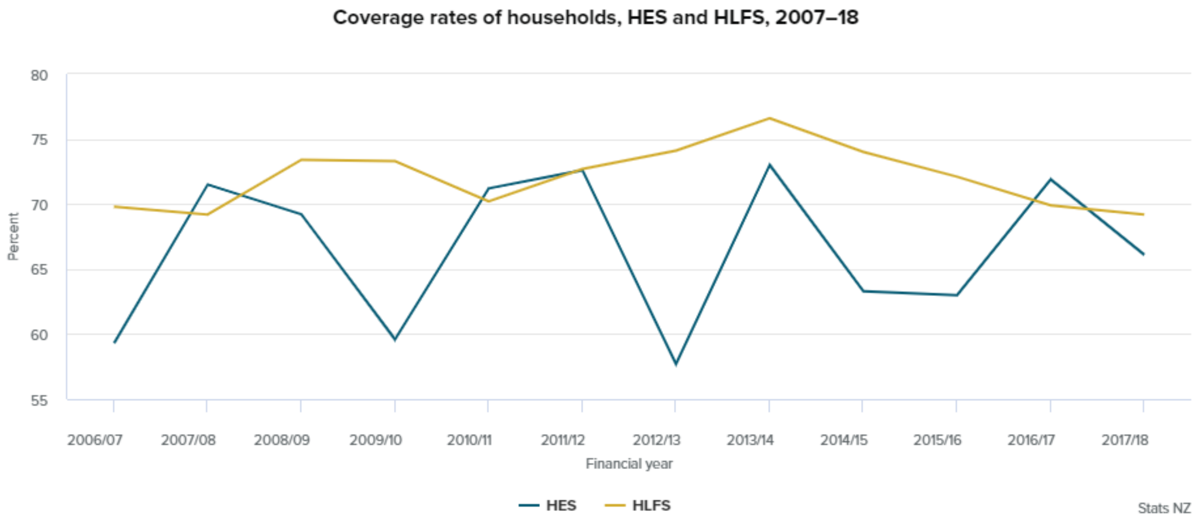
The benchmarks used initially for HES are:

- five-year age groups, by sex (although 0–14 years is not broken down by age)
- Māori, by 0–29 years and 30+ years
- two-adult households, by region; other households, by region
- region (Auckland, Wellington, Canterbury, Rest of the North Island, Rest of the South Island).

In this work we have retained the age-by-sex and Māori benchmarks, and added additional benchmarks.

Figure 5 shows that HES consistently has a lower response rate than HLFS and that this response rate varies a lot year to year. In particular, the coverage of single-adult households with children and lower-income households in HES in some years is lower and more variable than in the HLFS.

Figure 5



To improve the stability of the HES series we have added two population benchmarks to the weighting of HES, which are estimated from the larger HLFS sample. We did this to:

- calibrate the HES household income distribution to the HLFS household income distribution. We have used vigintiles (20 categories each representing 5 percent of the population) with benchmarks calculated using equivalised disposable household income vigintiles from HLFS admin data
- calibrate the HES sample so it matches the household-type distribution in HLFS. We created benchmarks for household type (1 adult, 2 or more adults with 0, 1, 2 or more children) by three broad equivalised disposable household-income categories (decile 1, decile 2, and remaining deciles) calculated from HLFS data.

The household income benchmarks are more highly correlated to income poverty measures than to material hardship measures. Therefore we expect this to have a bigger impact on the AHC income measure than on the material hardship measure.

For these new benchmarks we would ideally use just one income by household-type benchmark, but this would result in a large number of categories. The HES sample numbers limit the number of categories we can use as some may have no sample in them. Because of this we chose one set of benchmarks to capture the complete income distribution, and a second set that cross some key household types with broad income categories.

The HLFS estimates used as benchmarks will have sampling error associated with them. The sampling errors of the HLFS benchmarks were estimated using a bootstrap approach. We include this uncertainty in the HES estimation by calibrating each HES bootstrap sample to a benchmark generated from an HLFS bootstrap sample.

Reducing non-response bias

While the primary motivation for including benchmarks generated from the HLFS is to reduce the volatility in the series, the HLFS also has a higher response rate than HES and therefore may not have the same level of non-response bias as HES. Using the benchmarks discussed above allows us to partly correct for some non-response bias in HES.

The benchmarks we have used (income and household type) help to adjust for the lower response of low-income and single-parent households to some degree. However, because income and material hardship are not well correlated, the gains in using these benchmarks for the material hardship measure are not as large.

