US Historical Time Use Data File Production

Report on Activities, February to October 2004

Muriel Egerton, Kimberly Fisher, Jonathan Gershuny,
Andreas Pollmann and Nuno Torres

Essex October 24 2004
Revised July 2005

A Report to the Glaser Progress Foundation

US Heritage Time Use Data File Production
Report on activities, February to October 2004

Contents

Data Disk contains:

- Data files
- Documentation of history of harmonization activities
- Syntax files used in data production

US Heritage Time Use Data File Production
Report on activities, February to October 2004

Muriel Egerton, Kimberly Fisher and Jonathan Gershuny,
Andreas Pollman and Nuno Torres

Essex October 22 2004

Executive Summary.

Developments in source data

We have acquired new or improved source versions of the 1965 and 1985 datasets, and identified errors in the data recorded for spouses in the 1975-76 survey respondents. Further work depends on the release of the 2003 ATUS micro-data.

Background variables

The file of background variables has 36 variable harmonized across data for the 1960s 1970s, 1980s and 1990s datasets. A rather larger group of variables is available for comparison between the 1970s and the 2003 ATUS alone.

Time use variables

We have prepared a number of harmonized time use files in three distinct formats (as "episode", "diary" and "summary" files). The first two of these formats include both primary and secondary data (though the 1990s data has no secondary activity material) both coded to a standardized 2-digit activity classification. An additional "secondary care" variable corresponding to the ATUS "caring" will be added to all three sorts of file once the ATUS becomes available.

Quality profiles.

We have carried out a quality analysis of the harmonized variables: a summary of these activities is provided in Appendix 1 of this report.

Codebook and documentation

Appendix 2 provides a detailed codebook for the harmonized files. Extensive documentation on harmonization procedures, together with the SPSS syntax files used for transforming the source materials into the harmonized variables is provided in the data disk attached to this report.

US Heritage Time Use Data File Production
Report on activities, February to October 2004

1 Introduction

This interim report describes the activities and outcomes of the harmonization undertaken by the Essex team for the 'heritage' USA Time Use datasets. A preliminary assessment of the quality of the data will be given, however a detailed quality report will not be submitted until after the final harmonization with ATUS 2003 and subsequent testing.

1.1 Principles for harmonization

Before detailing the activities completed for the production of this report, it may be helpful to restate the principles which have governed the work for this phase.

- The data were harmonized in order to retain the most detailed and accurate information possible.

- In most cases, variables were not edited, rather flag variables were constructed to identify cases where inconsistencies had been found between variables.

- All steps of the harmonization process were recorded in SPSS syntax files and where more complex issues arose were recorded in written reports. The identifiers on the harmonized files were constructed so that the harmonized data can be matched back to the original data files.

- The most detailed codebooks available for the original datasets were acquired and are included in the data disk appended to this report.

A lowest common denominator dataset covering all four decades has been produced at this stage. We plan in the next phase to produce a more detailed dataset with additional variables specifically matching the 1975-76 data with the 2003 ATUS data.

1.2 Developments and improvements in the source data files since the previous report

Since the first report was submitted more data has become available, either from John Robinson and Timothy Triplett from University of Maryland Time Use Research Project (TURP), or from Andrew Harvey and Aimee St. Croix at St Mary's University. In particular, for 1985, a version of the mail survey with more accurate matching between demographic and diary data replaced the previous files (though as discussed below, this file remains somewhat problematic in terms of the quality of matching of the demographic and time diary data). We still do not have the 1995 dataset, but once we receive this, its close similarity to the 1992-4 materials means that a harmonized file incorporating it can be constructed quite speedily.

As in the previous report, the data for 1965-66 was for a sample of urban workers with a restricted age range (19 to 65). The 1965-66 data made available by the TURP team also included variables for race and for region, and these have been included in the dataset submitted.

The 1975-76 study followed respondents over four waves. The questionnaire varied substantially between each wave, necessitating consistency checks across waves. The 1975-76 study also collected diaries and limited additional information from the cohabiting partner or spouse of those respondents living with a partner. Fewer separate background questions for spouses were collected, and these also varied across waves.

