Australian Bureau of Statistics

CAPACITY TO CONTRIBUTE: INCOME IMPUTATION – DISCUSSION PAPER

Direct measure of income refinement working group paper

January 2021

CAPACITY TO CONTRIBUTE: INCOME IMPUTATION – DISCUSSION PAPER

Executive Summary

Introduction

This paper presents preliminary findings for discussion, in respect of analytical projects the Department of Education, Skills and Employment (DESE) engaged the Australian Bureau of Statistics (ABS) to undertake to inform the imputation strategy for the Direct Measure of Income (DMI) method for calculating Capacity to Contribute (CTC) scores. These projects were introduced in the paper entitled 'Capacity to Contribute: Introduction to income imputation', presented by ABS at the November DMI refinement working group meeting.

The first project examines 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. Though there is a wide range of possible data sources and imputation methods that could be used, due to time constraints, this report presents the initial results from a single, initial approach based on the linear regression parameters of a model that made use of data from the ABS Survey of Income and Housing (SIH).

These complementary pieces of work are important to ensure the CTC income imputation strategy incorporates, to the extent practically possible, available information relevant to deriving or predicting parental incomes, and thus supports the robust estimation of school scores.

Summary of preliminary findings

- Government payments data provides income information for approximately 10% of parents in the 2020 CTC Address Collection.
    - For approximately two-thirds of these parents, the government payments data complements income information available in Personal Income Tax (PIT) data.
    - For approximately one-third of these parents, representing 3.4% of all parents in the 2020 Address Collection, an income amount was not available in the PIT data, but can be sourced from government payments data.
- Overall, ABS found that there was a high degree of variation in parental incomes among the CTC population, and this presents challenges for modelling those incomes. The initial statistical model developed explains less than half of the variation in parental income (adjusted R^2 value is 0.44), which is considered relatively low. Further refinement to the model may improve its explanatory power.
- Testing of the model identified large differences between actual income and modelled income for some parents. Nearly 75% of parents had a difference of greater than 20% between their predicted and actual income values.
- The impact on school scores of incorporating modelled incomes was assessed for two distinct groups. For parents who link to the Multi-Agency Data Integration Project (MADIP) spine and have Census data available, the statistical model incorporates information

available in Census, such as the person's occupation, into the predicted income value. When modelled incomes for parents who linked to MADIP and Census were included in the calculation of school scores, the majority (72%) of school scores did not change. Of the 28% of schools whose score changed as a result of including modelled incomes for this group of parents in the score calculation:
    - the majority (88%) of schools had a change in score of 1 point;
    - approximately 10% had a change of 2 points; and
    - approximately 3% had a change of 3 or more points.
- For parents who do not link to MADIP, there is little information available to incorporate into a statistical model. When modelled incomes for parents who did not link to MADIP were included in the calculation of school scores, about half (54%) of school scores did not change, while 46% experienced some change to their score. For some schools (1%), this resulted in a decline in their score of 5 points or more.

Preliminary recommendations

1. Government payments data should be incorporated into the income imputation strategy for CTC, to complement the existing data sources and provide a source of income values for parents across a range of income and labour force participation categories.
2. Given the variety of factors that can influence a person's income, a multi-stage imputation strategy should be used which incorporates available data sources, such as government payments data, to derive ATI before modelling is applied to impute for missing ATI values.
3. The initial SIH model should be further refined and evaluated, before assessing the value of this approach.
4. Due to the limited amount of information available to be incorporated into a statistical model for parents who do not link to the MADIP spine, further work is recommended to investigate approaches to imputing missing income values for this population group.

Next steps

- Further analysis of income values for members of the CTC population for whom income information is available from multiple data sources is in progress. This analysis will inform future recommendations as to how income amounts in government payments data should be incorporated into the income imputation strategy for CTC, where multiple sources of income information are available for a parent.
- Further enhancements to the initial SIH-based regression model can be made, subject to preliminary recommendations 3 and 4. This includes refinements to the model's specifications and analysis of its performance in the context of a revised imputation strategy which incorporates government payments data. If further work does not improve the model's goodness-of-fit, alternative imputation options can be considered.

Introduction

1. This paper presents preliminary findings for discussion, in respect of two projects which the Department of Education, Skills and Employment (DESE) engaged the Australian Bureau of Statistics (ABS) to undertake to inform the imputation strategy for the Direct Measure of Income (DMI) method for calculating Capacity to Contribute (CTC) scores.
2. The first project examines 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.
3. These complementary pieces of work are important to ensure the CTC income imputation strategy incorporates, to the extent practically possible, available information relevant to deriving or predicting parental incomes, and thus supports the robust estimation of school scores.
4. This paper builds on the Capacity to Contribute: Introduction to income imputation paper presented by ABS at the November DMI refinement working group meeting.