Testing against alternative benchmarks

The amount of calibration occurring is more than we would normally accept for a survey of this size.

To decide which benchmarks to use, we tested the impact on the AHC and material hardship measures of collapsing the HLFS benchmark categories into broader groups. This included looking at using separate income and household-type benchmarks (ie not crossing these two benchmarks together) and collapsing categories such as one-adult-with-child(ren) households into broader categories.

While the trend and latest estimates were similar when making these changes, using the reduced benchmarks introduced higher levels of year-on-year volatility into the measures. For this reason, we decided to stay with the initial HLFS benchmarks without collapsing them further.

Another option would be to calibrate to totals, formed directly from the IDI, that did not use the HLFS sample to form household estimates. Examples are the total number of adults on a benefit, or the income distribution of individuals classified as being in the IDI estimated resident population.

However, because household composition in the IDI is not of high quality, it is difficult to form household-level benchmarks from the IDI directly. While exploring these options further would be useful we decided to use the HLFS sample linked to the IDI data sources for the benchmarks – because this allowed household-level variables to be included in the calibration (eg household type by household income). Also, the sample size still enabled us to produce accurate benchmarks.

Outcome of the introduction of new benchmarks to HES

We have used the new population benchmarks described above for both the AHC income measure and the material hardship measure.

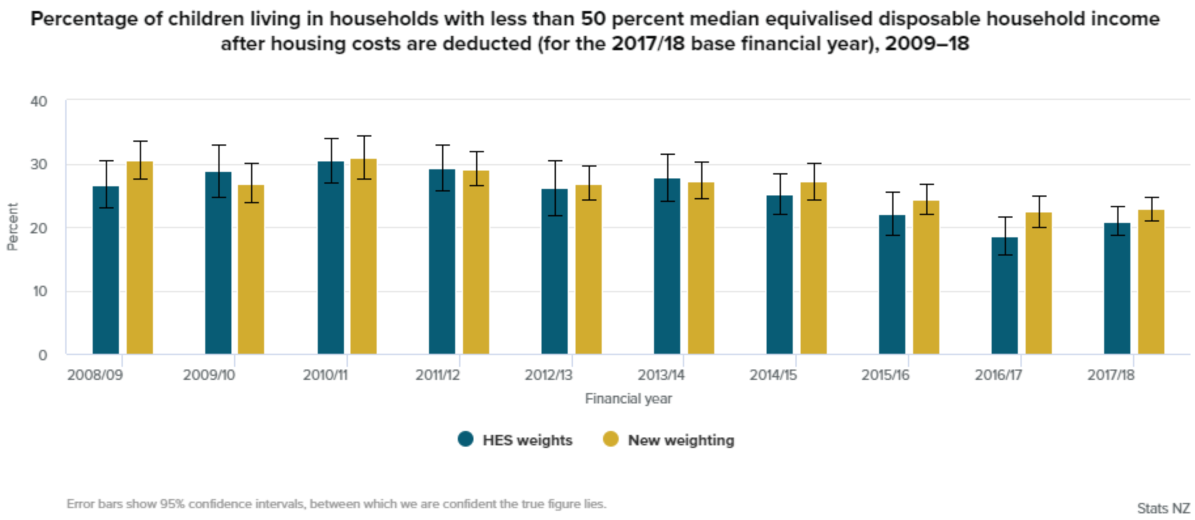
For the AHC measure (see figure 6) doing this has smoothed volatility and reduced the sample errors on the annual movements. Sample errors on rates are now between 2 and 3 percent (they were previously between 2 and 4.5 percent) The AHC income measure replaces income with admin data – as in the BHC measure, rates are a bit higher than those calculated from the original HES.

For the material hardship measure the results of the reweighting had less impact than for the income measures. There have been slight changes to the estimates but a large drop in rates between 2014/15 and 2015/16 remains. This is not a surprise given the additional benchmarks we used are more closely related to income than to material hardship. It is harder to find benchmarks that would have a larger impact on the material hardship estimates.

The reweighting of the material hardship measure has not improved the sample errors for this measure; for some years they are slightly larger than previously calculated. This means this series needs to be used with caution, particularly for early years.

Across all three measures we found that estimates for the 2017/18 year have been robust to the methods and testing we have done. This provides confidence that the 2017/18 year estimates are robust for setting targets.

