Executive Summary

1. Optimum imputation methodology for the Scottish House Condition Survey income data was explored by application of nine appropriate methods to the 2003/4 dataset. These methods comprised both single and multiple imputation methods.

2. The single imputation methods included unconditional mean imputation, simple hot deck imputation (with no donor subgroups), hot deck imputation using a regression model to select donors, hot deck imputation using a logistic regression model to select donors and predicted mean imputation (which involved substitution of auxiliary variable values into a regression model to obtain imputed values).

3. The multiple imputation methods used included multiple simple hot deck imputation (which involved repeating the simple hot deck imputation five times with different random seed values each time) , multiple regression imputation (which involved using random draws of linear regression parameters from their posterior distributions), multiple propensity score imputation (which involved calculation of propensity scores of missingness to impute missing income values), and multiple Markov Chain Monte Carlo (MCMC) imputation (which involved several iterations of imputation and posterior-parameter steps until it converged to determine imputed values).

4. The missing income data (approximately 6% of all observations) differed from complete case data (observations with no missing income) with respect to some attributes. This helped establish the missing at random data mechanism (allowing relationships between missing values and variables other than the missing variable itself).

5. It was found that the missing data had more highest income householders and partners in the age range 55-64, whereas the modal age class for the complete case data was the age group 35-44. It was also found that the missing income data observations had comprised more of single adult and single pensioner families, whereas the complete case data contained more small adult and 'older smaller' families (that is households with two adults and two pensionable-aged adults with no children). It was also found that the missing data had more people with long-term sick/disabled people in the household than the complete case data.

6. The single hot deck regression and logistic regression methods, together with the multiple regression and propensity score imputation methods involved creation of appropriate intermediate logistic regression and multiple linear regression models. These were created in a restrictive manner with the primary purpose of being donor selector models.

7. The logistic regression model modelled income missingness and resulted in independent variables of tenure, whether the household had children, the number of householders in a household and the respondent person number. It was significant at the 5% level and gave an area under the ROC curve that was reasonably close to 1 (indicating a good fit).

8. The multiple linear regression model modelled the income variable itself and resulted in independent variables of occupational/employer pension, investment income, working status (in reference week) of respondent and partner age group. It was significant at the 5% level. 30% of the variance was explained by these models.

9. The mean income and standard deviation was calculated for each imputation method and these were compared to those for the complete case income data (with no missing income values). The difference between mean and standard deviation were used to assess which method would be appropriate for imputation.

10. In addition, the mean income and standard deviation for each method was split into council tax bands for post-imputed and complete case income data. The conformity of pattern of the post-imputed dataset to that of the complete case data was also used a criterion to assess choice of method to implement. Practicality of implementation of the methods was also considered as a factor to decide on an optimum method.

11. In terms of the mean and standard deviations, the single unconditional mean imputation underestimated the variance the most (which was expected of the deterministic nature of the method) whereas most of the multiple imputation methods seemed to overestimate the variance. The latter was also expected since methods like the MCMC imputation incorporate uncertainty due to parameter estimates and the imputation procedure.

12. It was not known whether positive or negative deviations were expected from the complete case mean, so the conformity to the complete case pattern of mean income for each council tax band was used to choose an imputation method.

13. Based on this, all the single imputation methods and the MCMC method conformed to the pattern of the complete case data. Although the MCMC seemed most appropriate to implement based on the extra variability for imputation and that it was better suited to the (arbitrary) missing data pattern, it would be more practical to

implement a single imputation method because only one dataset would need to be maintained (rather than five).

14. The simple hot deck imputation method was not chosen because it does not enable checking of sufficient donor subgroup group members to impute. Hence, the hot deck logistic regression method was recommended. This method is consistent with the theory of previous methods applied on this data (and so has the additional merit of consistency) and also assuages distortion of variables and their interrelations when compared with model based methods (such as regression) because only donor selection is based on modelling methods (hence it is semi-parametric).