Missingness and income imputation in 2020 DMI scores

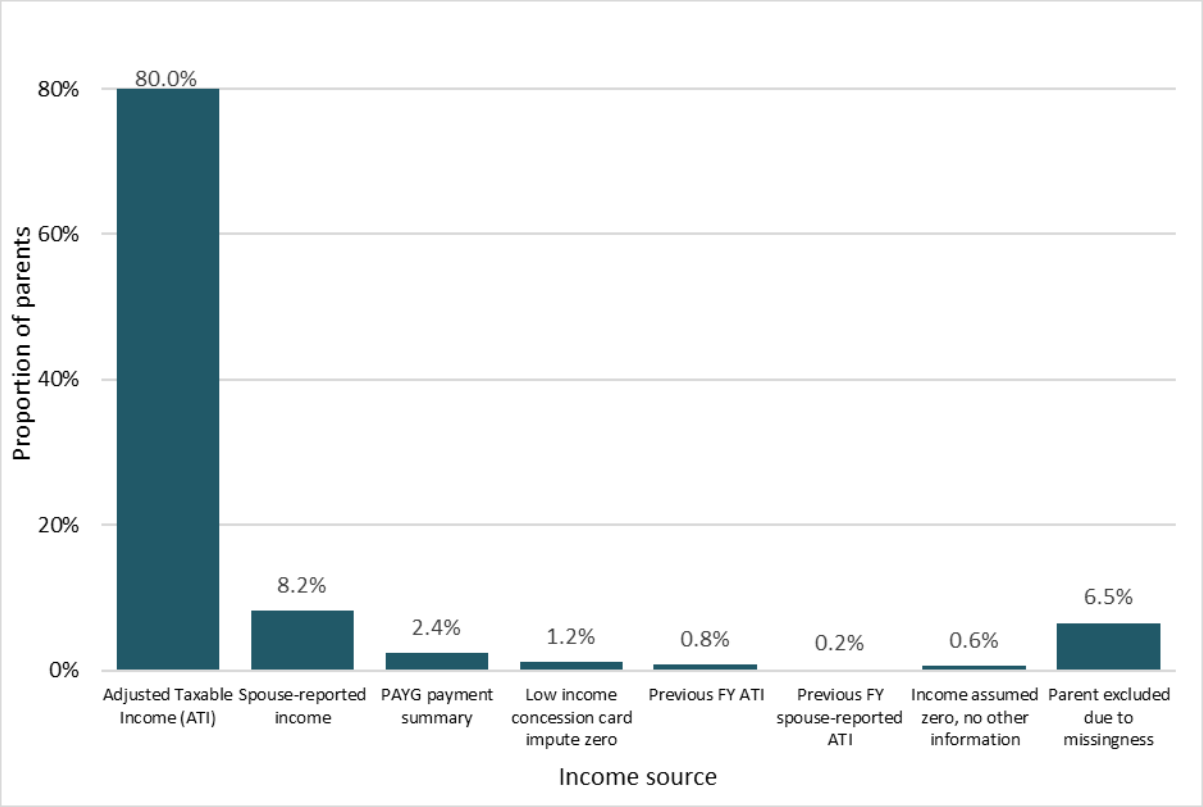
5. For CTC, income imputation refers to the methods used to determine a value of ATI for those parents whose ATI is missing in the linked administrative data available via the Multi-Agency Data Integration Project (MADIP).
6. For the calculation of 2020 DMI scores, as in previous years, a multi-stage approach was used to impute for the missing ATI values.
    - First, an income value is sought from Personal Income Tax (PIT), spouse-reported PIT, and payment summary data, in that order. Of all parents included in the 2020 Address Collection, an income value was available for 90.7% of parents[1] from these sources.
    - Second, if the above data sources are unavailable and the parent has a low income concession card flag, then zero income is imputed for that parent. This imputation occurred for 1.2% of parents in 2020.
    - Third, if no income has been assigned, an income value is sought from the previous year's PIT, previous year's spouse-reported PIT, or previous year's payment summary data, in that order. Income values were sourced for 1% of parents from these sources in 2020.
    - Fourth, if no information is available for a parent, they are:
        - imputed zero income if the student has two parents in the Address Collection and the other parent has an income value, which occurred for 0.6% of parents in 2020; or

[1] Throughout this paper, references to analysis of 'parents' refers to parent and guardian records in the 2020 Address Collection. Parent records are not the same as unique parents, as parents with children at multiple non-government schools are counted multiple times in this figure.