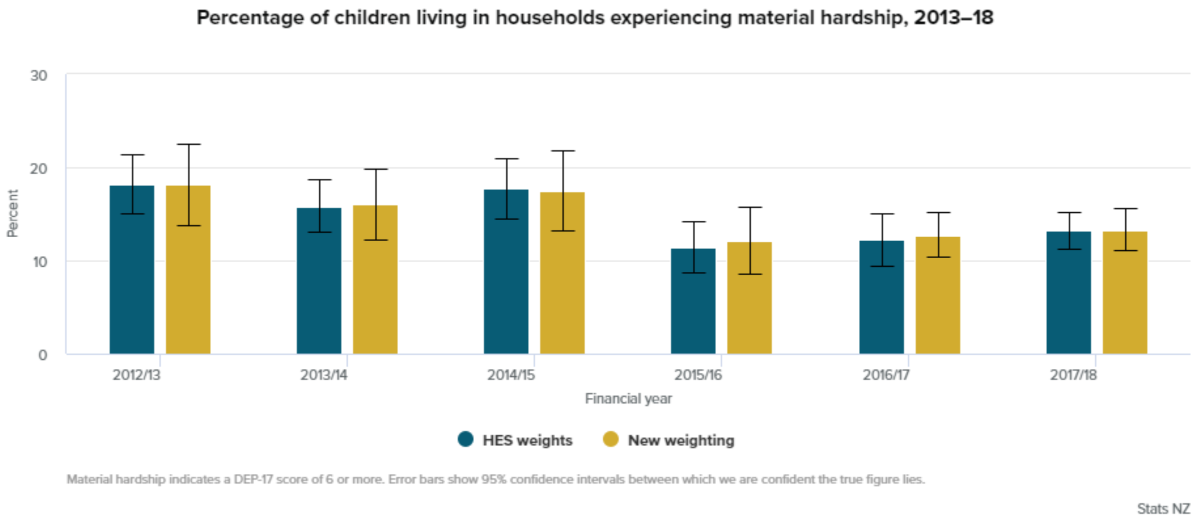
Figure 6



Error bars show 95% confidence intervals, between which we are confident the true figure lies.

Stats NZ

Figure 7



Material hardship indicates a DEP-17 score of 6 or more. Error bars show 95% confidence intervals between which we are confident the true figure lies.

Stats NZ

Alternative approach to AHC income measures

We tested an alternative approach to producing AHC income measures by imputing housing costs for the HLFS component of the sample. We did this using a combination of area-level housing-cost information, and individual-level rental information sourced from tenancy data.

In preliminary results we found that:

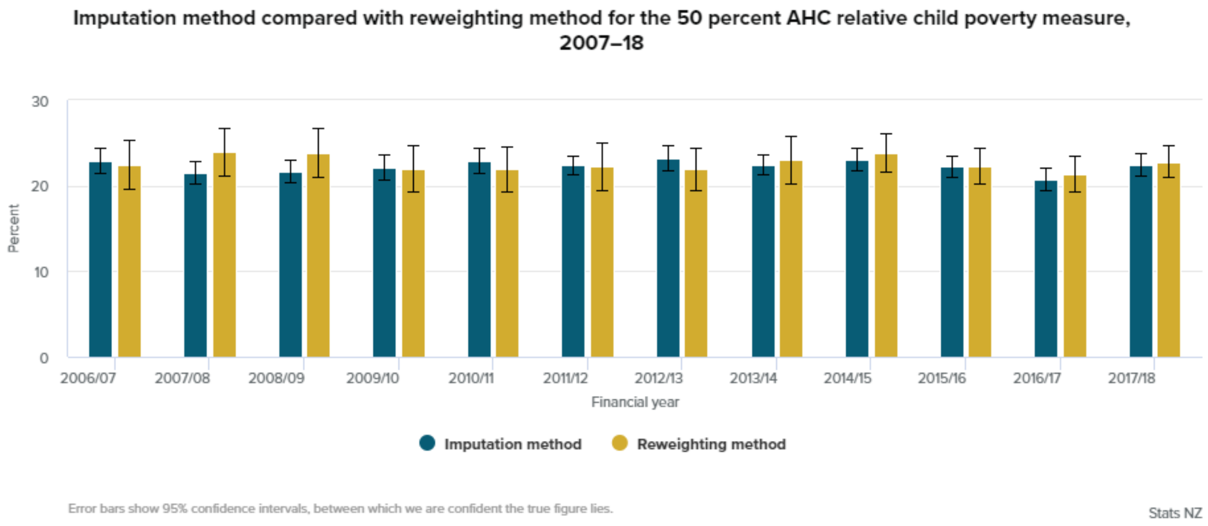
- the housing-cost model failed to adequately model the distribution of housing costs for people who were not found in tenancy bond data – this led to a large amount of housing-cost information being pulled towards the centre of the distribution
- it was not clear how the model would perform on HES years before 2016/17, when tenure information was not available in the HLFS. Tenure information was a key variable in the housing-cost model – the tenure question was introduced to the HLFS in 2016

- it was likely that the estimation error associated with modelling housing costs would be significant; we would need a methodological approach to combine this with our existing measure of sampling error.

