

Evaluating the use of admin data as a substitute for Household Labour Force Survey income variables



These results are not official statistics. They have been created for research purposes from the Integrated Data Infrastructure (IDI) which is carefully managed by Stats NZ. For more information about the [IDI and/or LBD] please visit <https://www.stats.govt.nz/integrated-data/>.

Access to the data used in this study was provided by Stats NZ under conditions designed to give effect to the security and confidentiality provisions of the Data and Statistics Act 2022. The results presented in this study are the work of the author, not Stats NZ or individual data suppliers.

The results are based in part on tax data supplied by Inland Revenue to Stats NZ under the Tax Administration Act 1994 for statistical purposes. Any discussion of data limitations or weaknesses is in the context of using the IDI for statistical purposes, and is not related to the data's ability to support Inland Revenue's core operational requirements.

Copyright: This work is licensed under the Creative Commons Attribution 4.0 International licence. You are free to copy, distribute, and adapt the work, as long as you attribute the work to Stats NZ and abide by the other licence terms.

Liability: While all care and diligence has been used in processing, analysing, and extracting data and information in this publication, Stats NZ gives no warranty it is error free and will not be liable for any loss or damage suffered by the use directly, or indirectly, of the information in this publication.



Crown copyright ©

[See Copyright and terms of use](#) for our copyright, attribution, and liability statements.

Citation

Stats NZ (2026). *Evaluating the use of admin data as a substitute for Household Labour Force Survey income variables*. Retrieved from www.stats.govt.nz.

ISBN 978-1-991431-13-4 (online)

Published in March 2026 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 and summary	5
Purpose	5
Summary of key points	5
Current state: HLFS income	7
What we collect	7
Earnings from employment	7
Income from government transfers and ACC	7
Future state: An admin data alternative?	8
Why move to admin data now?	8
Key considerations for using admin data to replace HLFS income variables	8
Admin income variables and Māori data sovereignty	11
Future state options.....	12
Option 1: Current state – HLFS income	12
Option 2: Linked HLFS and admin data.....	13
Option 3: Admin data only.....	16
Results	18
High-level results: All options	18
Detailed results: Linked HLFS and admin data	19
Conclusion	29
References and further reading.....	31
Appendix 1: Māori data governance and Te Tiriti o Waitangi	32
Appendix 2: Median weekly income by source, 2007–2024 June quarters.....	34
Appendix 3: Creating HLFS-comparable weekly measures from admin income.....	36
Winsorisation of extreme values	36
Fortnightly pay periods in monthly tax data	38
Annually reported tax data – self-employment and Inland Revenue tax credits.....	39
Appendix 4: Imputing ‘admin’ wage and salary income for HLFS paid employees.....	41
Working the reference week	41
Imputation method.....	42

List of tables and figures

List of tables

1 Benefits and limitations summary Option 1: Current state – HLFS income	12
2 Benefits and limitations summary Option 2: Linked HLFS and admin data.....	14
3 Benefits and limitations summary Option 3: Admin data only.....	16
4 Summary of relationship between HLFS income results and linked or admin only estimates of median weekly income by source.....	18
5 Weekly earnings from wages and salaries, self-employment and government transfers using HLFS income, linked HLFS and admin income, and admin data only	19
6 Median weekly earnings by source, by sex, June 2024 quarter	25
7 Median weekly earnings by source, by age group, June 2024 quarter	25
8 Median weekly earnings by source, by ethnicity, June 2024 quarter	26

List of figures

1 People in the HLFS sample receiving income by source of income, HLFS and IR, unweighted	20
2 Mean and median wage and salary hourly earnings, HLFS and linked IR estimates, 2007–2024 June quarters.....	21
3 Mean and median self-employment hourly earnings, HLFS and linked IR estimates, 2007–2024 June quarters.....	22
4 Median hourly earnings by sex and gender pay gap, HLFS and linked IR estimates, 2007–2024 June quarters.....	23
5 Weighted distribution of hourly earnings, HLFS and linked IR estimates, June 2024 quarter	24
6 Median weekly earnings by industry (main job) and source of income information, June 2024 quarter	27
7 Median weekly earnings by occupation (main job) and source of income information, June 2024 quarter	28
8 Median weekly income by source, 2007–2024 June quarters	34
9 Illustration of how weekly income could be skewed (fake data)	36
10 Illustration of full weekly income before and after winsorisation (all fake data)	37
11 Number of wage and salary earners by source, 2007–2024	41

Purpose and summary

Purpose

Evaluating the use of admin data as a substitute for Household Labour Force Survey income variables explores whether admin data¹ could be used, alone or linked to the Household Labour Force Survey (HLFS), to create HLFS income survey measures.

Summary of key points

Stats NZ's strategic priority of becoming an organisation that uses administrative (admin) data first relies on evaluation of current measures against admin alternatives so potential changes are well understood. Even where the underlying admin data quality is high, as is the case for income data, comparisons need to be created and shared to determine whether newly derived measures based on admin data would be timely, high-quality, and comprehensive enough to meet customer needs.

Our comparison of HLFS income measures with those reproduced using a linked HLFS and admin income dataset, and a full-coverage admin income dataset, indicate that:

- current HLFS income measures are high quality and highly comparable with admin data results where substitution is viable
- linked HLFS and admin data for wage and salary, government transfer, and ACC income appear to be viable substitutes for HLFS income measures
- unlinked admin data is not a viable substitute for HLFS hourly income measures
- self-employment income records are not timely enough for replacement using linked or standalone admin data.

We recommend the following two future state options are considered further:

- **Option 1:** Continue with the current state of collecting HLFS income measures through survey questions in June quarters.
- **Option 2a:** Shift to a linked HLFS and admin data model in which wage and salary, government transfer, and ACC income are provided through admin data, while self-employment continues to be collected in the HLFS.

Option 1 will remain the default position if there is no further work on option 2a.

Work towards option 2a will need to include:

- customer consultation (to confirm whether linked admin income data would be a viable alternative)
- a project to build and test how linking these sources would work in a production environment, along with further testing of results

¹ Referred to interchangeably throughout as 'admin data', administrative data refers to information captured for operational purposes that can then be turned to statistical use. In this case, the admin data used includes tax information from Inland Revenue as well as records of government transfers, such as main benefit receipt, superannuation, ACC, etc, to provide an indication of individual income. This was accessed under research conditions in the IDI using the October 2024 refresh.

- adjustments to HLFS income survey content.

We recommend one year minimum in which both options are run simultaneously before introducing any changes that would affect the collection of HLFS income survey questions.

Current state: HLFS income

This section summarises the data currently collected in the HLFS income module in June quarters.

What we collect

The HLFS currently collects data on wages and salaries, self-employment income, and government transfers in June quarters. Published series include mean and median weekly earnings/income (for each income source) and mean and median hourly earnings (for employment income).

The HLFS is the only regular source of median hourly earnings from paid employment, a critical measure of quantity-controlled 'typical' pay for New Zealand workers.

This information can be output by all demographic and employment variables collected as part of the core HLFS content (subject to sampling error and confidentiality requirements).

Sex and ethnicity breakdowns are key wage gap indicators, while industry and occupation breakdowns are important for considering median or market wages in a migration context. Age breakdowns are key when considering minimum wage impacts, due to differences in adult and youth minimum wage thresholds.

Earnings from employment

Information on the weekly and hourly earnings from employment is collected for all respondents employed at the interview time. Respondents can talk about their income as an annual salary, hourly rate, latest pay, or profit or loss.

Results on earnings from employment are published for employees, through wage and salary earnings, and for self-employed people.

We calculate weekly and hourly income from main job or business and, where applicable, from second job or business. For income from a second job or business, total earnings are calculated by adding together main and secondary earnings.

Income from government transfers and ACC

We also collect information on weekly income from government transfers (superannuation, tax credits, paid parental leave, benefit income, and student allowances) and Accident Compensation Corporation (ACC) payments for all respondents receiving these at the interview time.

Future state: An admin data alternative?

Why move to admin data now?

Since the New Zealand Income Survey ended in 2016, admin data has been the preferred long-term option for sourcing income data for the HLFS. Admin data has the potential to improve data quality, and reduce respondent burden and collection costs. Until now, we have not progressed beyond exploring options for this work.

Several factors make it timely to proceed. First, this work aligns with Stats NZ's strategic goal of becoming an 'admin-data-first' organisation and with our key performance indicators (Stats NZ, 2025). Making greater use of admin data in social statistics, and household surveys more specifically, has been identified as an important part of this strategy (Ramsay, 2023). Replacing income questions in household surveys with admin tax data is one way we can make real progress toward this. The feasibility of using admin data has been successfully demonstrated in the Household Economic Survey and the Household Income and Living Survey. These surveys provide examples for the HLFS to follow.

We also have escalating collection costs for household surveys and growing difficulties in meeting targets for achieved sample rates and response rates. Removing some, or all, of the income questions from the HLFS would help by reducing interview times, reducing pressure on collection resources in June quarters, and freeing up staff to spend more time on hard-to-reach respondents. Reducing respondent burden – not just interview times but also in terms of the sensitive and hard to answer nature of income questions – may help address declining response rates.

The adoption of the Māori Data Governance model (Kukutai et al, 2023) and the Ngā Tikanga Paihere framework (Stats NZ, 2020b) also make this an opportune time to proceed. We now have more clarity and guidance on how to use admin income data in a way that is appropriate and respectful of Māori data sovereignty.

Key considerations for using admin data to replace HLFS income variables

While there are compelling reasons for moving forward with this work, it is not straightforward. Although there are clear advantages using admin tax data in place of survey questions, it is not designed for that purpose. Several issues need to be addressed if we are to move to an admin data model.

Paid hours

A critical issue for current users of HLFS income data is that admin data on paid hours is not yet of sufficient quality to produce hourly earnings data. Hourly earnings are used for measurement of the gender pay gap and for MBIE's annual minimum wage review. Paid hours have been collected by Inland Revenue (IR) as part of their payday filing system since mid-2020. However, provision is optional and admin paid hours data is available only for a non-representative minority of businesses.

To continue producing hourly earnings estimates we need to link tax data to HLFS data on hours worked – there is currently no suitable direct replacement. The quality of admin paid hours data should continue to be monitored, especially if collection moves from optional to compulsory.

In this research, usual hours were used to derive hourly earnings from weekly earnings, which in turn were derived from average earnings across the full quarter (or period spent receiving income from that source, if less than the full quarter).

This matches the approach taken in HLFS income for people who report their wages on an annual basis, for whom paid hours are assumed to equal usual hours.

However, employed people can also provide HLFS income information through hourly rates (in which case hours paid is collected directly) or with their latest pay (in which case paid hours are created by combining actual hours worked with any paid leave hours). Our research approach will not match these current methods of deriving paid hours.

Regular income

HLFS income focuses on regular income, instructing respondents (where necessary) to exclude bonuses. Taxable income, on the other hand, will include irregular income like bonuses, commissions, and back pay. Lump sum payments that aren't liable for ACC levies, such as redundancy payments, can be identified and removed in the PAYE data, but others will remain in our dataset.

We winsorised extreme wage and salary earning values for each June quarter's full IR wage and salary records prior to linking them to the HLFS. Winsorised data puts maximum and/or minimum values on extreme outliers, to reduce their impact on final results. In this research, anything over the 99th percentile was capped at the 99th percentile value, and anything under the 1st percentile was capped at the 1st percentile value ([Winsorisation of extreme values](#) has more detail).

Linkage errors

We consider HLFS linkage rates in the Integrated Data Infrastructure (IDI) to be acceptable, at 91.4 percent in the October 2024 refresh, but this dataset may contain 'false negatives' (records for the same person are not able to be linked) and 'false positives' (records are linked but actually belong to different people). This is an issue whenever multiple sources are linked, including within the IDI spine.

This may become more of a concern if subsequent productionisation projects look to link HLFS and admin income data outside of the current IDI process. Regardless, linkage rates need to be monitored and should be reported on whenever linked statistics are produced.

These linkage rates also need to be monitored for bias to ensure high linkage rates in some groups don't mask lower rates for smaller or marginalised populations. IDI bias analysis reports are shared with IDI researchers each refresh, including linkage rates and false positive rates by sex, year of birth, and ethnicity.