The documentation to the 1975-76 survey states clearly that the spouse diary instrument did not collect secondary activity. We have found nevertheless that the spouse lines interpolated between respondents within the data file deposited in the Michigan archive do apparently include some a considerable amount of secondary time. It appears that this is the misleading result of a simple error in a transposition procedure described in the documentation, leading to the attachment of respondents' secondary activity to their spouse's diary records in the spouse lines, and these data should be ignored. (We have prepared files with 1975 spouse data with the same structure as the respondent data, but excluding the secondary activities). As is usual with panel studies, attrition between the 1st and 2nd panel was substantial, but decreased for the 3rd and 4th waves. Our proposal for this study is to weight cases as the inverse of their response probabilities on the basis of wave 1 characteristics (these weights are however not included in this preliminary release of the harmonized data).

The 1985 survey was, in effect, a household survey, with all household members asked to respond. This survey also had a panel element. However, in what follows only the first survey (a mail survey) was used. Only responses from household members aged 18 and over are included in the harmonized dataset. The 1992-94 data made available to us by the TURP team included two extra variables; number of adults in the household, and the zip code of the home from which an urbanicity variable could be derived. More detail on survey and sampling characteristics are given in the codebook.

1.3 Using the example data files

Users will find three sets of SPSS data files with this report. The first set of data files covers the harmonized background variables from the 1960s through the 1990s. The structure of these files follows the following format, with the yy representing the decade in which the data was collected: USAyySQUEST (eg. USA60SQUEST.SAV, USA90SQUEST.SAV).

The second set of files contain the time use diary variables, and have the following structure, with the yyyy reflecting the first or central year of data collection, HF standing

for heritage files, and x reflecting the nature of the time variables in the file: USAyyyyHFx.SAV

We have created three types of harmonized time use files:
EP – episode level files - each row case represents one episode of activity in a diary.
D – diary files - each row case represents one 24-hour time diary organized in 5 minute time slots.
SUM – aggregated (summary) time file - total minutes spent during the diary day in each of the 1-digit time use codes (main activity only), and total minutes spent during the diary day in each 2-digit main activity code where no secondary care time was recorded followed by a separate set of variables showing the total minutes spent during the diary day in each 2-digit main activity code where secondary care time (either child care or adult care) was also recorded. The sum of time spent in all 1-digit time use activity codes plus total missing time equals 1440 (ie a complete day). The sum of time spent in all 2-digit codes with no secondary care added to the sum of time spent in all 2-digit codes occurring alongside secondary care time plus total missing time also equals 1440 minutes. Each row case represents one 24 hour time diary.

To use the files, the user should open the questionnaire file from the particular year and match-in the preferred format of time diary data. The key matching variables are SURVEY (survey identifier), WAVE (wave identifier), HHID (household identifier) and PID (person identifier). The user should initially ensure that both the questionnaire file and desired diary format are sorted by these identifiers in this order (survey, wave, hhid, pid) before matching the files together.

When matching individual surveys, the PID (person identifier) suffices for matching the data from 1992/94, as this survey collected only one diary from one person per sampled household. The PID and WAVE variables are necessary for matching files for 1975/76, as this study collected diaries from the same people over four waves The 1985 study collected diaries from all older children and adult household members. Thus for this study, the household identifier and person identifier are needed to match in the data. As the 1975/76 study collected diaries from the same people over four waves, the WAVE identifier is needed to match the data from this year. The 1965/66 surveys include two separate samples followed over the same period, a sample of Jackson and a near-by rural area, and a national sample. We include data from both studies, as the two samples are both small and as they have been analyzed together in the past. Nonetheless, the SURVEY variable marks the difference between the samples for researchers wishing to consider only one of these studies or two consider the samples separately. Sorting all files by these four identifiers before matching files for any year will lead to a successful match of the datasets.

The third set of files contain additional information offered by the 1975/76 longitudinal time use study. The file called COMBWAVE75 displays the participation of each respondent across the four waves of this study in one file. This survey also collected limited background information as well as diaries from spouses or cohabiting partners

from those respondents who had a live-in spouse or partner. The three files called USA1975HFSPOUSEx.SAV contain the limited background variables and the diary data in the three formats for the spouses. These extra files contain the same identifier variables found in the main datasets.

Users should also note two problems when matching the files. First, in the case of the 1985 data, the demographic details were poorly matched to the diary data. We are still working to reduce this problem by detailed inspection and editing of individual cases; at the present moment users will encounter a number of cases where diary and demographic data do not match at present. In the case of the 1992-94 data, there are 91 cases of child diarists whose demographic details are recorded, but for whom age is missing. In 14 of these cases, the child data will be removed in the next version. In the other 77 cases, it appears that 18 and 19 year old diarists completed child diaries rather than adult diaries by mistake, and diary data for these cases will be recovered in the next release.

