MELC Developer Questions and Notes

# The Basefile

## Are the basefile variables at birth or at year 1?

They are at year 1 except those explicitly defined at birth or prior to birth (e.g. SESBTH, MAGE, fage, z1single0, pregalc, pregsmk, bwkg, ga).

There are also other unique variables that are not at birth or year 1 in the basefile:

* BREAST which spans the first year (number of months breast-fed)
* at the moment INTERACT and PUNISH are in the basefile and these variables represent mother’s characteristics assessed on the basis of interviewer observations averaged over the period from age 3-5 yrs. These are in the basefile because they are not modelled but just used. In the future these may be modelled which might mean they are no longer needed in the basefile.
* NPRESCH: the total number of years preschool received by the child. This is used in the same way as INTERACT and PUNISH.

The do\_iteration\_year1 function does the iteration for year 1; however all this function actually does is store the current values in the outcomes. This is because it assumes that all variables in the basefile are for year 1.

## What are the values of age in the basefile?

In the basefile, age is 1 for everyone

## Can any variables be blank in the basefile?

Yes, typeofchange is currently blank in the basefile and everything still runs fine. It is possible that things would still run OK if certain other variables were also blank (such as sptype, mumgroup, dadgroup), but this has not been tested.

## What variables must I have in the basefile?

See file H:\workspace\MELC\CHDS\docs\ basefile\_NecessaryVariables.xlsx

Burt and conduct are not in the basefile. These variables are simulated at their starting ages (8 and 6 respectively) using cross-sectional models.

# Simulation Outputs

## Means by overcrowding

* The mean for the ‘All Years’ row for overcrowding can appear much lower than the mean for any of the individual years
* This is because there are many children who in an overcrowded home for only 1 of their 13 years
* So they contribute to the mean for that year
* As overcrowding is a time-variant variable so a decision had to made about how to classify each individual for the all years row
* we decided to take the mean over all years for an individual and then round it
* This meant that the grouping classification for overcrowding is mostly in an overcrowded home over the 13 years or mostly not in an overcrowded home over the 13 years
* This subgroup is different to the subgroup used at each year, where people who were just in an overcrowded home for one of their 13 years were included
* Hence in the all years overcrowding group we have the people in the very worst situation for this variable and the difference between overcrowded and not overcrowded is greater than that for any one particular year

## All years row for quantiles

* Are the quantiles of the averages across the 13 years for each individual
* Hence the maximum mean is much smaller for this row is much smaller than the mean maximum of the entire distribution within one year

# Scenario Guidelines

* Can’t do subgroup scenarios on age, i.e. using age as a subgroup
* Scenarios on at birth or pre-birth variables should not be subgrouped by a time-variant variable
  + I think currently users can actually do this but the ideal would be to not allow them to

# Known Bugs/Problems and Future Improvement Plans

## Current Known Bugs:

* Subgrouping by pregalc in tableBuilder
* In the scenario weightings screen, define a subgroup scenario bwkg<2.5, then click preview on bwkg. This calls tableBuilder. We would expect a table showing 100% of people in the <2500 category but this does not work.

## (Possibly) Plan To:

* Have a button in tableBuilder for turning on or off confidence intervals
* Insert automatically generated “by\_subgroup” tables in the tree
  + These are tables by a subgroup that is defined by a formula by the user in the scenario weightings screen
  + These are generated in R but have not yet been added to the tree the user sees
* Make it so if the user does a scenario on z1single0, that z1single1 also changes, along with householdsize and any other relevant variables
  + At the moment, if a user performs a scenario on z1single0 (and leaves z1single as is) then this has very little effect in the simulation as changing z1single0 does not change z1single at year 1 and it is z1single that is (possibly) used more throughout the simulation.
* Means and frequencies regressing to the mean of the data which was used in the statistical analyses. E.g. the not owning a home percentage starts at 50% and by year 13 has reduced to 18%.
  + Idea to fix this in next round of statistical models
  + We could set the coefficients for the intercept and age and then the coefficients for other variables would have to be estimated using effect parameterisation in SAS (difference from overall mean rather than from mean of the reference category) for categorical variables. The continuous variables could be entered into the model as centred variables (difference from the mean) but this would just give the same beta coefficient for that variable anyway.

## To Check:

* The totals row for means (by year) of fathers age
  + These are not being included in the tool so are a low priority