In the October 2024 refresh, there were no differences in linkage rate by sex. By year of birth, linkage rates tended to range from approximately 87 percent to 95 percent, with lower linkages for people in their mid-teens to mid-thirties. By ethnicity, linkage rates were:

- European – 93.18 percent
- Māori – 90.54 percent
- Pacific – 91.19 percent
- Asian – 91.79 percent
- MELAA (Middle Eastern, Latin American, and African) – 91.96 percent.

Lower link rates among non-European ethnic groups likely reflect the younger age structure of these populations compared with Europeans. More detailed bias analyses between the HLFS and the IDI spine are available on request.

Timeliness

The timeliness of tax and linked data is also an issue. IR data on wages and salaries, and government transfers is highly timely, provided to Stats NZ daily with a two (business) day lag (that is, data provided on a Wednesday will reflect any payday filings made for Monday). However, there is a significant lag in the availability of data on self-employment income. This is because self-employment income requires information from annual tax returns and can take up to 24 months before all data is available, limiting its utility as a replacement for HLFS data.

Issues with timeliness are exacerbated by solutions that require linking to the HLFS through the IDI, which updates data on a fixed refresh schedule. Though payday files are now provided to Stats NZ daily, they are rolled into monthly collections of pay periods (to mimic the old Employer Monthly Schedule) before inclusion in the IDI. The IDI is refreshed three times a year – at an October refresh, for instance, tax income data from payday filing is available up to the August month.

October is the first refresh at which HLFS June quarter data is available in the IDI. A linked solution for June quarter income production that relied on current IDI timetables could only begin being produced in an October refresh and wouldn't be published until late November/early December.

Under-reporting

Another issue with data on income from self-employment – as well as informal employment – is that tax data excludes any income not reported to IR. Self-employed people have been estimated to under-report their income to IR by around 20 percent (Cabral and Gemmell, 2018).

However, under-reporting of income can also be an issue with survey data, so it is difficult to quantify the change in quality compared with our current approach.

Reduced back series

Linked HLFS survey and admin data is only available back to 2007, rather than 1998. If used as is, it would be a significant reduction in the available back series.

However, the strong linear relationship between surveyed and admin data over that time means it should be possible (though not yet tested) to backcast key wage and salary, and government transfer published series back to 1998.

Admin income variables and Māori data sovereignty

HLFS income data is very important to Māori, iwi, and hapū. It contains sensitive data from all populations within New Zealand, including those of Māori descent and ethnicity, and is used to publish annual snapshots of incomes from paid employment and government transfers. That published data is then used to calculate ethnic and gender pay gaps; the effects of minimum and living wage reviews on workers of all kinds; and to inform fair pay agreements and reduce income inequality.

Our research evaluating the use of admin data as a substitute for HLFS income variables needed to be underpinned by a broader understanding of when, and how, it was appropriate to use admin data to produce Māori statistics.

For the current project and baseline evaluation of data quality, our researchers relied on the guidance of Ngā Tikanga Paihere – a framework of five key principles, each related to two tikanga that help researchers practically apply those principles. The principles are:

1. Have appropriate expertise, skills, and relationships with communities
2. Maintain public confidence and trust to use data
3. Use good data standards and practices
4. Have clear actions and purposes
5. Balance benefits and risks.

Our researchers are experienced with HLFS data, tax data, and the linking of HLFS and admin data sources. They also understand how users interact with HLFS income variables and the burden on respondents to work through relevant questions.

Results below have been openly and honestly presented – both where admin data could perform well as a substitute for HLFS income and where survey data remains a necessity for publication of timely and high-quality results. We welcome comments and critiques of our approach – this is an initial evaluation, and there is room in subsequent iterations to adapt our methodology and areas of investigation.

The results are low risk for publication, comparing already published results with similar values produced using admin data or with linked survey and admin data. However, we ensured the highest level of security, consistently abiding by the strongest levels of confidentiality required by any of our underlying data sources.

There are also important potential benefits. This research is the first step in considering whether we could replace HLFS income variables with admin data. Additional work would be required to conclusively prove the feasibility of such a change, including reproducing results following another HLFS income release in production type conditions. Potential rewards include lower regular respondent burden, higher quality income estimates, and the ability to continue improving how and when we produce HLFS income measures.

Based on these Ngā Tikanga Paihere principles, we believe this project has made balanced and appropriate use of admin data in the IDI.

We also used the 2023 Māori Data Governance (MDG) model when considering the potential future of this project. [Appendix 1: Māori data governance and Te Tiriti o Waitangi](#) has more detail.

Future state options

Our evaluation of admin data as a potential substitute for HLFS income variables investigated the following options.

Option 1 (Survey): Continue with the current state – all HLFS income measures collected through survey questions.

Option 2 (Linked): A linked HLFS and admin data model, using the HLFS sample as the base population for which income estimates are produced, but replacing variables collected in the income supplement with admin data from IR.

- **Option 2a:** Wages and salary, government transfer, and ACC income provided through admin data while self-employment continue to be surveyed.
- **Option 2b:** All income sources (wage and salary, government transfer, ACC, and self-employment) replaced with linked admin data.

Option 3 (Admin only): Standalone admin income data used to create measures currently produced through HLFS income questions. This had no connection to HLFS survey data, trading a full coverage population for a reduced ability to investigate income by person and job demographics, including hours worked.

Option 1: Current state – HLFS income

Income data is currently produced through HLFS questions for respondents on income from select sources – specifically wage and salary earnings, self-employment income, government transfers (such as, student allowance, benefits, tax credits, and superannuation and pensions), and ACC income compensation.

As this information is collected alongside core HLFS questions, income from paid employment can be linked back to job data. This limits the number of questions asked (for example, skipping questions on self-employment income for respondents who were not self-employed) and allows variables like hourly earnings to be derived from weekly income through HLFS questions on hours worked. It also means questions about the respondent only need to be asked once to produce key labour statistics, including those on income, by demographics of interest like age, gender, ethnicity, disability status, or region.

Table 1

Benefits and limitations summary	
Option 1: Current state – HLFS income	
Benefits	Limitations
Relatively short question set – approximately 6 to 7 additional minutes per household in quarters without system challenges. Key demographic breakdowns available. Designed to match ideal concepts. Updatable alongside HLFS core survey content (eg, when new sex/gender guidelines are introduced).	Sample limitations inherent to sample surveys – decreased precision of estimates. Potential for respondent error, including proxy respondents. Although short per household, the cumulative time of income questions are significant across approximately 15,000 households

No linking required through the IDI for production. Comparable time series back to 1998. No barriers to ongoing production, beyond those faced by the core HLFS.	
---	--

Method for testing

Option 1 simply recreated published HLFS results within the IDI. This set the baseline for all subsequent approaches and reduced the risk of errors going unnoticed in combined approaches.

This used all responses provided to the HLFS – even those unlinked to the IDI spine – and included imputed income and hours data, as used in production.

If applied correctly using June 2024 quarter weights, the results should perfectly match those published as part of the June 2024 quarter release. Results were produced for 2007 to 2024 following the October 2024 refresh – those were the available June quarter HLFS datasets in the IDI at the time.

Option 2: Linked HLFS and admin data

Option 2 used the HLFS survey as the base population by which income estimates are produced, replacing variables collected in the income module with those from admin data from IR.

Linking the HLFS sample to the IDI spine, and therefore to IR information for respondents, means that estimates can easily be produced on income by demographics of interest in the HLFS, like age, gender, ethnicity, disability status, or occupation. HLFS usual and actual hours worked questions can be used to create estimates of hourly earnings for income from employment. This linking also limits the population to the HLFS working-age population, excluding those with taxed income who are outside the target population of the usually resident, non-institutionalised population aged 15 years and over.

Theoretically, though not investigated here, this approach could produce more detailed and frequent income information than is currently collected in the HLFS. For example, income while on paid parental leave could be published using linked admin data, which would fill a valuable information gap in the current HLFS. Similarly, benefit income could be split by type of benefit and superannuation could be split from veterans' pension income, though the quality of these estimates would depend on sample size.

Admin income data could be attached to the core HLFS each quarter without additional respondent burden, allowing for more frequent measurement of average incomes. This would be valuable for more precise measurement of the speed of wage and salary increases and to improve estimation of annual salaries, something regularly requested but unable to be officially calculated using a single quarter's measure.

This option was split into two sub-options, depending on income sources replaced:

- Option 2a – wages and salary, government transfers, and ACC income replaced
- Option 2b – all sources collected (wages and salary, government transfers, ACC, and self-employment) replaced.

Wages and salary, government transfers, and ACC income are all available sooner in the IDI than HLFS quarters. However, there is a significant time lag on admin sources of self-employment data.

As expressed in Stats NZ's 2023 evaluation of the quality of admin-sourced income data:

With the current IDI refresh cycle, income data is available in the IDI around four months after the end of the tax year for data that is collected by agencies on a monthly or more frequent basis. This includes income from wages and salary, New Zealand superannuation, and working-age benefits. There is a longer delay of around 16 and up to 24 months before all income derived from annual tax returns is available. This includes self-employment [income] (Welsh, 2023).

Option 2a did not reduce as much income content from the HLFS but produced relatively timely results with little effect on quality.

Option 2b removed all HLFS income content but results that included self-employment income would be far less timely and have lower quality data in recent years.

Table 2

Benefits and limitations summary	
Option 2: Linked HLFS and admin data	
Benefits	Limitations
Reduced or no income questions – benefits respondents and collection staff. Key demographic breakdowns available. All current measures available. Potential to expand current income breakdowns without increasing respondent burden. Updatable alongside HLFS core survey content (eg, when new sex/gender guidelines are introduced). Links to HLFS provide ability to restrict to HLFS sample.	Publication lag based on IDI refresh cycle. Sample limitations inherent to sample surveys – decreased precision of estimates. Potential for linkage error. Require modelling to produce estimates prior to 2007. Can include irregular income. Quarterly averages may smooth genuine variability within quarterly income. Excludes any income not registered with IR. Relies on consistent admin data collection – changes to reporting requirements could impact results. Restricting to the HLFS sample excludes valid admin income information that could be used for full coverage results.

Option 2a: Wages and salary, government transfers, and ACC

Method for testing

Option 2a aimed to recreate published HLFS results using admin income variables for:

- wages and salary income

- government transfers (calculated as the sum of student allowance, Work and Income payments, Inland Revenue tax credit, and superannuation and pension income)
- ACC compensation.

These variables were attached to the base HLFS sample each June quarter to reproduce weighted mean and median weekly and hourly earnings results within the IDI. We compared results to those produced using option 1 (current state). We limited our work and income transfers to taxable benefits, but plan to include non-taxable benefits in subsequent iterations.

After the October 2024 refresh, data was available from 2007 to 2024 for both HLFS and most replaced admin income variables in option 2a.

The only exception was IR tax credits, which are reported each financial year. From our testing, receipt and quantity of tax credits were steadier year on year than self-employment income, and could provisionally represent the upcoming June quarter in a financial year. Tax credits are published as part of a broader government transfers total and make up a relatively small proportion of income from this source, so minor fluctuations have limited impacts on published results.

In this model, self-employment income alone would continue to be collected directly from HLFS respondents. All other information currently available through HLFS income would be replaced with either admin income information, core HLFS variables, or statistically imputed data.

Option 2b: Wages and salary, government transfers, ACC, and self-employment

Method for testing

Option 2b aimed to recreate published HLFS results using admin income variables for:

- wages and salary income
- government transfers (calculated as the sum of student allowance, Work and Income payments, Inland Revenue tax credit, and superannuation and pension income)
- ACC compensation
- self-employment income.

These variables were attached to the base HLFS sample each June quarter to reproduce weighted mean and median weekly and hourly earnings results within the IDI. We compared results to those produced using option 1 (current state). We limited our work and income transfers to taxable benefits, but plan to include non-taxable benefits in subsequent iterations.

The only difference between results from options 2a and 2b should be self-employment income and, subsequently, income from paid employment and total income (each of which include self-employment income).

In this model, no income information would be collected directly from HLFS respondents. All information would be replaced with either admin income information, core HLFS variables, or statistically imputed data.

Option 3: Admin data only

If it were possible, creating HLFS income measures using admin data alone would be the best option in many ways. It would introduce flexibility in production timing, remove survey and linkage errors, and completely remove the need for HLFS collection of income information.

However, at present, the production of HLFS measures is severely limited by the lack of comprehensive collection of paid hours, meaning data on hourly earnings cannot be produced. In addition, data on incomes is needed by variables like occupation, which cannot be produced using admin data alone.