Weighting variables, to compensate for sampling and non-response differences between the various datasets are currently under development, and will be included in the next release of the data.

2 Background (non-time-use) variables

A brief overview of the harmonization and checking of background variables is set out in this section. (further documentation and SPSS syntax files detailing the derivation of variables will be found in the directory /History of Harmonization/questionnaires/ in the data disk supplied with this report).

2.1 Development of background variables

The datasets varied substantially in the numbers of variables held, and their structure. Relevant variables in the 1965-66, 1985 and 1992-94 datasets were coded in broadly similar ways however. Therefore the team concentrated on wild code and consistency checks for these datasets. The data has previously been scanned for wild codes and the main activity here was recoding missing values consistently over the harmonized datasets. More detail on consistency checks is given below.

Variables for the four waves of the 1975 dataset were derived from the original data and could be derived to be consistent with the other surveys. However, the greater detail of the 1975 data meant that more detailed checking was possible to resolve apparent inconsistencies, such as those arise with variables which may change from survey to survey, (e.g. marital or employment status).

Table 1: Variables subject to change between waves of the 1975-76 data[1]

Variable		Wave 1	Wave 2	Wave 3	Wave 4
Individual					
	Qualifications	Only held on Wave 1			
	Marital status	Y	Y	Y	Y
	Family status	Y	Y	Y	Y
Household					
	Household type	Y	Y	Y	Y
	Number of adults	Y	Y	Y	Y
	Number of children	Y	Y	Y	Y
	Age of children	Y	Y	Y	Y
Employment					
	Economic activity	Y	Wave 2 is dependent interviewing	Y	Y
	Work hours	Y	Y	Y	Y
Income					
	Household income	Only held on Wave 1			

[1] Less detail is held for many variables on Wave 2 (see /History of Harmonization/).

2.2.1 Geographical information

Urbanicity

Information on the urbanicity (i.e. economic/social integration of a location into a metropolitan area, irrespective of whether the immediate environs are urban, suburban or rural) of the respondent's home was held on 1965, 1975 and 1985**.** The zip code of the respondent's home was held on the 1992-94 dataset. It was possible to check these zip codes against the 1990 census tables of zip codes by urbanicity and derive an urbanicity variable (see directory /History of Harmonization/questionnaires/). About 14% of values were missing, either because the zip code was a missing value in the dataset or because it was not in the census tables, due perhaps to respondent or data error. Over one third of the missing values on urbanicity occurred in the final tranche of the survey (approx one eight of the sample). The missing value rate was much higher among older people. Over 10% of people aged over 65 had missing values on the zip code of home variable compared with about 5% of respondents aged between 18 and 65.

Region and state

Census region information is held on 1965, 1975 and 1992-94. The region variable in 1965 was unlabelled and the regions were identified by adding together the frequencies of the sampling locations given in the codebook by region and matching these frequencies to the frequencies of the variable on the dataset. However, there were missing cases in this dataset and the identification of region may have some error.

The state location of the respondent's home is held in 1975-76 and 1992-94. Minor recoding of the 1992-94 data was necessary to bring it into line with the FIPS code frame. Consistency checks between state and census region were carried out and showed the variables to be consistent. EPA (Environment Protection Agency) regions were held on the 1992-94 data. Since this is a more detailed variable than Census region, an EPA region variable was derived from state for 1975.

2.2.2 Individual characteristics

Gender

Checking showed that the coding of sex changed for 10 respondents between waves of the 1975 data. No conclusive determination of the correct value could be found. Therefore, since it is known that analysts are particularly sensitive to this particular error and since these cases could not be used for weighting, the cases were excluded from the dataset. Over the four waves, this means the loss of 38 cases (less than 1% of the total). No problems were identified on any other survey.

Age

Child diaries were held on the 1985 and 1992-94 surveys and these were removed from the dataset. In both these surveys there were non-trivial percentages of missing values for age (between 5% and 10%). Other information suggests that many of these cases are adults and an attempt will be made to impute age in the next stage of the project, in order to retain these diaries in the datasets. In the first wave of 1975 six cases were found with missing values on age and a further 5 were found to be inconsistent on age-based filters. Age was checked across waves of the 1975 dataset. Ages which were consistent within a five year range were allowed, i.e. age could not be more than 2 years younger or 2 years older than the age given in another wave. However, after selecting out the cases with inconsistent values on sex, only 7 values of age were found to be missing or inconsistent over the four waves (.02% of the data).