Because of these issues we decided the approach that weighted the HES sample gave us the most robust estimates. In making this decision, we accepted that our chosen approach would result in larger sample errors.

As figure 8 shows, despite using very different methodologies, the two approaches to the AHC income measures produced similar estimates across the time series. This was further encouragement that we have produced robust estimates.

Figure 8


Imputation method compared with reweighting method for the 50 percent AHC relative child poverty measure, 2007–18

Error bars show 95% confidence intervals, between which we are confident the true figure lies.

Stats NZ

Appendix 3 details the methodology we used to impute housing costs for HLFS respondents.

Median income

This section provides information on the median household equivalised disposable income used in calculating estimates of child poverty measures.

For the tables and graphs below:

- the HES dataset is the dataset used previously for income distribution and low-income analysis. It is a combination of HES data and data from modelling work done by Treasury
- the HES-IDI dataset is a HES sample where most of the income data is replaced by admin data. The weights used are the original survey weights
- the HLFS-IDI dataset is the HLFS sample created for this work where the income is from admin sources. The weights used are those created for this work.
- the pooled HES-HLFS dataset is the HES-IDI and HLFS-IDI datasets pooled together. The weights used are pooled weights.

Before housing costs income measures

The median household equivalised disposable incomes before housing costs (BHC) are deducted are detailed in table 5. These are the medians used to produce thresholds for the BHC poverty measures.

The estimates of the medians produced from the different datasets are very similar. A general pattern shown in the data is that the HES-based estimates tend to be slightly higher than the HLFS and pooled estimates.

Table 5
Median household equivalised disposable income before housing costs

HES year	Original HES weights	Pooled HES-HLFS
	Median ($)	
2007	26,500	25,300
2008	28,200	27,000
2009	29,900	28,700
2010	30,300	29,200
2011	30,900	29,600
2012	31,600	30,800
2013	32,400	32,800
2014	34,500	33,100
2015	36,000	34,700
2016	35,700	35,800
2017	38,200	36,500
2018	39,900	38,800

Source: Stats NZ

After housing costs income measures

The median household equivalised disposable incomes after housing costs (AHC) are deducted are detailed in table 6. These are the medians used to produce thresholds for the AHC poverty measures.

Table 6
Median household equivalised disposable income after housing costs

HES year	Original HES weights	Pooled HES-HLFS
	Median ($)	
2007	20,500	19,300
2008	21,300	20,400
2009	23,400	22,200
2010	23,700	22,900
2011	23,700	22,700
2012	24,800	23,300
2013	26,000	25,000
2014	26,800	25,800
2015	27,700	26,000
2016	27,100	27,300
2017	29,700	28,100
2018	30,500	29,300

Source: Stats NZ

ABS review

The methodology used to improve the estimates of child poverty outlined above was reviewed by colleagues at the Australian Bureau of Statistics (ABS). The review found that an appropriate and rigorous methodology has been used to develop the estimates. The review made several very helpful suggestions on the main risks and areas for potential improvement in the methods. We made many of these improvements before finalising this report.

The ABS's final review comments are available on request to Stats NZ.

Appendix 1: Tax and ACC earners' levy scales

Tax rates

Table 7 details the tax rates used in the 2007/08 to 2017/18 tax years. These rates are used in the calculation of disposable income.

Table 7
Tax rates 2007/08–2017/18

Tax year	Income range 1	Tax rate 1	Income range 2	Tax rate 2	Income range 3	Tax rate 3	Income range 4	Tax rate 4
	($)	Percent	($)	Percent	($)	Percent	($)	Percent
2011/12 – 2017/18	Up to 14,000	10.5	Over 14,000 and up to 48,000	17.5	Over 48,000 and up to 70,000	30	Remaining income over 70,000	33
2010/11	Up to 14,000	11.5	Over 14,000 and up to 48,000	19.25	Over 48,000 and up to 70,000	31.5	Remaining income over 70,000	35.5
2009/10	Up to 14,000	12.5	Over 14,000 and up to 48,000	21	Over 48,000 and up to 70,000	33	Remaining income over 70,000	38
2008/09 (see Table 8)	…	…	…	…	…	…	…	…
2006/07 – 2007/08	Up to $9,500	15	Over 9,500 and up to 38,000	21.00%	Over 38,000 and up to 60,000	33	Remaining income over 60,000	39

Symbol: … not applicable

Source: Stats NZ

2008/09 tax year

The tax rates used for the 2008/09 tax year are rather messy as a new tax system was introduced mid-year. Table 8 amalgamates the two different tax systems.