While admin-only estimates were produced for data confrontation where possible, it was not considered a feasible option for full replacement. A shift to solely producing income estimates using tax data, without linking to another source, would only replicate a limited subset of current HLFS income measures. We believe this would not meet key user needs.

Table 3

Benefits and limitations summary	
Option 3: Admin data only	
Benefits	Limitations
Removal of HLFS income questions – benefits respondents and collection staff. No survey errors – sample errors, respondent errors, etc. No linkage errors if tax data alone used. Timelier than HLFS income for income from most sources. Potential to expand on current income breakdowns. Can be produced up to monthly for wage and salary, and government transfers income. Captures jobs outside the HLFS working-age population (eg, those held by people living and working in New Zealand for less than 12 months). Could be produced outside the IDI	No hourly earnings, due to insufficient data on hours. As a result, no gender pay gap or minimum wage earner calculations. Limited demographic breakdowns available, especially if tax data alone used. Reduced timeliness for self-employment and tax credit income. Can include irregular income. Excludes any income not registered with official sources. Relies on consistent admin data collection – changes to reporting requirements could impact results

Method for testing

Option 3 aimed to recreate published HLFS results using admin data, primarily from IR (so excluding non-taxable government transfers). This was done to compare with published and linked approaches, and determine whether admin data alone could be used in place of surveyed data for measures which can be produced using tax data alone (that is, mean and median weekly incomes by source). We wanted to prioritise standalone admin data, so focused on results that can be produced using single sources (for example, admin tax data) and did not use additional information found on the IDI spine or restrict our population further.

No weights were used in the creation of these measures, since this was based on all admin records for a given period. This period aligned as closely to HLFS June reference quarters as possible for comparison, though if this option were adopted, that could shift to improve estimate frequency and/or provide full-year coverage.

Results

High-level results: All options

Estimates of median weekly income by source for each of our three main options (HLFS income, linked HLFS and admin income information, and solely admin data) showed similar levels and changes over time among all measures except self-employment income.

Table 4

Summary of relationship between HLFS income results and linked or admin only estimates of median weekly income by source				
Comparison with HLFS estimates				
Income source	Correlation coefficient, 2007–2024		Difference ¹ (\$), latest comparable data	
	Linked	Admin only	Linked	Admin only
Pensions	0.999	0.999	-1	-1
Benefit receipt	0.919	0.921	22	22
Inland Revenue	0.924	0.900	-34	-45
Student allowances	0.987	0.986	5	-11
ACC claims	0.953	0.941	135	136
Self-employment ²	0.672	0.633	325	579
Wages and salaries	0.994	0.999	42	-43
1. A negative value indicates that the linked or admin data is lower than HLFS income estimate. 2. Only available to 2023. Source: Stats NZ				

Correlation coefficients show the strength of the linear relationship between two series – a value of 1 or -1 indicate perfect correlations (positive or negative), while the closer the value is to 0, the weaker the relationship is.

Most income sources had correlation coefficients ≥ 0.9 , indicating strong linear relationships between HLFS income, linked, and admin measures of median weekly income by source. Time series of median weekly earnings by source of income can be examined in graph form (see Appendix 2).

Self-employment income stood out with the weakest relationship between HLFS income results and linked or admin only results, at 0.672 and 0.633 respectively. This was also reflected in the large absolute difference between all three options for median weekly earnings from self-employment.

In absolute terms, the difference between ACC claims from HLFS income and admin data sources was also large (over \$100 per week), though there remained a strong linear relationship between HLFS results and linked or admin results. As this is reported directly by ACC and entirely taxed, admin records of income

amounts are more likely to be correct than respondent provided data. HLFS respondents may have misunderstood the source or quantity of their ACC income (for example, reporting ACC income as paid employment income, after-tax income as pre-tax, or 100 percent of income as coming from ACC or employer where this should be an 80:20 split).

Table 5 summarises pensions, benefits, tax credits, and student allowances into government transfers, showing mean and median weekly incomes from our three broadly published sources for the most recent comparable June quarter available (2023 for self-employment income and 2024 for all other sources).

Table 5

Weekly earnings from wages and salaries, self-employment, and government transfers using HLFS income, linked HLFS and admin income, and admin data only				
Weekly income (\$)	Option 1: Current state	Option 2a: Linked wages, salaries, and government transfers	Option 2b: Linked income from all sources collected	Option 3: Admin data only
Mean wages and salaries	1510	1556	1556	1461
Mean self-employment ¹	1264	1264	1609	1477
Mean government transfers	438	448	448	453
Median wages and salaries	1343	1385	1385	1300
Median self-employment ¹	767	767	1346	1092
Median government transfers	453	460	460	460
1. Only available to 2023.				
Source: Stats NZ				

Detailed results: Linked HLFS and admin data

In addition to the mean and median weekly earning analysis which can be performed for all options, further analysis was done for linked HLFS and admin data:

1. Proportion of HLFS respondents receiving income from each source.
2. Hourly earnings by source, including impacts on the gender pay gap.
3. Median weekly earnings by source; by sex, age, ethnicity, full-time status, industry, and occupation.

Proportion of HLFS respondents receiving income from each source

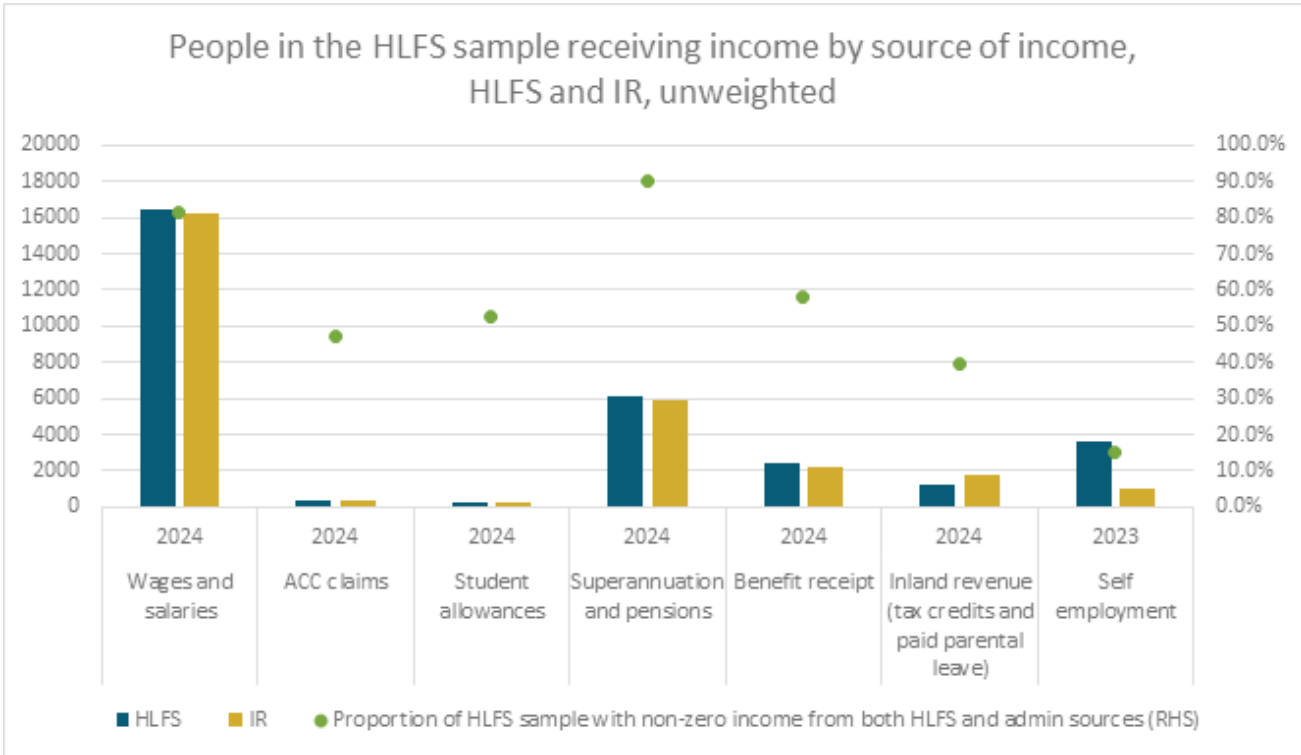
Of the income sources collected in HLFS income, wages and salaries was the most commonly received, followed by superannuation and pension income, and self-employment income.

A similar number of people were linked to income from each source using tax data for all sources except self-employment income, which was substantially less likely to be linked.

Wage and salary, and pension reporting matched between HLFS income and tax data for most HLFS respondents. Of people in the HLFS sample who received superannuation income from either source, 90.3 percent reported pension income in HLFS income and had matching tax records.² Wages and salaries also had a high overlap rate – 81.6 percent of those who had wage and salary information from either source both self-reported and had tax data.

Most other sources had overlap rates of 40 to 60 percent, apart from self-employment income where there was a 15.3 percent overlap rate.

Figure 1



Low overlap rates minimise the value of linking HLFS and admin data for income from employment. Respondents with IR employment income but no employment information in the HLFS cannot be included in any production of final estimates, since they cannot be used to produce job level or hourly income information.

Respondents with HLFS employment information and no IR income, however, can be used to produce income information by imputing income based on the IR income of those who have both. In this research, we attempted to impute this missing IR income information for wage and salary earners (see Appendix 4). If option 2 is selected as a preferred approach, imputation should be a key area for consideration and improvement.

² Using superannuation as an example of how to calculate this proportion: Approximately 6,000 people reported receiving super income in the HLFS, and approximately 6,000 have IR records of receiving super. A 90% proportion indicates that approximately 5,700 people appear in both HLFS and IR; 300 in just HLFS; and 300 in just IR, as $5,700/6,300 = 90\%$ and $5,700 + 300 = 6,000$.

Income from government transfers is not connected to any employment variables in the HLFS, and therefore low overlap rates are less concerning, provided overall response numbers are similar. No income from government transfer sources was removed or imputed based on HLFS responses.

Hourly earnings by source, including impacts on the gender pay gap

HLFS income produces hourly earnings estimates for all income from paid employment – wage and salary, and self-employment.

This can be roughly replicated using linked IR and HLFS data, in which tax data on weekly earnings is divided by HLFS usual hours.

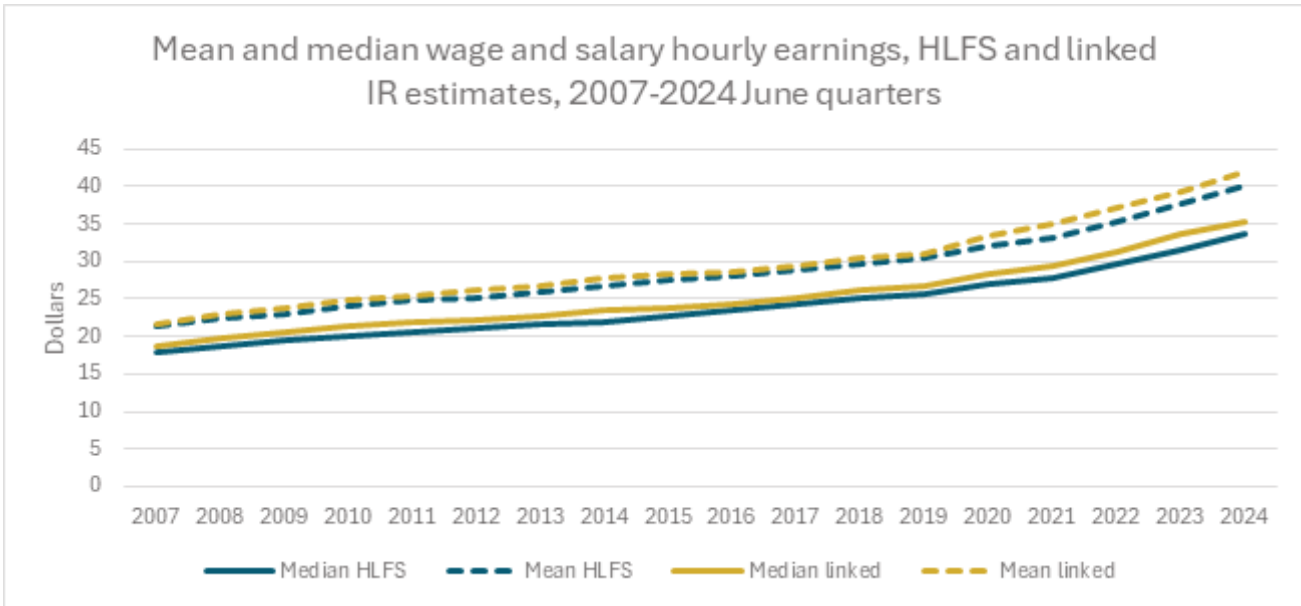
HLFS income paid hours are derived from the HLFS usual or actual hours, depending on how income is provided. There are also income specific questions about paid leave hours and hours worked at specific pay rates.

