Australian Bureau of Statistics

CAPACITY TO CONTRIBUTE: INTRODUCTION TO INCOME IMPUTATION

Direct measure of income refinement working group paper

November 2020

CAPACITY TO CONTRIBUTE: INTRODUCTION TO INCOME IMPUTATION

1. Introduction

1.1. For capacity to contribute (CTC), income imputation refers to the methods used to determine a value of Adjusted Taxable Income (ATI) for those parents whose ATI is missing in the linked administrative data.

1.2. This paper describes two projects which the Department of Education, Skills and Employment (DESE) has engaged the ABS to undertake to inform the imputation strategy for the Direct Measure of Income (DMI) method for calculating CTC scores.

1.3. The first project aims to examine the fitness-for-purpose of using government payments data to derive or improve estimates of parental Adjusted Taxable Income (ATI). The second project explores the use of statistical modelling to impute for missing income values.

1.4. These projects both have the potential to inform the imputation strategy for CTC. Should both approaches prove fit-for-purpose, the derivation step would likely be applied first to derive an income for a parent with government payments information. The modelling step would then be applied to any records where the derivation step could not be used.

1.5. In addition to investigating ways to improve data quality and reduce missingness, these projects also aim to support the analysis of alternative summary statistics, such as tri-mean and mean, as part of the DMI refinement forward work program.

2. Background: missingness in CTC data and current imputation strategy

2.1. For some parents in the Address Collection, the value for ATI is missing. This can occur for different reasons. In particular, some parents from the Address Collection cannot be linked to the spine of the Multi-Agency Data Integration Project (MADIP) data asset. Some other parents can be linked, but there is no tax information available for them. This can occur if the parent was not required to submit a tax return, or if the parent lodged their tax return too late for it to be included in the linked data.

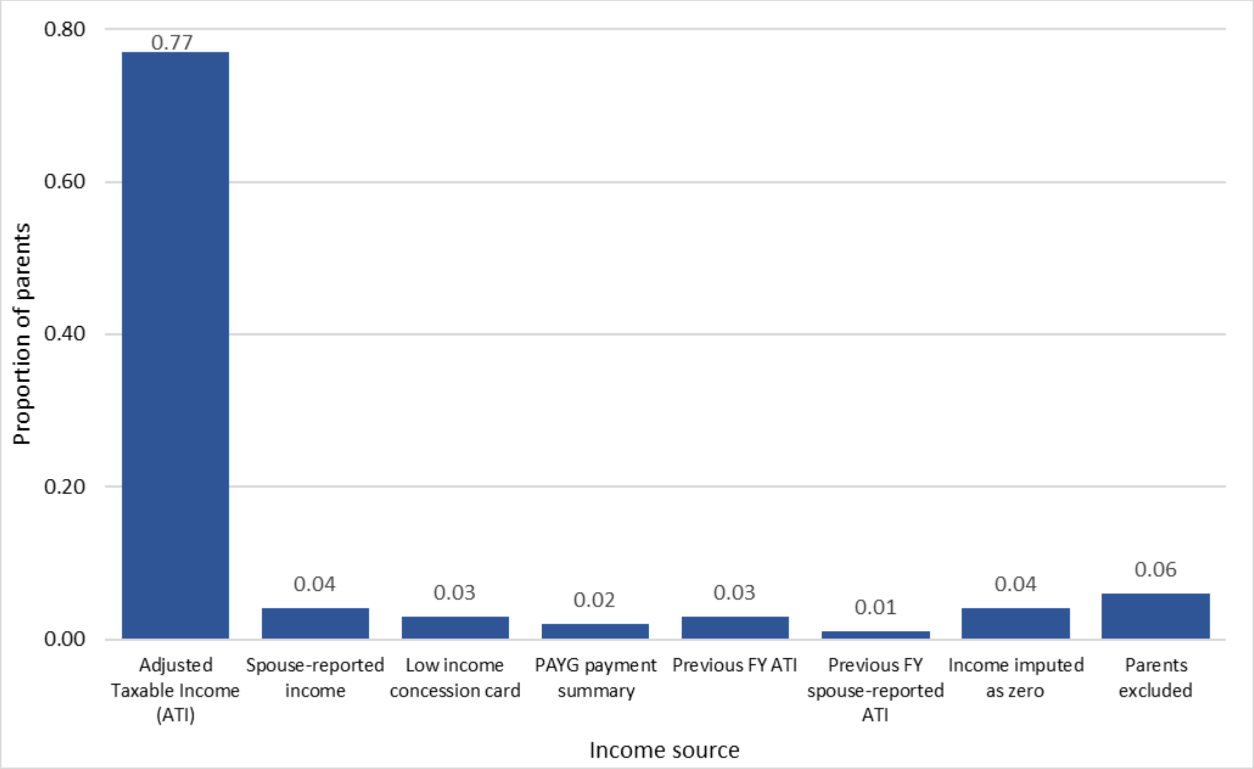
2.2. For the calculation of 2019 DMI scores, a multi-stage approach was used to impute for the missing ATI values. Alternative sources of income values, such as PAYG payment summary data, were used when available. In some cases, a value of zero was imputed for a missing ATI value, such as when:

- no other sources of income data were available but the parent was a low income concession card holder; or
- income data was available for one of a student's parents but not the other.