Table 8
Tax rates 2008/09

Income range ($)	Tax rate
	Percent
Up to 9,500	13.75
Over 9,500 and up to 14,000	16.75
Over 14,000 and up to 38,000	21
Over 38,000 and up to 40,000	27
Over 40,000 and up to 60,000	33
Over 60,000 and up to 70,000	36
Remaining income over 70,000	39

Source: Stats NZ

ACC earners' levy rates, maximum liable income, maximum levy

Table 9 details the ACC earners' levy rates, maximum liable income, and maximum levy for the 2006/07 to 2017/18 tax years. These rates are used in the calculation of disposable income.

Table 9
ACC earners' levy rates, maximum liable income, maximum levy

Tax year	Levy rate	Maximum liable income	Maximum levy
	Percent	($)	($)
2007	1.3	96,619	1,256.04
2008	1.3	99,817	1,297.62
2009	1.4	102,922	1,440.91
2010	1.7	106,473	1,810.04
2011	2.02	110,018	2,222.36
2012	2.04	111,669	2,278.04
2013	1.7	113,768	1,943.05
2014	1.7	116,089	1,973.51
2015	1.45	118,191	1,713.76
2016	1.45	120,070	1,741.01
2017	1.39	124,053	1,724.33
2018	1.39	126,286	1,755.37

Source: Stats NZ

Appendix 2: Income data sources

Table 10 details the sources of the income data extracted, rules relating to the extraction, and which high-level component of disposable income the data feeds into.

Appendix 3: Housing-costs modelling

Information on housing costs is required for the after-housing costs (AHC) measure. The IDI has no admin data on housing costs that could be used to produce direct housing costs for households in the HLFS sample. However, there are other variables that could potentially be used to model housing costs for these households. The model used is described below.

Methodology and variables

Housing costs were modelled using a Random Forest. This model uses a group of predictive decision-trees to determine which variables provide the best fit. The population was split into two groups – renters and non-renters; we modelled housing costs separately for each group.

Variables come from several sources, including the HLFS/HES, tenancy bond, and accommodation supplement information.

Key variables include:

- tenancy information – owner occupied or rented
- regional council for household
- year of survey collection*
- number of bedrooms and rooms from bond data
- household disposable income
- percentage of females and minority ethnicity individuals*
- number of adults and children in household*
- average age of individuals in household*
- rent and bond amounts for household and geometric mean rent for meshblock and territorial authority – we took only bond records from the last two years since records do not reflect rent increases in after the tenure start date
- housing costs from accommodation supplement; annualised, weekly, and at time of survey interview date
- subsidised rent from Housing New Zealand (HNZ) for people living in social homes – we replaced a bond record with HNZ rent information; these people were included in the renters model
- New Zealand Deprivation Index (NZDep) for the household's meshblock.

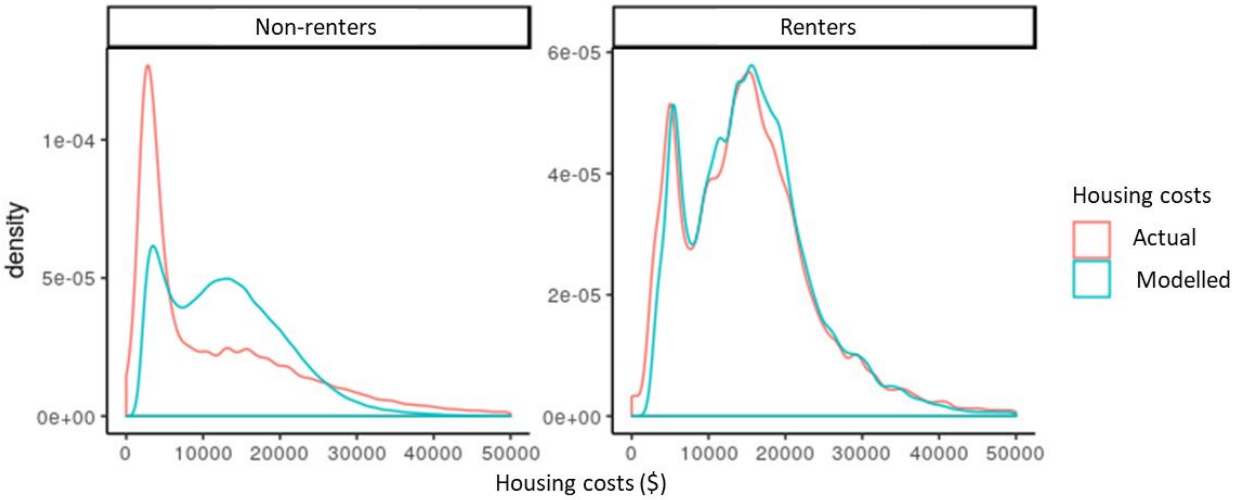
***** These variables come from surveys.