Ethnicity

Only minor recoding was necessary. Two variables for ethnicity were produced. A three category variable which was consistent over 1965,1975 and 1992-94 and a more detailed variable which was consistent between 1975 and 1992-94.

Education

More detailed codings of education were held on the 1975 and 1992-94 surveys. These were condensed to harmonize with 1965-66 and 1985. Some checks were carried out.

2.2.3 Family and household characteristics

Marital status

The question was not asked in 1992-94. Therefore it was not possible to construct household type for this survey. A partial household type variable was constructed identifying single person households by gender. The question about marital status was only asked in Wave 1 of the 1975 survey. However, changes in marital status were traced over the panels of the 1975 survey. Except for Wave 2, (2 to 3 months after wave 1) where only filters for marital status were found, it was possible to identify becoming married with some certainty through the relationship of household members to the respondent. It was not possible to identify with certainty respondents who had separated. A flag variable was created to identify respondents whose status had or seemed to have changed.

Number of children

Three variables were held, although not all of them were on all datasets (see Directory /History of Harmonization/questionnaires/ for more detail). Children aged under 18 was on all datasets, however children aged under 5 was not held on the 1992-94 datasets. The

age of the youngest child was held on this dataset and was also held on 1975. Consistency checks between these variables were carried out. Flag variables were created where inconsistencies had been identified.

Number of adults

This variable was only held on 1975 and 1992-94. Consistency checks were carried out across waves of the 1975 data, particularly where apparent changes of marital status were found.

Household type

A derived variable for household type was supplied with all surveys except the 1975-76 survey. Consistency checks were carried out with the marital status, number of adults and numbers of children variables. The variable was derived from these variables for the four waves of 1975. It was found that the two categories 'married without children' and 'married with children' were households which also contained other adults.

Family life cycle

This variable was derived from the age of the respondent and the age of children. It refers only to the respondent's status, not to the household.

2.2.4 Economic characteristics

Economic activity

The questions and filters varied across surveys, necessitating complex sets of checks (for more detail see Directory /History of Harmonization/questionnaires/). Different sets of variables were supplied with the datasets used. Eventually a set of dummy variables identifying the respondent's main economic activity were constructed, a variable identifying whether or not the respondent did any paid work or not, and whether or not this was fulltime or part-time was constructed. Work hours variables were harmonized. The dummy variables include people who might also be doing paid work (e.g. retired, students, housewives). Flag variables were constructed to identify respondents with non-paid work main economic activities who had fulltime work hours. Codes were constructed to identify respondents in 1965-66 and 1992-94 who had been routed out of questions about either main economic activity or work hours.

Household Income

Household income was not held on 1992-94. On 1965-66 it was a 10 category banded variable, while on 1975 it was an 18 category variable. On 1985 only quartiles were held. Therefore it was necessary to derive a four category variable. The bands did not break into exact quartiles and there is variation of about 3% around the quartiles. The 1965

variable had an extra code '10' accounting for 2.2% of the data. The variable was unlabelled and the labels were found from the codebook. It was unclear how to treat the extra category. However, cross-tabulating it with education level (see below) suggested that it did not indicate a very high income. It seems likely that it arose sometime during data processing, and it was treated as missing (see Directory History of Harmonization/questionnaire for more detail). Table 2 shows that earnings rise with qualification level as would be expected. However respondents with code 10 have a high percentage of missing values on qualifications and no trend by qualifications.

Table 2 Crosstabulation of household income in 1965-66 by education qualifications

	Education							
Household Income	Missing	0 – 8th Grade	9 – 11th Grade	High School Graduate	Some College	College graduate	Post College	All
0 under $1000	65.6	2.7	0.5	0.1	0.7			1.6
1 $1000-1999		2.7	2.3	0.5	0.3	1.0		1.1
2 $2000-2999		3.9	4.0	0.9	1.0	0.5		1.8
3 $3000-3999	3.1	7.8	7.5	4.0	2.8	1.5		4.7
4 $4000-4999	6.3	12.5	8.3	6.3	5.2	1.0	2.2	6.7
5 $5000-5999		12.5	12.0	9.6	9.7	5.9	2.2	9.7
6 $6000-7499		21.6	20.0	19.9	12.8	7.3	6.5	17.2
7 $7500-9999	3.1	17.6	26.0	28.6	22.6	18.0	21.7	24.2
8 $10000-14999	3.1	12.9	14.3	22.0	28.1	31.2	28.3	21.0
9 $15000 or over	3.1	2.0	3.5	6.4	14.9	32.2	37.0	9.7
10	15.6	3.5	1.8	1.8	1.7	1.5	2.2	2.2
	100	100	100	100	100	100	100	100