- excluded from the calculation, which occurred for 6.5% of parents in 2020.

7. The assignment of income sources to parents in the calculation of 2020 DMI scores is summarised in figure 1. This includes all parents, including those who did not link to MADIP.

Figure 1: Proportion of parents by income source used in 2020 DMI score calculation.



8. It should be noted that tax information is not expected to be available for all parents in the Address Collection. Some parents are not required to submit a tax return, for example if they earn no income or earn less than the tax-free threshold. ABS Survey of Income and Housing (SIH) data indicates that, among households with a student attending a non-government school, approximately 12% of parents and guardians earned less than the tax free threshold in 2017-18. Also, some parents may lodge their tax return too late for it to be included in the linked data.

Analysis of government payments data

9. Data relating to a range of government payments is sourced from the Department of Social Services DOMINO Centrelink Administrative Dataset and linked to MADIP[2]. This data is of interest because of its potential to complement PIT and payment summary data as sources of income values for parents across a range of income and labour force participation categories.

[2] 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'.

10. ABS has undertaken 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; and
    - analysis of the coverage of government payments data in the CTC population.

Conceptual review of government payments included in ATI

11. Both taxable and (some) non-taxable government payments are included in the Australian Taxation Office's (ATO's) definition of ATI[3]. However, some government payments included in the definition of ATI are not available in the DOMINO Centrelink Administrative dataset or in MADIP. Examples of these include payments paid by the Department of Veteran's Affairs (such as Defence Force Income Support Allowance). There are also a number of government payments in DOMINO, such as Family Tax Benefit and Child Care Benefit, that are not included in the definition of ATI.
12. In the analysis described below, 'government payments data', unless otherwise stated, refers to government payments included in the DOMINO data available via MADIP. The reference period used was 2017-18, which aligns with other income data used in 2020 DMI scores. In-scope government payments refers to those government payments included in the definition of ATI.
13. A list of the government payments included in, and excluded from, this analysis, is provided in Appendix 1.
14. Due to the timeframes in which this analysis was undertaken and the ongoing nature of analysis, ABS considers the results presented in this paper to be preliminary.

Coverage of government payments data in the 2020 Address Collection population

15. Preliminary analysis of the 2017-18 financial year data indicates that additional income data, in the form of in-scope government payments, is available for approximately 10% of parents in the 2020 Address Collection.

[3] ATI is defined by the Australian Taxation Office as the sum of the following amounts:
- taxable income
- adjusted fringe benefits (total reportable fringe benefits amounts multiplied by 0.51)
- reportable employer and deductible personal superannuation contributions
- certain tax-free government pensions or benefits received by the person
- target foreign income (income and certain other amounts from sources outside Australia not included in your taxable income or received as a fringe benefit)
- net financial investment loss (the amount by which the person's deductions attributable to financial investments exceeded their total financial investment income)
- net rental property loss (the amount by which the person's deductions attributable to rental property exceeded their rental property income)
- less any child support payments the person provided to another person.

For more information, see: www.ato.gov.au.

- For approximately two-thirds of these parents, the government payments data complements income information available in Personal Income Tax (PIT) data.
- For approximately one-third of these parents, representing 3.4% of all parents in the 2020 Address Collection, an income amount was not available in the PIT data, but can be sourced from government payments data.

Next steps

16. Further analysis of income values for members of the CTC population for whom income information is available from multiple data sources is in progress. This includes analysis of the differences between government payments data and income amounts available via other data sources, such as PIT and payment summary data, for parents who have both.
17. This analysis will inform ABS' recommendations as to how income amounts in government payments data should be incorporated into the income imputation strategy for the DMI methodology. In particular, it will inform recommendations regarding the inclusion of government payments data in a parent's ATI estimate if the parent also has income amounts available in one or more of the other data sources.
18. This analysis and subsequent recommendations for the income imputation strategy are also necessary for calculating how DMI scores would change if the government payments data were included.

Estimating ATI using statistical modelling

19. This section describes research into a possible approach for using a statistical model to estimate, or impute, missing income values for the DMI methodology for calculating CTC scores. This section provides:
    - an overview of the modelling approach undertaken;
    - preliminary findings arising from this analysis;
    - indicators of the performance of the model; and
    - preliminary recommendations arising from this analysis.

