Growing Up in Australia: **The Longitudinal Study of Australian Children**

Cleaning of income data

Following the original release of the data, users reported problems with outlying values in the continuous income variables (i.e. afn09a, afn09b, afn09m, afn09f, cfn09a, cfn09b, cfn09m, cfn09f). While this is not unusual for income, it appeared that some of these cases had unusual responses to other questions for those with such high incomes (e.g. more modest incomes reported when asked about combined yearly income at K20 of the face-to-face interview, more menial occupations). It appears that many of these are due to discrepancies between amount and time period when reporting income (e.g. giving yearly income as weekly). Many of these outliers have been subsequently cleaned up, although certain assumptions have been made to do so.

The process for cleaning the Wave 1 data used adaptations of the data query rules coded into the Wave 2 CAPI instrument. As well as providing a logical framework to underpin the investigation, this will also help in making the data more consistent longitudinally

The rules used were as follows:

1) If a respondent's only source of income is Government benefits or salary they should not report an income of $0 or a loss.
2) If profit or loss is a source of income then incomes >$200,000/year should be queried unless they also have a salary
3) Where Government benefits are the main source of income, incomes >$750 a week should be queried.
4) For all other combinations of income types, incomes >$260,000/year should be queried.

Cases identified by the first of these rules were all set to missing. Most of these seem to be due to respondents not counting government benefits as income. In the B-cohort file there were 49 cases of this for Parent 1 and 6 for Parent 2, while for the K-cohort 31 cases were identified for Parent 1 and 6 cases were identified for Parent 2.

For those identified by the other 3 rules, if the categorical annual income for Parent 1 and Parent 2 at K20 was consistent with the continuous values they were left as is. If there was an obvious correction that could be applied (e.g. deleting a zero from an income figure, changing the time period from weeks to year) to bring the income into or close to the range specified at K20 then this was applied. If there was no way that the continuous income values could be made reasonably consistent with the combined parental yearly income then the response was assumed to be an error and was made missing.

Restrictions on publishing case level information limit what can be disclosed about these cases. However, for Parent 1s in the B-cohort, of the 31 cases identified by rules 2 to 4, 6 were made missing, 8 were corrected and 17 were left as is. For B-cohort Parent 2s, of the 48 cases identified, 4 were made missing, 10 were corrected and 34 were left as is. For K-cohort Parent 1s, of the 30 cases identified, 5 were made missing, 6 were corrected and 19

cases were left as is. For K-cohort Parent 2s, of the 42 cases identified, 5 were made missing, 8 were corrected and 29 were left as is.

In Waves 2 and 3, suspicious cases were identified using the above rules. These cases were checked against their income data from earlier waves, plus other information such as work hours and occupation. As would be expected, data collected with the CAPI instrument was cleaner, and fewer imputations had to be made. In Wave 2, 7 corrections made to Parent 1 income, 4 corrections to Parent 2 income, and 1 to the income of other adults in the home. In Wave 3, only one Parent 1 and one Parent 2 required correction.