The household income variable for the 1975 dataset was the yearly household income in 1974. It did not seem to be possible to construct an accurate household income variable from the data on household members earnings and benefits, however it does seem to be possible to construct an accurate variable for the respondent's earnings (see Directory /History of Harmonization/ questionnaires/).

2.3 Summary of data quality issues for background variables

An extended discussion of data quality issues is found in Appendix 1 to this report. We report here currently unresolved issues.

We have decided for the moment not to use the 2nd and 3rd waves of the 1985 data, since the original datasets for these surveys seem to have been corrupted. When clean source material becomes available we will be able to produce harmonized data simply by using the current syntax for this year.

It might have been expected that differences in questions and filtering on employment questions in the 1992-94 would have some effect. However, if so, the differences are small and would need more detailed investigation. The second wave of the 1975 survey used dependent interviewing techniques for employment questions. However this wave has most missing values on employment variables, and possibly either interviewers or respondents found the filtering confusing.

An as yet unresolved ambiguity concerns the first two categories of the household type variable 'married without children' and 'married with children' which could in principle include adult children or other adults. We hope to resolve this once we have access to the 2003 micro-data. Another possible problem in harmonizing with the 2003 data is the composition of the Hispanic variable. In 1992-94 it represents Hispanic origin or descent, rather than current self-defined ethnicity and it has higher rates of Hispanic origin than the Hispanic ethnicity category of the ethnicity variable. Although a Hispanic ethnicity variable could be derived from the 1975 ethnicity variable, this was based on interviewer observation and again care will need to be taken in harmonizing this variable with the 2003 data.

The largest error which is likely to be due to data processing is on the urbanicity variable for 1992-94, with nearly 14% of missing values. The variable was derived at Essex from zip codes held on the dataset. About half of the error is due to zip codes defined as missing on the dataset. However, the rest (6.6%) occurred because the zip codes in the dataset were not held in the Census tables. This might be due to respondent error, as suggested by the higher rate among older people, or due to interviewer mistranscription or due to mispunches. Since about half the error occurred in the final tranche of the survey, this might suggest some change of survey procedure which affected the processing of the data, assuming that new zip codes were not introduced in that period.

Generally, the quality of the data seems good, in particular that of the 1975 survey. The 1975 survey is very detailed, therefore it was possible to do detailed consistency and filter checks, documented in directory /History of Harmonization/questionnaires/, and for the most part, identifiable error was of the scale of 2 to 3 cases, excepting sex and age, where about 10 cases seem to have been mismatched over the waves (comprising less than 1% of cases).

2.4 Summary list of background variables

With the exception of the more detailed ethnicity variable, which may not be available for ATUS 2003, Table 3 lists the variables that we have harmonized for comparison with the new dataset.

Table 3 : Questionnaire variables in the harmonized dataset						
Variable		Variable name	1965-66	1975-76	1985	1992-94
Match variables						
Survey identifier		*Survey*	Y	Y	Y	Y
Wave identifier		*Wave*	Y	Y	Y	Y
Household identifier		*Hhid*	Y	Y	Y	Y
Person identifier		*Pid*	Y	Y	Y	Y
Survey						
Census Region		Regionc	Y	Y	Y	Y
EPA Region		Regione	N	Y	N	Y
State		State	N	Y	N	Y
Urbanicity		Urban	Y	Y	Y	Y
Demographic						
Sex		Sex	Y	Y	Y	Y
Age		Age	Y	Y	Y	Y
Age condensed		Agecat	Y	Y	Y	Y
Ethnicity (5 category)		Ethnic	N	Y	N	Y
Ethnicity (3 category)		Ethnic2	Y	Y	N	Y
Hispanic		Hisp	N	N	N	N
Education		Educ	Y	Y	Y	Y
Household						
Marital status		Civstat	Y	Y	Y	N
Flag for change in marital status or inconsistencies		Marrflag	N	Y	N	N
Family life cycle		Famstat	Y	Y	Y	Y
Household type		Hhtype	Y	Y	Y	N
Flag for edited inconsistency between household type and children variables		Kidflag	Y	N	N	N
Number of adults		Nadult	N	Y	N	Y
Number of children aged under 18		Under18	Y	Y	Y	Y
Number of children aged under 5		Under5	Y	Y	Y	N
Age of youngest child		Ageyngst	N	Y	N	Y
Flag for inconsistencies between number and age of children		Nkidflag	N	Y	N	N