Overview of model and approach

20. The statistical modelling described in this paper used data from the 2017-18 ABS SIH Basic Confidentialised Unit Record File (CURF) to create a linear regression model to predict parental income. Linear regression models use available information, or predictor variables, to estimate an outcome variable.
21. The SIH data was subdivided to include only people aged 16 or more, who were categorised as "husband, wife or partner" or "lone parent", and who lived in a household where at least one child attended a non-government school. This sub-sample is considered to be representative of the population of parents in the Address Collection for CTC.
22. The SIH data was not directly linked to the CTC data at the unit record level for this analysis.

23. ATI, the income concept on which DMI scores are based, is defined by the ATO and is not directly collected in the SIH. For modelling purposes, ATI was approximated in SIH by subtracting certain income amounts, which are not included in the definition of ATI, from the value of total income from all sources in the SIH data.

24. People with very low incomes were excluded from the analysis, because it has been found that very low reported incomes often do not accurately reflect the true financial status of the respondent. For this analysis, the lowest 5% of people by approximate ATI were removed from the dataset before the model was constructed.

25. Income values were transformed by taking the natural logarithm. This greatly improves the model fit and diagnostics. However, it means that the model cannot predict negative values of ATI. Other transformations could be explored to determine whether they provide more flexibility while maintaining the beneficial properties of the logarithmic transformation.

The initial model

26. The modelling identified several significant predictors of income – such as sex, occupation, whether a person has a tertiary education and whether a person received government benefits. When these predictor variables are available for a parent via MADIP, they can be used to produce an estimate of the parent's income that takes that extra information into account. The variables assessed for and used in the model are provided in Appendix 2.

27. Including occupation information improved the performance of the model. However, it should be noted that in the SIH, occupation data is collected at the same time as income data. In the linked CTC data, the occupation data for people who did not link to PIT comes from Census 2016 and may no longer be current. This may be a particular issue for parents who were employed at the time of the Census, but left the labour force (to care for children or for other reasons) before the reference period for the CTC score calculation.

Overall model performance

28. Overall, the initial statistical model developed explains less than half of the variation in parental income (adjusted R^2 value is 0.44), which is considered relatively low. Further refinement to the model may improve its explanatory power (Preliminary Recommendation 3).

29. It should be noted that there will be practical difficulties in developing models that can predict income values with a high degree of precision at the individual level for records where income information is not available. This is because income is highly variable and cannot be determined directly from variables such as age group, education level, occupation and geography. For example, two people may have very similar characteristics – the same age, the same gender, the same education and the same occupation – but have very different incomes because they work at different levels in an organisation. Without information directly related to the income received, it will be impossible for a model to predict which of the two people will have the higher income.

30. The performance of the model was tested by comparing the income estimates predicted using the model with the actual ATI of parents whose ATI value is known. The average difference between the modelled ATI and the actual ATI across all parents with a known ATI value was small but positive, implying that the approximate ATI calculated using SIH data may be overestimating ATI slightly on average. Consequently, ABS recommends that further refinements to the model be made, including to the derivation of the approximate ATI (Preliminary Recommendation 3).

31. The difference between the predicted and actual ATI was found to be large for some parents. Specifically, for:

    - 7% of parents, the predicted income was within 5% of the actual income;
    - 14% of parents, the predicted income was within 10% of the actual income; and
    - nearly 75% of parents, the difference was greater than 20%.

32. Linear regression models can provide useful insight into relationships between other factors and income at an aggregate or overall level. Despite the challenges associated with predicting each person's income, a model that may not necessarily provide strong predictive power for an individual can nevertheless make good predictions at the aggregate level. For CTC, this means that, where income data is available for a parent from another source, such as government payments data, that value should be used where it is considered fit-for-purpose (Preliminary Recommendation 2). It also means that assessing the impact of the modelled estimates on school scores is important.

Model performance: Impact on school scores

33. ABS assessed the impact of including the modelled estimates on school scores separately for two groups. The first group consisted of those parents for whom a relatively large amount of information was available to incorporate into a modelled income, as they linked to MADIP and Census data. After including the modelled income values for these parents, 2020 school scores were recalculated. While the majority (72%) of school scores did not change as a result of including modelled incomes, about 28% of schools had some change to their score. Almost a quarter of schools (24%) had a change in score of 1 point, 3% had a change of two points, and 1% of schools had a change of 3 or more points as a result of including modelled incomes in the score calculation. School scores which changed were more likely to increase (65%) than decrease (35%) as a result of incorporating modelled incomes into the score calculation. This is consistent with the model slightly overestimating the approximated ATI on average. (The methodology for calculating DMI scores is provided in Appendix 3.)