With earnings representing a quarterly average, and assuming no income specific questions about how paid hours were split, usual hours were considered the most appropriate numerator available for hourly earnings calculations.

As with weekly earnings, hourly wage and salary earnings trends aligned extremely well between our linked and HLFS income estimates.

In 2024, linked IR estimates of median wage and salary hourly earnings were \$1.72 higher than HLFS income estimates, and linked IR mean estimates were \$2.06 higher.

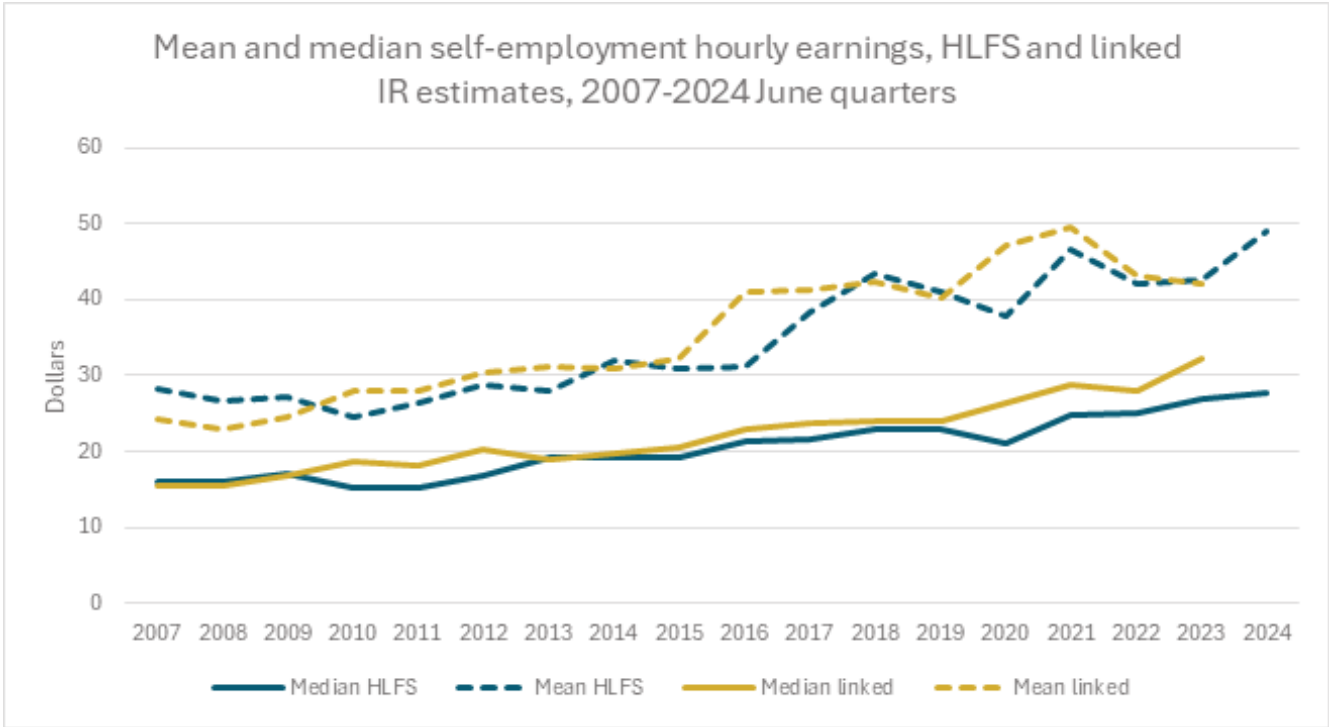
Figure 2



Given the limitations of linked self-employment IR data, much of which is only available annually, mean and median self-employment hourly earnings were closer than anticipated between linked IR and HLFS income results. The wide gap between mean and median in self-employment income was preserved, and general trends over time were similar. Self-employment income from either source was also more prone to fluctuation generally than wage and salary earnings, due to the volatility of self-employment earnings and the smaller number of people receiving income from this source.

In 2023, the latest year with linked IR estimates of self-employment hourly earnings, median estimates were \$5.46 higher than HLFS income estimates, and mean estimates were \$0.60 lower.

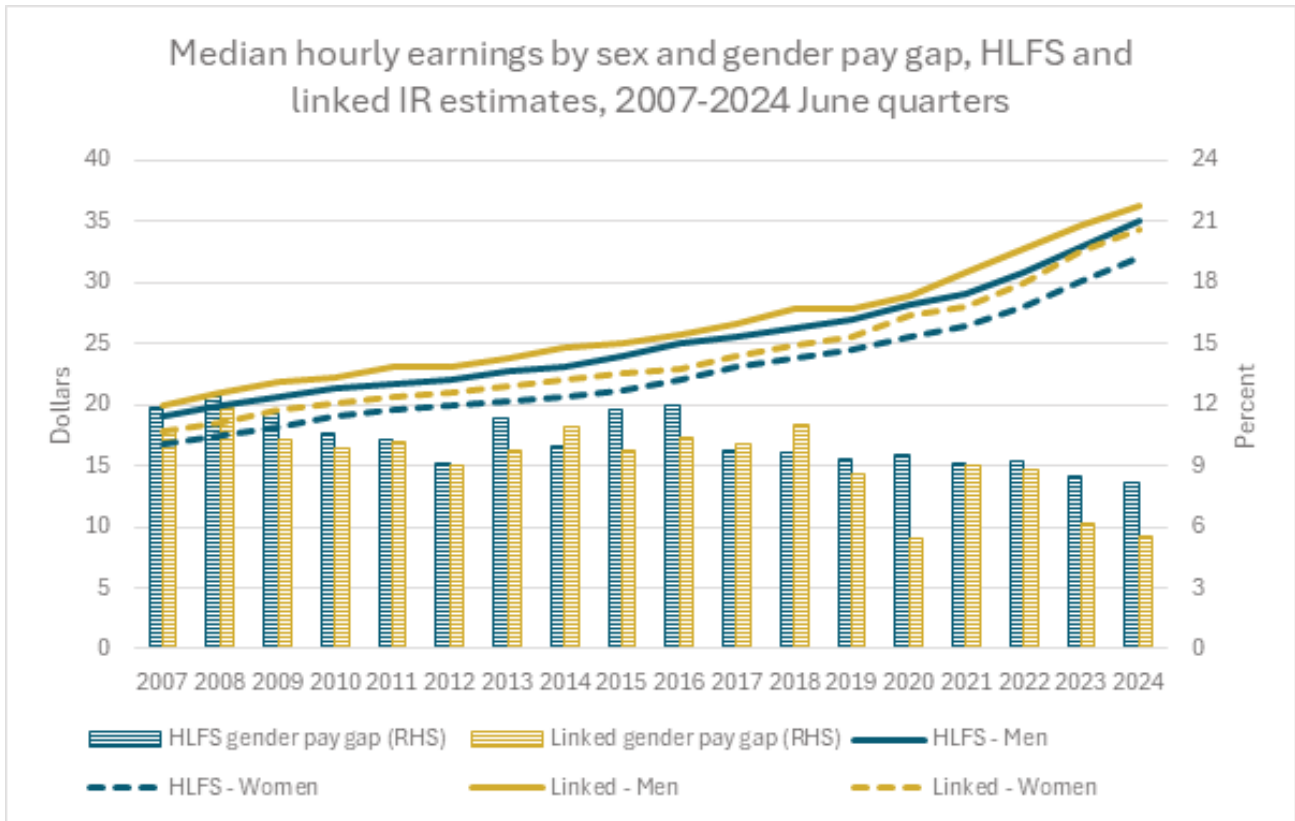
Figure 3



Median hourly earnings from wages and salaries are important to disaggregate by sex to produce the national gender pay gap – one of the key measures produced using HLFS income.

In most quarters, there was relatively little difference between the HLFS income and linked estimates, both in means and medians. The median IR pay gap was slightly lower than the HLFS income pay gap in most quarters, but usually within 1 percentage point, and the trend from 2007 to 2019 was fairly similar.

Figure 4



In recent years, however, there were two main periods of difference between HLFS income and linked estimates of gender pay gaps.

In 2020, hourly earnings for women in the linked dataset grew 7.4 percent annually while male earnings only rose 3.7 percent, driving the gender pay gap in our linked data down to 5.4 percent. Meanwhile, HLFS income showed male and female earnings rose similar amounts (4.6 percent for men; 4.4 percent for women), leaving the official gender pay gap at 9.5 percent.

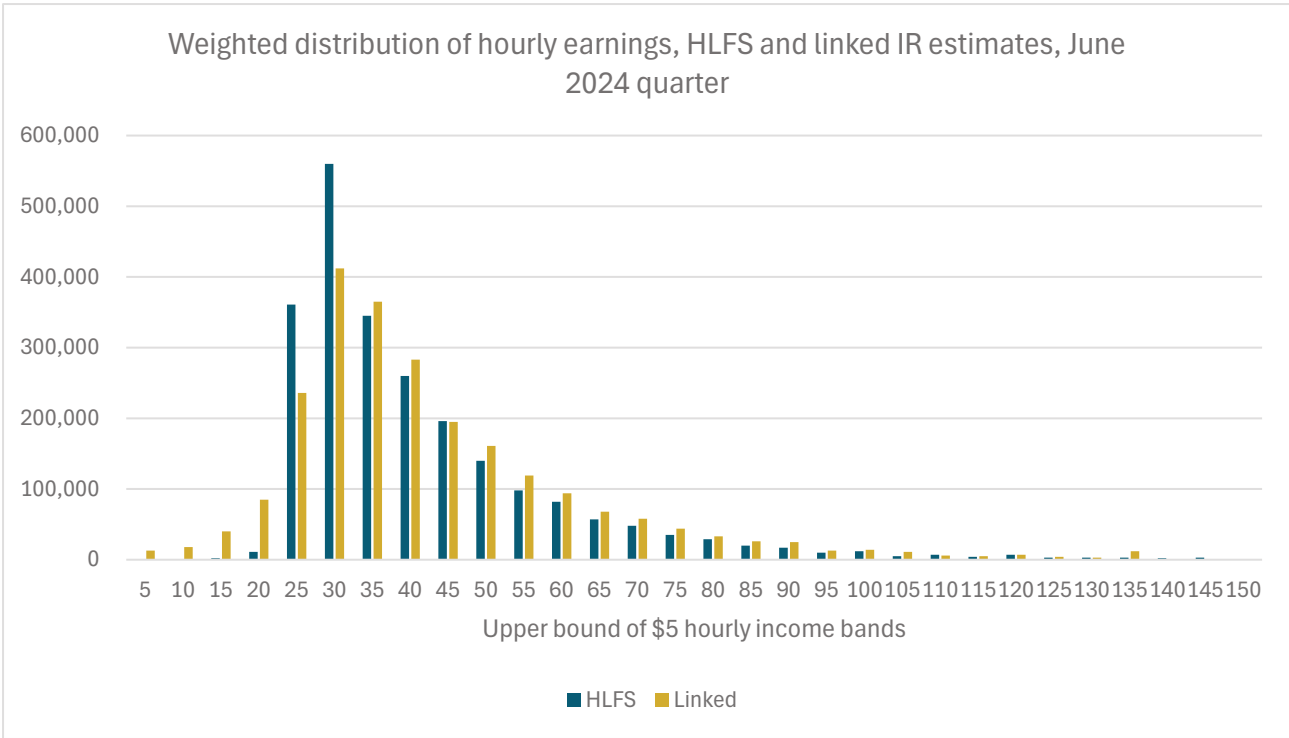
This may be related to our imputation approach, which did not impute income for employees not working the reference week. While that affected a similar proportion of male and female employees in the June 2020 quarter, we might expect higher paying jobs in female-dominated industries to have been less affected by COVID-19 restrictions (for example, healthcare, an industry that employs almost 20 percent of female employees in New Zealand) while higher paying male-dominated industries, like construction, were less likely to have continued operating. As a result, average wages for women may have been temporarily buoyed by the exclusion of a greater proportion of not working lower earners.