Model performance using HES data

The renters model performed very well, explaining approximately 65 percent of the overall variance in the model. The non-renters model explained around 30 percent of variance. The rent indicates the housing costs quite strongly, but the non-renters model also has a number of sub-populations that also would have lowered the accuracy. For example, consider two households with the same structure and similar income, but one household does not have a mortgage.

Figure 9 compares the modelled housing costs with reported (actual) housing costs for each household. The renters model tends to perform evenly from zero to $50,000 in housing costs. The non-renters model shrinks estimates slightly towards the average. This reflects the fact there are almost indiscriminable sub-populations that may have incredibly low or high housing costs.

Figure 9



Despite some evidence of the non-renters model reducing estimates towards the population average, this appeared to have little effect on the AHC indicators for children. This is likely because the medians we used to determine poverty lines are still the same; those in poverty are significantly more likely to have a bond or social home and would be modelled to an acceptable standard.

We used multiple test-train sets of data to determine how differences between modelled and reported housing costs affect poverty rates. Specifically, we trained a model on half the HES data and calculated poverty rates with modelled and reported housing costs separately for comparison. We repeated this process 50 times; that is, we produced 50 AHC child poverty rates using reported housing costs and 50 rates using modelled housing costs. All estimates were unweighted as using full survey weights for half the sample would not have appropriately controlled for survey design. As a result, the unweighted estimates may be more volatile.

Across most years there is very little evidence of bias due to modelled estimates. The only time-point of concern is the 2006 year, with slightly more than 3 percentage points difference. All other years are well within our expected sampling error. We also expected the differences to become smaller when we produced a model using the entire HES dataset to predict housing costs in HLFS. Test-train sets halved our training data and led to some of the volatility we see in the rates.

Applying the model to the HLFS data

We split the HES data into tenancy bond and non-bond groups and then trained random forests to predict HLFS housing costs. When pooling, we also predicted housing costs for the HES sample, so we subjected both survey sources to the same model uncertainty. That is, we did not want all HLFS samples to be slightly over- or under-estimated, therefore making HES samples look proportionately lower.

We considered the impact of missing a survey tenure indicator before the HLFS redevelopment in 2016. We removed the tenure indicator from HES data and produced another set of unweighted child poverty rates using multiple train-test iterations. There didn't seem to be any systematic difference between the two approaches. The differences we observed were all less than 1 percentage point between modes with and without a tenancy indicator in HES data.

We cannot rule out that the absence of a tenure indicator is having an effect, but it does not seem to have a systematic effect on the estimates.

References

Ball, C, and Ormsby, J (2017). Comparing the Household Economic Survey to administrative records: An analysis of income and benefit receipt. Wellington: The Treasury. Available at: https://treasury.govt.nz.

Black, A (2016). The IDI prototype spine's creation and coverage. Wellington: Stats NZ. Available at: http://archive.stats.govt.nz.

Brieman, L, Friedman, JH, Olshen, RA and Stone, CJ (1984). *Classification and regression trees,* Wadsworth, Belmont.

Chambers, R (2001). Evaluation criteria for statistical editing and imputation, National Statistics Methodological Series, 28, London.

Gath, M, & Bycroft, C (2018). The potential for linked administrative data to provide household and family information*.* Retrieved from www.stats.govt.nz.

Ishwaran, H, & Kogalur, UB (2019). Random forests for survival, regression, and classification (RF-SRC), R package version 2.8.0.

Ishwaran, H, Kogalur, UB, Blackstone, EH & Lauer, MS (2008). Random survival forests. *Ann. Appl. Statist. 2(3),* 841–860.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Tang, F, & Ishwaran, H (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining, 10*, 363–377.