34. The second group consisted of parents who did not link to MADIP. Relatively little information – only that which is collected in the Address Collection – is available to incorporate into a modelled income for these parents. These parents were assumed to have lower incomes, and this was reflected in the application of the model.

35. After including the modelled income values for the parents who did not link to MADIP, 2020 school scores were recalculated. In this case, a smaller majority (54%) of school scores

remained unchanged, with about 46% of schools having some change to their score. For some schools (1%), there would be a substantial decline in their scores of 5 points or more. Consequently, further work is recommended to investigate approaches to imputing missing income values for this population group (Preliminary Recommendation 4).

Summary of Preliminary Recommendations

Preliminary recommendation 1

36. Government payments data should be incorporated into the income imputation strategy for CTC, to complement the existing data sources and provide a source of income values for parents across a range of income and labour force participation categories.

Preliminary recommendation 2

37. Given the variety of reasons as to why a person's income may vary, a multi-stage imputation strategy should be used which incorporates available data sources, such as government payments data, to derive ATI before modelling is applied to impute for missing ATI values.

Preliminary recommendation 3

38. The initial SIH model should be further refined and evaluated, before assessing the value of this approach. Areas for possible refinement of the model include the approximation of ATI using SIH data and refinement of the geographical information in the linked CTC dataset to better match the Greater Capital City/Rest of State split in the SIH data.

Preliminary recommendation 4

39. Due to the limited amount of information available to be incorporated into a statistical model for parents who do not link to the MADIP spine, further work is recommended to investigate approaches to imputing missing income values for this population group.

APPENDIX 1: GOVERNMENT PAYMENTS DATA

Government payments data available in the DOMINO dataset which are included in or and excluded from the definition of ATI are listed in table 1.1.

Table 1.1: Government payments – availability in DOMINO dataset and inclusion in ATI.

Payments data available in DOMINO and included in ATI	Payments data available in DOMINO and not included in ATI	Payment included in ATI but not in DOMINO dataset
Job Seeker Payment	Pensioner Education Supplement	Farm household allowance
Newstart allowance	Assistance for isolated children	MRCA Education payments
Youth allowance	Business Services Wage Assessment Tool payment	Veterans' Children Education Scheme
Austudy payment	Carer allowance	Community Development Employment Project (CDEP) payments
Parenting payment (Partnered)	Child care benefit (Formal)	Disaster Income Support allowance
Partner allowance	Child care benefit (informal)	Education entry payment
Sickness allowance	Low income supplement	Widow B pension
Special benefit	Double orphan pension	Age service pension
Widow allowance	Family Tax benefit	Veteran payment
ABSTUDY	Income management	Defence force income support allowance
Youth disability supplement as part of Youth allowance or ABSTUDY living allowance (see note)	Transition to independent living allowance	Defence force income support allowance paid by DVA
Disaster (Emergency) recovery allowance	DFaCS Pensioner Education Supplement	Income support supplement
Age pension	Mobility allowance	Invalidity service pension
Bereavement allowance	Senior health card	Partner service pension
Carer payment	Stillborn baby payment	
Disability support pension	Medical equipment payment	
Parenting payment (Single)	Youth training allowance	
Wife pension		
Parental leave payments		
Dad and partner payments		
Parenting payments (PGA) - Obsolete		

Education payments are included in ATI where the recipient was over 16 years old. For the purposes of this analysis, parents of school-aged children were assumed to be aged over 16 years and included in the analysis.

APPENDIX 2: VARIABLES ASSESSED FOR AND USED IN THE LINEAR REGRESSION MODEL

Table 2.1 provides the full list of explanatory variables included in the variable selection process.

Table 2.1: Explanatory variables considered for inclusion in the linear regression model.