In 2023 and 2024, gender pay gaps also differed between HLFS income and linked estimates. HLFS income estimates of the gender pay gap were 8.5 percent and 8.2 percent, which did not statistically significantly differ from estimates of 9.7 percent to 9.1 percent in the 2017 to 2022 period. Linked gender pay gaps for the same time periods, on the other hand, fell to 6.2 percent and 5.5 percent.

The divergence in gender pay gaps came from slower wage growth for men in our linked dataset compared with the HLFS income data. Between 2022 and 2024, hourly wage and salary earnings in the HLFS rose 13.3 percent for men and 14.6 percent for women. Over the same period, linked hourly wage and salary income rose 10.7 percent for men and 14.8 percent for women.

In addition to median and mean averages, hourly earnings distributions are important to monitor given the importance of HLFS income data in assessing minimum wage impacts.

Figure 5



The distributional graph above confirms the similar median hourly income band, with the bulk of people in the June 2024 quarter receiving over \$20 and up to \$40 per hour in wage and salary earnings. The shape and spread of the graph is very similar between the two, with the same basic unimodal left-hand skew you would expect from income data, just more exaggerated in HLFS data. Outliers >\$150 per hour were excluded from analysis for confidentiality reasons, but this affected relatively few respondents.

While still a minority, the jump in estimates of people earning extremely low amounts will need further investigation. Most of it is likely the shift from HLFS paid hours to usual hours, which affect those with variable hours more than those on standard or fixed amounts. Solutions could involve exploration of use of admin paid hours – particularly if the proportion of businesses reporting them increases – or consideration of inclusion of a question about hours paid for people working non-standard hours in the reference week.

Some, however, will be genuinely low hourly income earners. HLFS respondents in similar situations are investigated to ensure this reflects genuine low-pay situations (for example, priests or musicians, who might be paid for a very small proportion of the time they actually work in a given week or partly paid in non-monetary benefits). We could not investigate to this level in this research.

Median weekly earnings by source; by sex, age, ethnicity, full-time status, industry, and occupation

We also examined key demographic groups to ensure that national numbers weren't hiding inequitable data quality outcomes. We did so for HLFS income survey data along with linked HLFS and tax data.

We considered demographics that relate to people, like sex, age, and ethnicity, for both wage and salary and government transfer income sources, with similar results for each.

Table 6

Median weekly earnings by source, by sex, June 2024 quarter				
Sex	Wages and salaries (\$)		Government transfers (\$)	
	HLFS	Linked	HLFS	Linked
Male	1458	1510	441	460
Female	1197	1246	457	460
Total	1343	1385	453	460

Table 7

Median weekly earnings by source, by age group, June 2024 quarter				
5-year age group	Wages and salaries (\$)		Government transfers (\$)	
	HLFS	Linked	HLFS	Linked
15-19	370	337	298	303
20-24	1040	1040	341	349
25-29	1283	1315	420	404
30-34	1438	1494	350	327
35-39	1516	1577	350	306
40-44	1535	1633	325	328
45-49	1573	1660	304	391
50-54	1506	1577	374	404
55-59	1477	1536	404	404
60-64	1343	1439	400	404
65-69	1283	1356	450	460
70-74	921	950	461	460
75-79	792	873	462	460
80+	448	532	480	557
Total	1343	1385	453	460

Table 8

Median weekly earnings by source, by ethnicity, June 2024 quarter				
Ethnicity	Wages and salaries (\$)		Government transfers (\$)	
	HLFS	Linked	HLFS	Linked
European	1400	1425	454	460
Māori	1227	1257	437	460
Pacific	1200	1279	406	455
Asian	1280	1353	398	404
MELAA	1343	1354	409	404
Other	1265	1296	461	460
Total	1343	1385	453	460

For each of the demographic groups we examined, the relationship between groups was preserved. For example, wage and salary earnings by age formed a gentle negative curve, with younger and older workers earning less than those in the middle age groups regardless of whether income was sourced from HLFS income or our linked data.

However, there were some minor differences between our HLFS income and linked estimates worth highlighting for further consideration:

- Men reported lower median income from government transfers in the HLFS than indicated by our linked dataset, while women's rates were similar regardless of source.
- National wage and salary earnings were approximately 3 percent higher in our linked source. However, rather than all age bands showing similar differences, linked incomes were 8.9 percent lower than HLFS income for those aged 15 to 19 years.
- Pacific peoples incomes from wages and salaries, and government transfers were disproportionately higher in our linked data than the national differences.

In general, median weekly earnings from wages and salaries, and government transfers aligned well for each of the person-level demographic breakdowns examined.

For demographics that relate to a person's job (like full-time status, industry, and occupation), only wage and salary earnings were examined.

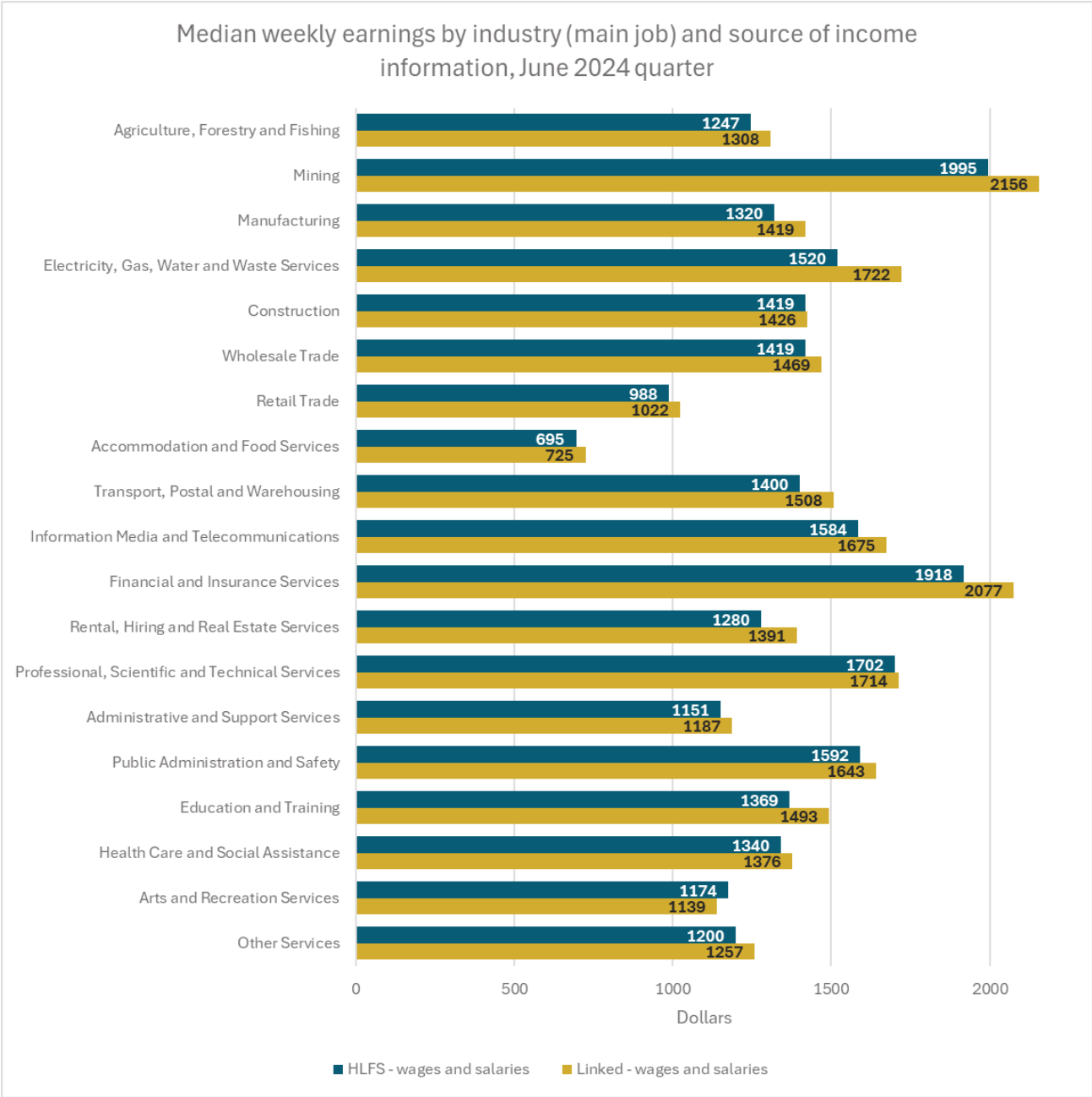
In the June 2024 quarter, full-time workers earned a median weekly wage of \$1,458 based on HLFS income and \$1,510 based on linked data. Meanwhile, part-time workers earned \$500 based on HLFS income and \$517 based on linked data.

By industry, most showed similar median weekly wages and salaries with linked earnings slightly higher than those sourced from HLFS income. There were a few exceptions with unusually high linked earnings – namely mining; electricity, gas, water, and waste services; and financial and insurance services. These are all small

industries that could be thrown by minor sample differences, but also industries in which bonuses could be more common. These bonuses are removed from HLFS estimates but may not have all been successfully cleared from our linked dataset.

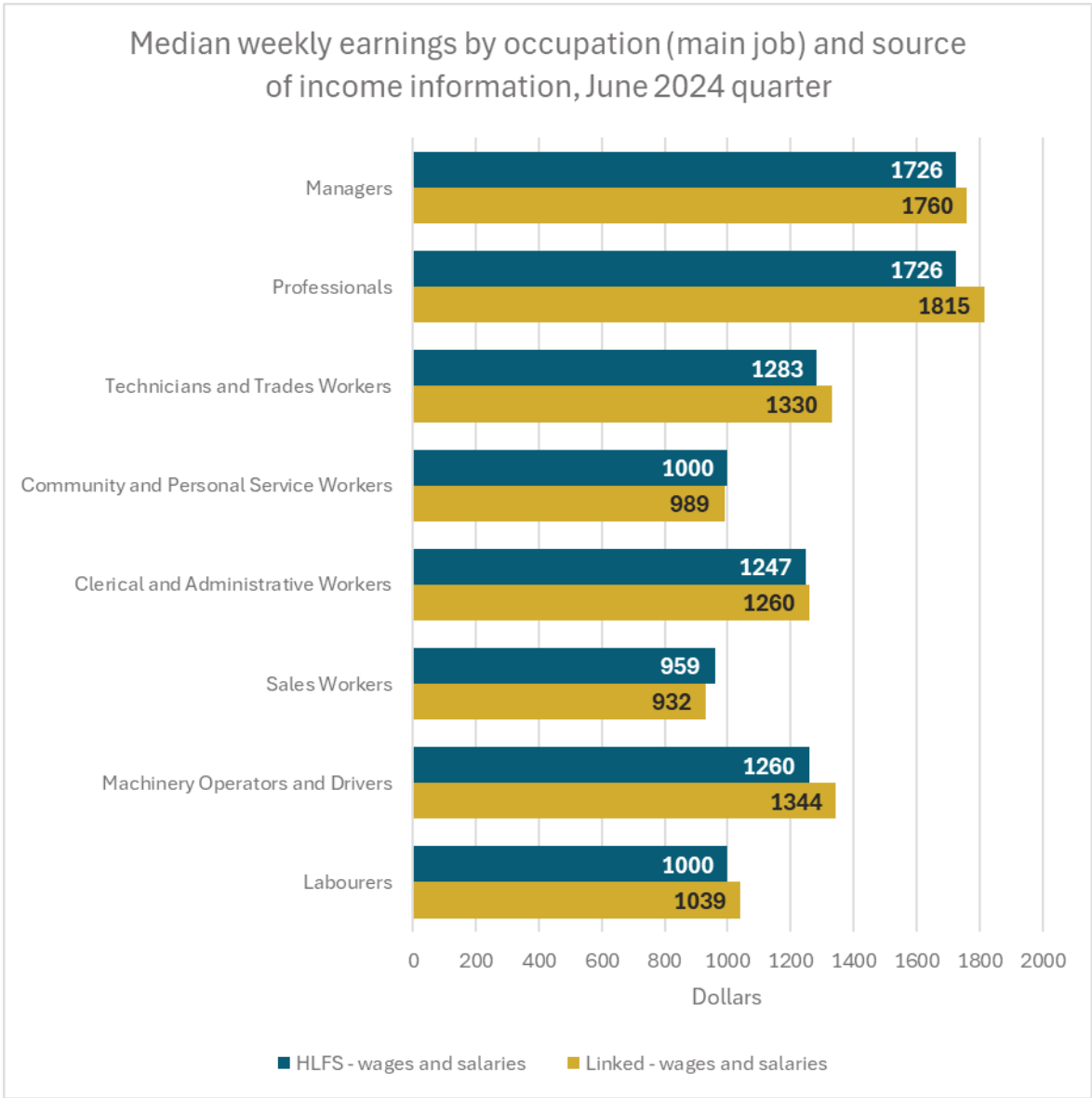
Overall, industry wage and salary differences were preserved between linked and HLFS estimates.