2.3. It should be noted that because the DMI score is calculated from the median income value for a school, the score will be relatively robust. That is, the score will not be changed if a parent whose income is low but non-zero is given an imputed value of zero.

2.4. In the calculation of 2019 DMI scores, ATI was available for 77% of parents (see Figure 1). Approximately 3% of parents had income imputed as zero due to having a low income concession card, 4% of parents had a value of zero imputed as income information was available for the student's other parent, and 6% of parents were excluded due to missingness.

Figure 1: Proportion of parents by income source used in 2019 DMI score calculation.



2.5. More information about the imputation used for the 2019 DMI scores can be found in A Data Quality Framework for the Australian Government's Direct Measure of Income for Capacity to Contribute.

3. Deriving ATI using government payments data

3.1. Data relating to a range of government payments is sourced from the Department of Social Services DOMINO Centrelink Administrative Dataset and linked to MADIP[1]. ABS will undertake preliminary data investigations to assess the fitness-for-purpose of using government payments data to derive or improve estimates of parental ATI. This includes:

- a conceptual review of the government payments included in ATI;
- analysis of the coverage of government payments data in the CTC population overall and among parents with a low income concession card flag; and

[1] This dataset was previously referred to as Social Security and Related Information (SSRI) and this name is still used in some MADIP documentation. DOMINO stands for 'Data Over Multiple Individual Occurrences'.

- assessment of the accuracy of ATI estimates derived using government payments data, such as analysis of differences between government payments data and ATI, for parents who have both.

4. Statistical modelling to estimate missing ATI

4.1. Preliminary research

4.1.1. The ABS has experience with imputing for income in its own survey data. In addition, to inform the statistical modelling for missing incomes for CTC, ABS has conducted a brief literature review into methods for imputing income. Key factors identified in the literature as associated with income are:

- industry of employment;
- geographic location;
- education level;
- age or length of tenure in employment;
- family composition; and
- disability and some illnesses.

4.1.2. It is also important to understand patterns of missing ATI values in the linked data, including any associations with other data items. An analysis of missingness is being conducted by the ABS.

4.2. Summary of the modelling approach under consideration

4.2.1. The method being developed uses data from the Survey of Income and Housing (SIH) to create a linear regression model that will predict an ATI value based on data items that appear in both the SIH data and the linked data being used to calculate DMI scores. The model parameters would then be applied to the linked data to impute values for records with a missing ATI. It is likely different sets of parameters would be used for different subgroups of people, based on patterns of missingness.

4.2.2. The SIH dataset provides a comprehensive measure of personal income, containing data items for different sources of income such as wages and salaries and government benefits. It also includes demographic data, such as whether members of a household attend a non-government school. This means the model can be developed using the most relevant subpopulation. The SIH data being used also contains data items for state and territory, and the capital city/rest of state division.

4.2.3. The accuracy of the model can be evaluated by applying artificial missingness to the SIH data in the same patterns that occur in the linked data, and comparing the predicted values with the reported values. However, the true values of ATI for all records in the linked data cannot be known, so this assessment will only be approximate. The impact of varying assumptions on the final DMI scores can also be checked, to determine how sensitive the scores are to the details of the model.

4.3. Caveats and limitations

4.3.1. Many factors influence a person’s income, yet the data available is limited. Therefore, it is possible any model developed for this project will have relatively low predictive power. Some records may still need to be excluded from the calculation of DMI scores if most other data items are also missing for that person.

4.3.2. It should be noted that the aim of this modelling is to improve the accuracy of the score calculation, not to determine an accurate income estimate for every person. If a robust statistical measure such as the median is used for the score calculation, then using a model to predict missing data may provide an improved method without necessarily having very high predictive power. Other statistical measures, such as the mean, may be more sensitive to the predictive power of the model and therefore would require a model with higher predictive power.

4.4. Other possible approaches for modelling Adjusted Taxable Income

4.4.1. Possible approaches for imputation that could be considered in future stages of this project include: alternative regression models, donor or hot deck imputation, using additional data sources (such as linked SIH data) and machine learning. These alternative methods also have associated limitations. For example, the SIH data comes from a survey and therefore will not provide an ATI value for all records in the linked dataset. Machine learning methods can be complex to apply and may not produce accurate results from a relatively small dataset. Machine learning methods can also be relatively opaque - that is, it can be difficult to determine and explain why a particular record received a particular predicted value. In contrast, the parameters from a regression model are transparent and easy to interpret and explain.

5. Timeframes

5.1. ABS will present the key findings of income imputation analysis and modelling to the working group at the next meeting. Given the timeframe for this phase of the work program, it will only be possible for ABS to develop and assess one modelling approach in this time.