Variable name	Description
male	Whether the person is male
gov_ben_flag	Whether the person received a non-zero amount of government benefits
DSSPENSN	Whether the person holds a low-income concession card
log_ben	The natural logarithm of the total amount of government benefits received (continuous variable), set to 0 if the total amount of government benefits received is less than 1
Geographic variables	State/Territory and Greater Capital City/Rest of State variables combined to create two-way interactions between State/Territory and Greater Capital City/Rest of State where applicable
Age group variables: agegroup2, agegroup3, agegroup4	agegroup1: 16-34 agegroup2: 35-44 agegroup3: 45-54 agegroup4: 55+ (agegroup1 was not included in the model – it is the base category which the others are compared against)
high_edu_flag	Whether the person has a tertiary-level education
no_yr12_flag	Whether the person's highest level of educational attainment was below Year 12
Occupation variables: occ_band_1, occ_band2, occ_band_3, occ_band_4, occ_band_5	Whether the person's 2-digit occupation falls within that group (occ_band_1 contains the occupations with the highest mean incomes according to SIH)
pp_flag	Whether the person received the Parenting Payment
own_home	Whether the dwelling is owned outright or with a mortgage by the household
lone_parent	Whether the person is a lone parent
ftb_flag	Whether someone in the household receives the family tax benefit
nsa_flag	Whether the person receives the Newstart Allowance

A forward selection process was used to identify a set of significant explanatory variables. At the end of this variable selection process, only explanatory variables that were statistically significant at the 5% level were kept in the model. These are listed in Table 2.2 along with their parameter estimates.

Variables listed in Table 2.1 which do not appear in Table 2.2 were not found to be statistically significant in the model-building process. For example, the Greater Capital City area of NSW was the only geographic variable identified as significant by the forward selection procedure, and the 55+ age group was the only significant age category.

Table 2.2: Explanatory variables and parameter estimates for linear regression model.

Variable	Parameter Estimate	Standard Error	Pr > |t|	Standardized Estimate
Intercept	5.29165	0.07553	<.0001	0
male	0.5643	0.04001	<.0001	0.25472
gov_ben_flag	-0.63745	0.17748	0.0003	-0.24508
DSSPENSN	0.24793	0.10539	0.0187	0.05229
log_ben	0.11036	0.03668	0.0027	0.22268
agegroup4	0.17994	0.06764	0.0079	0.04582
state1_cap	0.12897	0.05879	0.0284	0.03691
high_edu_flag	0.15865	0.04178	0.0002	0.07116
occ_band_1	1.958	0.07852	<.0001	0.81674
occ_band_2	1.73572	0.09102	<.0001	0.483
occ_band_3	1.64384	0.07902	<.0001	0.6181
occ_band_4	1.4185	0.08232	<.0001	0.47858
occ_band_5	1.15757	0.08925	<.0001	0.30307
lone_parent	0.41764	0.07161	<.0001	0.11273
ftb_flag	-0.43206	0.05656	<.0001	-0.17711

The adjusted R^2 value for this model is 0.44, which indicates the proportion of variance explained by the model.

APPENDIX 3: METHODOLOGY FOR CALCULATING DMI-BASED CTC SCORES

The Direct Measure of Income (DMI) score

The DMI score is based on the median Adjusted Taxable Income (ATI) of each school community. It is created by:

- calculating the total income for each student by summing the incomes of up to two parents or guardians;
- identifying the median family income for each school; and
- converting the median incomes for all schools into DMI scores via standardisation[4].

The resulting DMI score represents the anticipated capacity to contribute of a school community, relative to other school communities.

The DMI score uses data from the Student Residential Address and Other Information Collection (the Address Collection) to identify the school community population. Income data is obtained via the Multi-Agency Data Integration Project (MADIP) and includes Personal Income Tax (PIT) data, payment summary data and low income concession card information from the DOMINO Centrelink Administrative dataset (formerly provided in the Social Security and Related Information) data. These data sources enable the DMI to use the most accurate and timely income data available for school communities. The PIT and payment summary income data are from the financial year ended 18 months earlier (table 3.1). The DOMINO data aligns with this reference period.

For a detailed description of the DMI methodology, see www.education.gov.au/quality-schools-fact-sheets.

The CTC score

In 2020, a DMI-based CTC score is the average of DMI scores for 2018 and 2019. This is because the first Address Collection to which administrative data in MADIP were linked took place in 2018. From 2021, a DMI-based CTC score will be the average of the previous three years' DMI scores (table 3.1).

Table 3.1 Reference periods of income data used in DMI-based CTC scores.

		Address Collection and DMI score reference year			
		2018	**2019**	**2020**	**2021**
CTC Score	**2020**	2015-16 income	2016-17 income		
	2021	2015-16 income	2016-17 income	2017-18 income	
	2022		2016-17 income	2017-18 income	2018-19 income

[4] Standardisation is a common statistical process which converts a set of numbers, which may have any average and spread, into a pre-determined average and spread. It does not change the order of school communities in the distribution.