Figure 6



Patterns in weekly wages and salaries by occupation were even stronger than by industry. Managerial and professional occupations remained the highest paid, whether looking at linked or HLFS estimates, while sales workers; community and personal service workers; and labourers were lowest paid.

Figure 7



Conclusion

Results from our comparison of HLFS income measures with those reproduced using a linked HLFS admin income dataset and a full-coverage admin income dataset indicate that:

- HLFS income measures as currently produced are high quality and highly comparable with admin results where substitution is viable
- linked HLFS and admin data for wage and salary, government transfer, and ACC income appear to be viable substitutes for HLFS income measures
- unlinked admin data cannot be considered a viable substitute for HLFS hourly income measures
- self-employment income records aren't timely enough for replacement using linked or standalone admin data.

The future state options we **recommend** considering further are either:

- **Option 1:** Continue with the current state of collecting HLFS income measures through our rotating content.
- **Option 2a:** Shift to a linked HLFS and admin data model in which wage and salary, government transfer, and ACC income are provided through admin data while self-employment continues to be directly collected in the HLFS.

Both option 1 and option 2a are highly accurate, reliable, and relevant. They align well with desired income concepts and appear highly consistent with one another and unlinked admin data.

Option 1 would be timelier than option 2a, assuming linking using the IDI. Linking would cause a lag of approximately three months on current publication schedules (moving from late August to late November/early December). At present, it is unknown whether this lag would be prohibitive for key users.

We don't believe there would be significant differences in accessibility or interpretability between recommended options, and consistency is likely to be maintained.

Completeness would be limited in these options compared with option 3, as we are limited to HLFS surveyed respondents. We consider the trade-offs of being able to include survey content alongside income measures outweigh this limitation.

Option 2b, which relies on admin data for self-employment income, is **not recommended** due to the weak relationship between linked self-employment income and HLFS income results; difference in level compared with admin data only; and lack of timeliness, with the latest data available both a year behind other sources and of significantly lower quality than earlier years.

Option 3 excels when looking at information available from tax data alone. In isolation, it is likely a better source of weekly ACC and government transfer income (including all subgroups within this category) and at least as good a source of weekly wage and salary income when compared with HLFS income estimates. It is far more complete than our data under all other options.

However, the absence of hours information and crucial demographic breakdowns means it could not act as a full replacement for all relevant HLFS income data needs. Option 3 is therefore **not recommended**.

Option 1 will be the default position if there is no further work on option 2a.

Feedback from public consultation on option 2a will determine whether we commit to full project planning on replacing HLFS income content with admin data and how we might:

- link to admin data in our production processes
- impute incomes for people without admin income records who indicated employment in the HLFS
- produce back series for data prior to 2007
- compare all published measures with our linked project output across a full June quarter production cycle.

We recommend one year minimum in which both options are run simultaneously (as if for release production) before making any changes that would affect the collection of HLFS income survey questions.

Therefore, even if we move toward option 2a immediately, no changes to current HLFS income content will be made until at least 2027.

References and further reading

- Assendelft (1996). Project report on phase 1 of the development of the income supplement to the Household Labour Force Survey. [unpublished document]. Statistics New Zealand, Wellington.
- Cabral, A & Gemmell, N (2018). [Estimating self-employment income-gaps from register and survey data: Evidence for New Zealand](#). Working Papers in Public Finance 07/2018. Victoria University of Wellington. Retrieved from <https://ir.wgtn.ac.nz/>.
- Iusitini, L, Meehan, L, & Pacheco, G (2024). [Gender and ethnic pay gaps: An industry-level portrait of Aotearoa](#). New Zealand Policy Research Institute, Auckland. Retrieved from nzpri.aut.ac.nz/.
- Kukutai, T, Campbell-Kamariera, K, Mead, A, Mikaere, K, Moses, C, Whitehead, J & Cormack, D (2023). [Māori data governance model](#). Te Kāhui Raraunga. Retrieved from www.kahuiraraunga.io/.
- Kvalsvig, A, Gibb, S & Teng, A (2019). [Linkage error and linkage bias: A guide for IDI users](#). Wellington: University of Otago Wellington. Retrieved from www.vhin.co.nz.
- Merry, S & Wilson J (2024). Investigation into APC household income in comparison to surveys (Statistics New Zealand Working Paper). [unpublished document]. Statistics New Zealand, Wellington.
- Ramsay, D (2023). Becoming an admin data first organisation. [unpublished document]. Statistics New Zealand, Wellington.
- Reid, G, Bycroft, C, & Gleisner, F (2016). [Comparison of ethnicity information in admin data and the census](#). Retrieved from www.stats.govt.nz.
- Statistics New Zealand (2013). [Future of the New Zealand Income Survey: Report on the decision to integrate the New Zealand Income Survey into the Household Labour Force Survey](#). Retrieved from <https://statsnz.contentdm.oclc.org/>.
- Stats NZ (2017). [User guide for Stats NZ's wage and income measures](#) (4th ed). Retrieved from www.stats.govt.nz.
- Stats NZ (2020a). [Changes to the household economic survey 2018/19](#). Retrieved from www.stats.govt.nz.
- Stats NZ (2020b). [Ngā Tikanga Paihere: a framework guiding ethical and culturally appropriate data use](#). Retrieved from www.data.govt.nz.
- Stats NZ (2025). [Stats NZ's statement of strategic intentions 2025–2029](#). Retrieved from www.stats.govt.nz.
- Welsh, I (2023). [Estimating income from linked admin data: Impact of new sources](#). Retrieved from www.stats.govt.nz.

Appendix 1: Māori data governance and Te Tiriti o Waitangi

We used the 2023 Māori Data Governance (MDG) model when considering the potential future of this project. Where Ngā Tikanga Paihere is a framework for guiding ethical and responsible data use in research, the MDG model is focused on system change that respects Māori data sovereignty across the data life cycle.

Potentially incorporating admin income data into HLFS income measures is well over a year away, but some key pou and underlying guidance that would be relevant in our future planning are discussed below.

Pou 3: Data collection

3.2 Collect only what is needed. Return what isn't required

Evaluating whether new sources are possible substitutes for survey variables is a key part of ensuring we don't over-collect data. Continuing to revisit respondents – especially Māori respondents, who are regularly oversampled and revisited – when it may be unnecessary is against the spirit of collecting only what is needed. Proving it is unneeded is a key first step in changing that.

3.3 How data is collected matters

Attaining full, prior, and informed consent in data collection is crucial when strengthening Māori rights to their own data. If we were to shift away from collecting HLFS income directly toward a linked admin substitution, we would need to explicitly get the consent of respondents to use the combination of data for the regular production of official statistics. Doing so could also include capturing additional linking information (for example, IRD numbers), which would also improve linking quality.

Pou 4: Data protection

4.1 Privacy

The substitution of admin data for survey data could improve individual privacy, removing the need for people within a household to provide sensitive income data in their own home and possibly within earshot of their fellow residents. Privacy may need to be reassessed in other ways, especially if a tax-data-only approach is taken where results can be pushed below the quality suppression thresholds currently enforced by the HLFS (for example, comparing fine-grained geographic breakdowns). Any data produced would fall under Data and Statistics Act (2022) requirements to protect privacy, even from other government agencies.

Pou 5: Data access, sharing, and repatriation

5.2 Sharing

Even more explicitly, this pou makes it clear that “data linkage must be conducted with the informed consent of those involved. To do otherwise violates Māori data sovereignty principles.” Which variables are required should be carefully considered, since only variables that are necessary for planned outputs should be kept as linked data.

Pou 6: Data use and reuse

6.2 Ask the right questions

Linking HLFS and admin data would allow production of Māori-specific results of admin income measures, which would be a positive reuse of data for those of Māori ethnicity or descent who want to dig into their own results. However, results would need to be produced in a positive and mana-affirming way, taking our lead from what Māori need to pursue their own goals and reduce inequalities.

Pou 7: Data quality and system integrity

7.2 Monitoring

Any shift from surveyed data to admin data removes some control from Stats NZ to ensure the ongoing quality and comparability of data. Tax data – as used in this research – has proven high-quality and is well understood, but continued monitoring will be required to ensure data quality is being tracked and communicated in the future. This monitoring needs to be done for Māori-specific data and at a national level.

Aligning with the Māori Data Governance will be an excellent way to help ensure future work in this area supports the principles of Te Tiriti o Waitangi and The Treaty of Waitangi.

Were we to incorporate admin income data into our HLFS release, this would support:

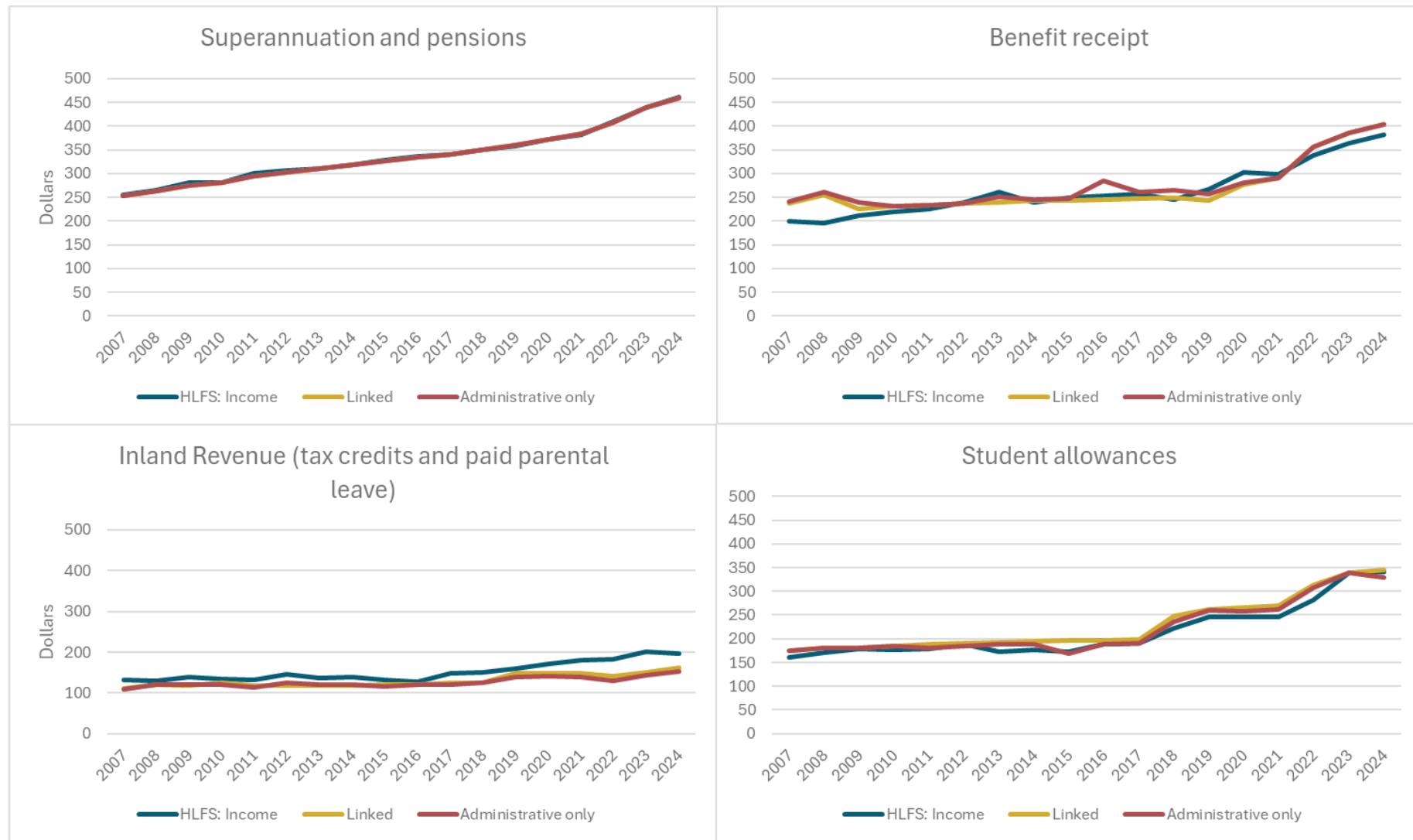
- **Article 1**, the provision of kāwanatanga and the principles of good governance and fiduciary duty, as it would ensure Māori perspectives are reflected in key income measures used in policymaking at a lower cost than surveying.
- **Article 2**, the provision of taonga and the principle of active protection, through ensuring the privacy and confidentiality of respondents when sharing results.
- **Article 3**, the provision of ngā tikanga katoa rite tahi and the principles of equity and equal outcomes, because Māori respondents – currently oversampled in the HLFS – would have their response burden reduced by the removal of HLFS income questions.

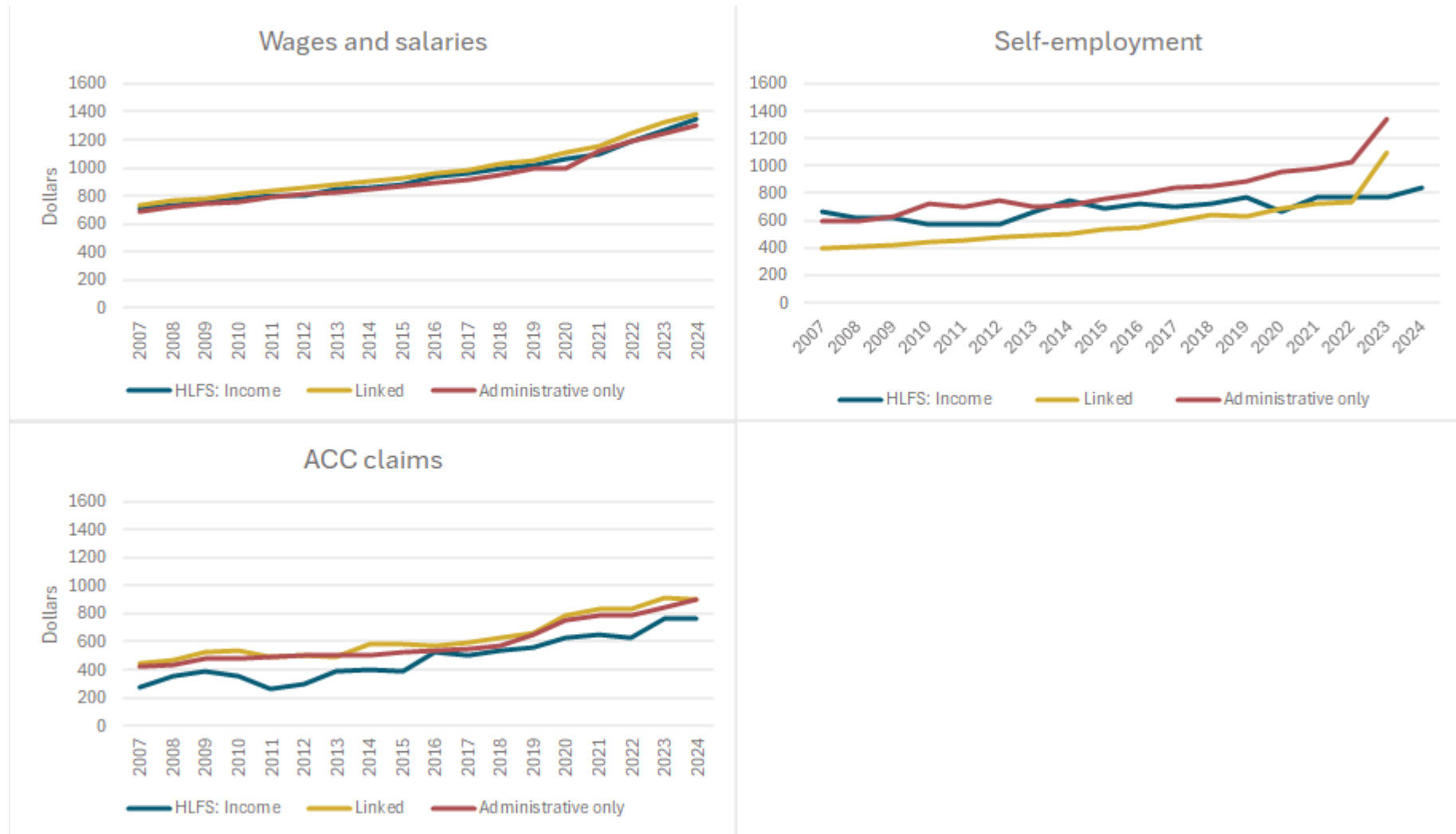
By continuing this work in partnership with Māori and in line with the MDG model, we would conduct ourselves in a way that supported:

- **Article 1**, the provision of kāwanatanga and the principle of partnership, by consulting with Māori about linking HLFS and admin income data together to produce HLFS income survey like measures.
- **Article 2**, the provision of tino rangatiratanga and the principle of self-determination, by ensuring all HLFS respondents – Māori and non-Māori – grant fully informed consent for the use of their admin income data for statistical purposes.

Appendix 2: Median weekly income by source, 2007–2024 June quarters

Figure 8





Appendix 3: Creating HLFS-comparable weekly measures from admin income

Admin information on taxable income is collected monthly and annually, including periods of irregularly high or low income.

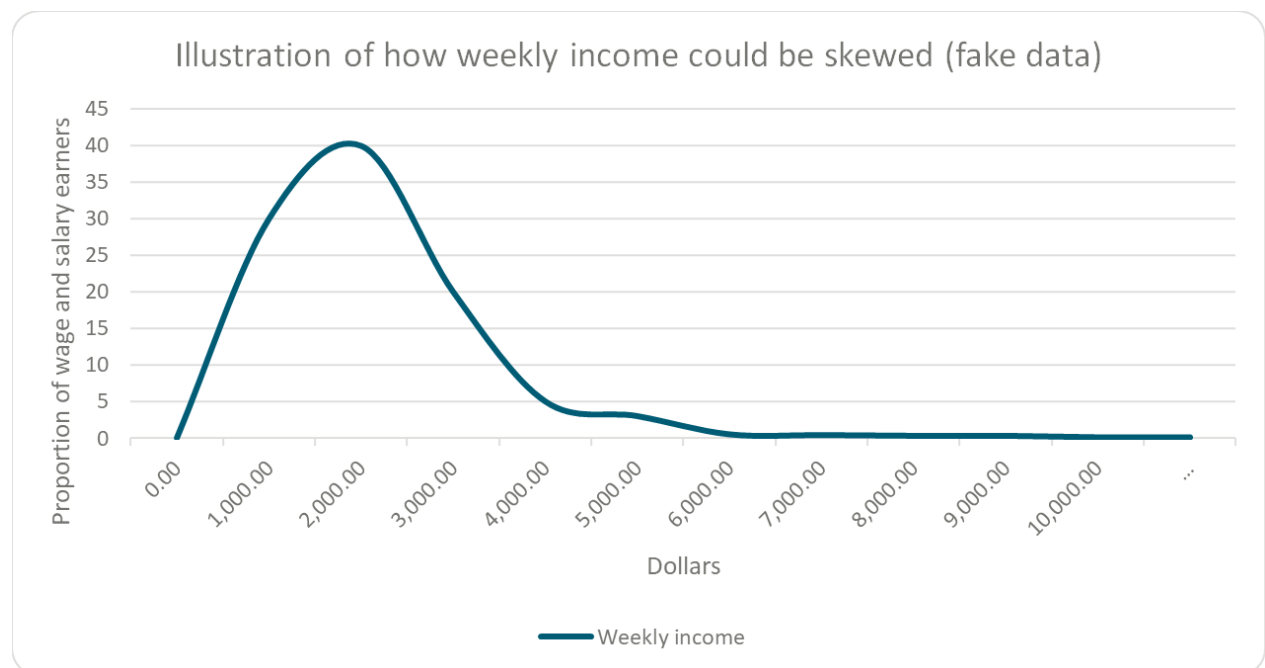
Admin income information was transformed in three ways to improve comparability with HLFS income measures:

1. Extreme wage and salary values were winsorised.
2. Income known to be provided entirely fortnightly (superannuation and paid parental leave) was transformed to approximate 6.5 pay periods in a 13-week quarter.
3. Annual tax credits were applied to the upcoming quarter.

Winsorisation of extreme values

Income measures are positively skewed – the minimum valid response for wage and salary income is zero, while there is no maximum value. The vast majority of people earn relatively similar amounts per week while a small number have much higher weekly incomes, especially if extra pay is provided during a week of interest.

Figure 9



In admin tax data, we believe the high earning tail of this distribution is genuine – they have not been provided in error and are genuinely representative of those individuals' incomes for that month.

However, two issues make this full tail less appropriate to include when attempting to replicate HLFS income.

First, this kind of fine tail can cause issues when combined with sample survey data that weights this income up to represent others for a national estimate because survey weights are a relatively blunt tool. If there were a handful of people in New Zealand regularly earning \$10,000 per week, but none of them responded to the HLFS, there would appear to be zero. If one of this handful responded to the HLFS and they had a weight of two hundred, that would lead us to calculate two hundred people earned this much. Both would misrepresent the real situation if published or used for subsequent analysis.

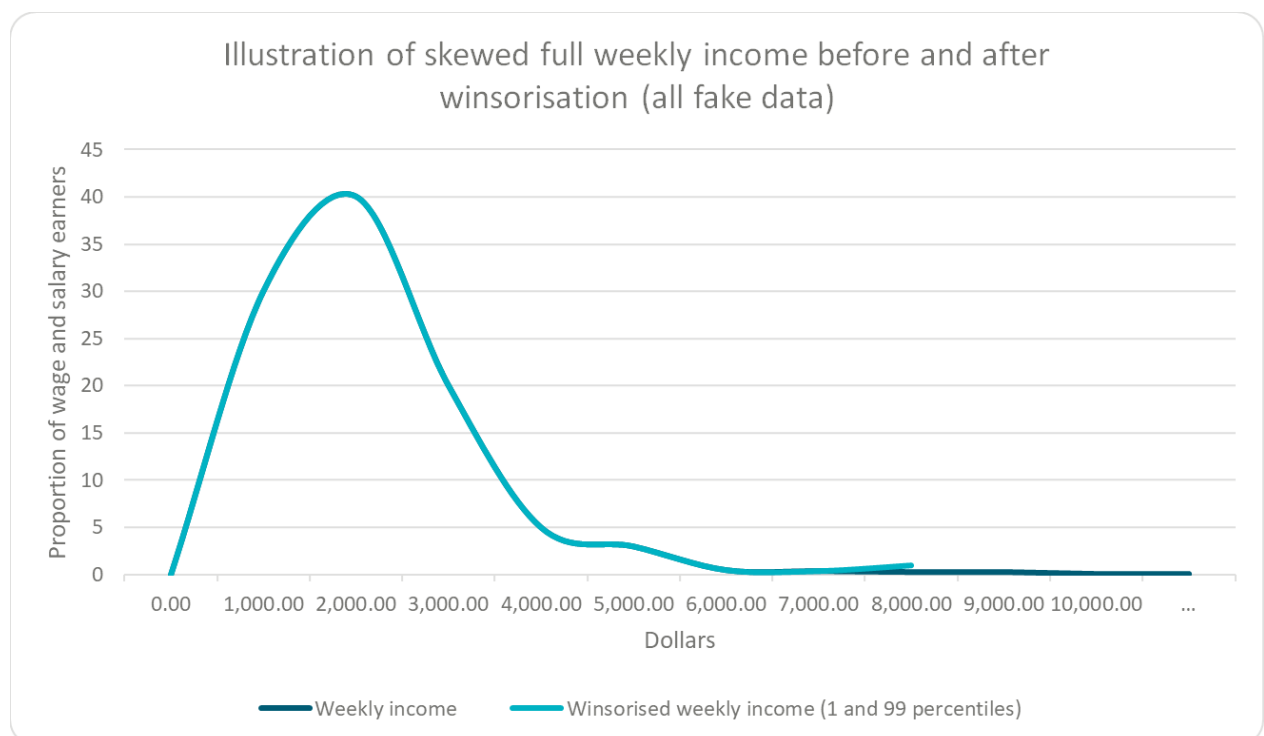
Second, HLFS income reports on regular weekly income for wage and salary earners, while high earnings can include both regular high earnings and extreme one-off bonuses or payouts when employees leave their role. There is a flag for lump sum payments in monthly PAYE tables, but this has not been used regularly since early 2019. Redundancy payments and retirement payouts, when correctly coded, should be removed as not liable for ACC levies, but leave payouts or incorrectly coded redundancy packages will be included.

We decided to winsorise weekly wage and salary data for each June quarter's full IR wage and salary records prior to linking them to the HLFS. From there, we calculate weekly and hourly income. Winsorising data means selecting a set proportion at the highest and lowest ends of the scale and setting any values above the high point or below the low point to that threshold point. When used on a positively skewed dataset, this has minimal effect on the lower end while shortening the top end tail.

Winsorisation respects that those tail values are likely genuine high earners, while setting the dataset up for more practical and realistic use. It also helps account for the potential inclusion of lump sum payments which should be removed prior to calculation of average wages and salaries.

Figure 10 shows the effect of winsorisation to the 1st and 99th percentiles on the distribution of weekly income, alongside our original illustration of weekly income.

Figure 10



A winsorised approach means there is likely to be a slight lump at the maximum weekly income value. An alternative approach could be to trim outliers above select maximum thresholds and remove their income value entirely. This would prevent the lump but makes less use of available data and would have a far greater depressive effect on mean incomes.

We believe these extremely high incomes do genuinely belong in the income distribution but want a less positively skewed dataset. We consider that trimming is less appropriate than winsorisation in this case.

Winsorisation should have no effect on medians at the national level. In extremely rare cases, winsorisation may have minor impacts when data is disaggregated, but only if those subgroups are disproportionately likely to be extremely high or low earners, that is, over half of all earners are in the lowest or highest 1 percent of the whole population.

Fortnightly pay periods in monthly tax data

Three months of tax data (April, May, and June) was used to inform most of our admin income.

This was fine for income received monthly or weekly and reported monthly. In either case, the sum of the three months income divided by 13 weeks gave an appropriate value for individuals weekly income from that role.

However, three months of data was insufficient to conclusively identify income received fortnightly. Fortnightly data should theoretically be paid 6.5 times over a 13-week quarter, but it is actually paid either 6 or 7 times depending on pay schedules.

When all data is known to be fortnightly, this can be solved relatively simply. If the income one month is approximately 150% the size of the income of the other two, it can be reduced by a third. Then, the totals of all three months' pay for known fortnightly payments can be assumed to cover 6 pay periods and multiplied up to 6.5 to cover 13 weeks.

This approach was taken for superannuation payments and paid parental leave, both of which are regularly provided fortnightly for all recipients.

However, identification of fortnightly pay periods where the pay period is not known can be more challenging, particularly when someone is paid 6 times a quarter. Without additional data, this looks identical to three monthly pay periods but is approximately 8 percent lower than it should be.

We assumed that higher and lower frequency fortnightly periods would roughly even out for wage and salary earners or those in receipt of other government income. In any given quarter, there is no reason to believe that the proportion of fortnightly earners receiving 7 sets of wages and salaries should be systematically higher or lower than those receiving 6.

However, this assumption may be false. If needed, we could use alternative approaches to better identify fortnightly wage and salary earners:

- Incorporate more months of tax data to identify fortnightly earners. On average, there are slightly under 2.2 fortnights per month. A minimum of six months would be required to definitively have at least one month with 3 pay periods for all wage and salary earners. This would double the number of records pulled in per year, with commensurate runtime increases.

- Assume that an employer-employee relationship, once identified as fortnightly, would always be fortnightly. This would likely work well for long tenure employees but will be impossible to implement for short-term or new employees. It may also cause revisions as employees are identified as fortnightly in subsequent years.

We don't believe these approaches are necessary at present, and haven't implemented either approach. This may change if further investigation reveals impacts of uneven fortnightly reporting.

Annually reported tax data – self-employment and Inland Revenue tax credits

As of the October 2024 refresh, data is available for the 2024 June quarter for the HLFS and most replaced admin income variables. Exceptions to this are:

- IR tax credits
- self-employment income from annual tax returns.

Both tax credits and self-employment income can be reported annually by financial year, which makes it challenging to use when creating June quarter estimates.

Tax credits will all be reported for the most recent financial year by the October refresh. There is no flexibility in reporting, as they come from IR. It is likely that recipients receive similar amounts in tax credits year on year.

In this research, we assumed that tax credits for any given financial year also represented the upcoming June quarter (that is, tax credits in the financial year of April 2023 to March 2024, which was available in the October 2024 refresh, also represented the June 2024 quarter).

Self-employment income from annual tax returns can be far less timely and stable year-on-year.

While some self-employment data will be in for the current financial year by the October refresh, data can have a time lag of up to 24 months after the end of the tax year. Previous research (Welsh, 2023) indicates that most data for this group is available approximately 16 months after the end of a given tax year.

Based on Welsh's assessment, we assumed self-employment income data in the October 2024 refresh should be considered high-quality up to the 2022 financial year; moderate quality for the 2023 financial year (including the June 2022 quarter); and provisional for the 2024 financial year (including the June 2023 quarter).

Income from self-employment can be far more volatile than tax credit receipt, both in terms of who receives that income and the quantities received.

We assumed that self-employment income for a given financial year should represent the June quarter it contains (for example, the 2024 financial year includes data from the June 2023 quarter) and that it would not be appropriate to cast forward, as was done for tax credit income.

There was no data for annual filers for the June 2024 quarter in the October 2024 refresh. As self-employed PAYE filers are not representative of all self-employed people, we treated 2023 as the latest valid year for self-employment income as of the October 2024 refresh, noting its provisional nature.

Another challenge when creating HLFS comparable measures from annual tax returns is no indication of the period in which income was earned. Any income earned needs to be assumed to be a full-year measure when transformed to weekly income.

This is unlikely to have a material impact on results from tax credits, as most recipients receive credits for a full year period.

For self-employment income, however, this may have a severe impact on weekly and hourly measures. It will also impact our ability to link admin data to survey respondents.

Self-employed people can spend short stints in self-employment for relatively high rates, which is then flattened by the annual report structure.

For example, someone who earned \$10,000 for two weeks work will be assigned a weekly rate under \$200 if that \$10,000 is assumed to be earned over a full year. They will also not be identified as self-employed in the HLFS if that two-week period did not overlap with the reference week for the June quarter in which they were interviewed.

There is no fix for this issue. As it stands, self-employed people with short tenures are far less likely to be linked between admin and survey data, and weekly and hourly earnings are significantly lower than they should be.

Overall, annual data is generally unsuitable to replace HLFS income measures. We believe this can be overcome with tax credit data but not self-employment income data.

Appendix 4: Imputing ‘admin’ wage and salary income for HLFS paid employees

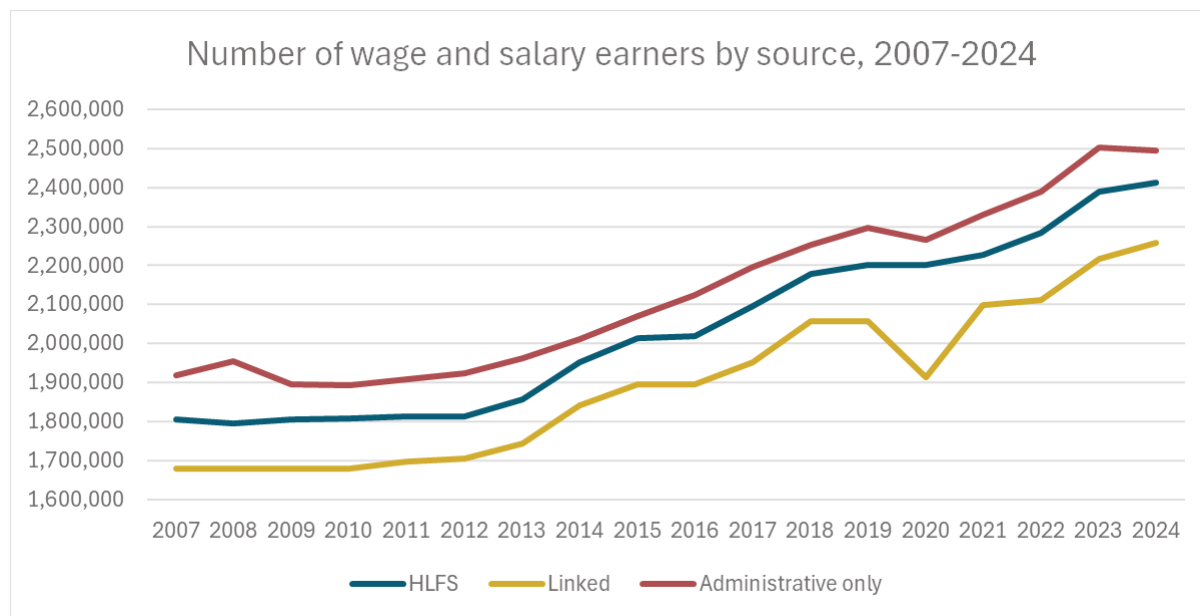
For this project, we imputed wage and salary income and hours for HLFS respondents who were paid employees and worked the reference week but had no income data.

Working the reference week

This caveat on ‘working the reference week’ recognised that two main groups of employed people have no IR income information – those who receive income and therefore should have data, and those who did not receive income that week and therefore were correctly excluded. This could include people working irregularly and not during the reference week; on unpaid leave; or receiving income from another source while temporarily absent from work (for example, ACC or paid parental leave). The assumption was made that this ‘working the reference week’ requirement would reduce incorrect imputation for those who received no income that week and therefore were correctly excluded, while still imputing incomes for the majority of those who should have received it.

However, the zero-hour limitation left the weighted count of people who received income from wages and salaries in our linked HLFS and tax admin data approach lower than either HLFS income or admin data alone, especially in 2020 when an unusually high number of employed people were unable to work due to COVID-19 restrictions. Admin data being higher than HLFS data makes sense, with no working-age population limits to wage and salary jobholders, but the additional gap between HLFS and linked (including imputed) data is a concern.

Figure 11



Some of these people may have received income from wages and salaries while not at work in the reference week, through paid leave or COVID-19 wage subsidies, and are simply excluded from our linked sample due to linkage errors, timing mismatches, misreporting income to IR, and so on.

An alternative approach could be to impute for all those identified as wage and salary earners in the HLFS who were not receiving IR wage and salary income, which would leave wage and salary earner

numbers identical in the HLFS and linked datasets. This would over-impute, as opposed to the current likely under-imputation.

Imputation method

Mirroring the imputation approach used in the current HLFS income survey, we used a rule that we would impute hours along with income, so a single donor provided both variables. This, combined with our preference to match the industry and occupation of donor to donee, meant that our donor pool had both employee information from the HLFS and wage and salary income from IR.

We used hot-deck imputation with imputation cells grouped by:

1. Quarter
2. 5-year age band
3. Industry
4. Occupation.

By imputing hours and incomes together we aimed to create a stronger relationship between income and hours, and create more plausible hourly incomes than might be expected when a donee kept their own hours and had someone else's earnings. We had limited success. The lack of prior processing on our linked dataset meant many of the donors still had unrealistic relationships between hours reported in the HLFS during the reference week, and average wage and salary tax data across the quarter.

An alternative approach was briefly explored where we imputed wages and salaries, and not hours. This used every person in the HLFS who had IR wage and salary income as our donor pool – not just those who were also identified as employees in the HLFS – and imputed based on individual characteristics only, rather than job characteristics like current industry and/or occupation.

This approach resulted in an average weekly wage and salary wage that was closer to that produced using HLFS income data and unlinked admin data, but a higher number of unrealistic hourly earnings at the individual level.

As such, we stuck to mirroring the current HLFS income imputation approach.

If we progress with option 2a, imputation should be a key area for consideration and improvement, including imputation of non-wage and salary income for unlinked respondent.