Table 3 (continued): Questionnaire variables in the harmonized dataset						
Variable		Variable name	1965-66	1975-76	1985	1992-94
Employment (full, part time or none)		Empstat	Y	Y	Y	Y
Not employed identifier		Noemploy	Y	Y	Y	Y
Dummy variable for full-time employed		Fulltime	Y	Y	Y	Y
Dummy variable for part-time employed		Parttime	Y	Y	Y	Y
Dummy variable for unemployed		Unemp	Y	Y	Y	Y
Dummy variable for retired/disabled		Retdis	Y	Y	Y	Y
Dummy variable for fulltime family care		Homemakr	Y	Y	Y	Y
Dummy variable for student status		Student	Y	Y	Y	Y
Work hours		Wkhrs	Y	Y	Y	Y
Number of workers in household		Nwork	Y	N	Y	N
Flag for inconsistencies between work hours and economic activity		Empflag	N	Y	N	N
Flag for missing work hours		Wkhrflag	N	Y	N	N
Number of workers in household		Nwork	Y	N	Y	N
Income						
Household income in quartiles		Incomeqt	Y	Y	Y	N
1965-66 Household income banded		Income65	Y	N	N	N
1975-76 Household income banded		Income75	N	Y	N	N

The 1975/6 dataset provides considerably more information about respondents than do the other heritage materials. So it has been possible to produce a number of additional variables for comparison with the ATUS from these materials. Table 4 lists some of the additional variables available for the 1975/76 – 2003 comparison files in the /1975 longitudinal/ directory on the data disk.

Table 4 Selection of extra variables for 1975/6 –2003 comparison	
1975	ATUS
Family status	Can construct
Household type	Y. Type of household in CPS. Household type is not recoded in ATUS because the roster information isn't complete enough to do so in all cases.
Number of adults	Y
Number of children	Y
Age of children	Y
Housing Tenure (1975 Wave 1)	It is in the variable list
Housing Assets (1975 Wave 1)	It is in the variable list
Occupation	Y
Industry	Y
Earnings	Y

3. Diary (time-use) variables

A brief overview of the harmonization and checking of diary variables is set out in this section. (further documentation and SPSS syntax files detailing the derivation of variables will be found in the directory /History of Harmonization/diaries/ in the data disk supplied with this report).

3.1 Development of diary activity variables

The initial phase of developing harmonized time use variables involved the creation of a concordance file of the time use information available in each study. The full concordance file is located on the data CD under History of harmonization/diary. This file considers the time use activities and contextual information separately. During the creation of the concordance files, we found that some original datasets no longer were attached to the detailed original activity coding frames – this is particularly a problem for 1985. We have reunited files and coding frames, and also placed the value labels for the 1985 time use activities on the Multinational Time Use Study website so that they are now freely available to any interested researcher.

We have developed two levels of codes. First, we adapt the broader groups of codes in the 2003 ATUS into a list of 10 1-digit level activities:

0 – personal care
1 – paid work
2 – education
3 – unpaid domestic work and child care
4 – adult care, civic, voluntary and religious activities
5 – out of home free time and leisure

6 – sports, exercise and outdoor activities
7 – in-home free time and leisure
8 – media and computer use
9 – travel

We then break these 10 1-digit level activities down to 91 two-digit level codes. In principle, we carry forward as much detail as is possible between all the studies (from 1965-66 through the 2003 ATUS). Once the ATUS data becomes available, we will develop a more detailed activity list for the 1975-76, 1985 and 2003 codes, as these studies offer more extensive ranges of codes of activities.

The one activity distinguished in the original older files which we have not retained is the distinction between routine indoor and routine outdoor cleaning activities. The reason we do not include this distinction in the harmonized data codes is that we have developed a more detailed indoor and outdoor activity variable, which will allow users to continue to identify this distinction when they analyze the data. We do note that in the older data sets, some activities, including pet care, care of adults, some forms of training, and use of libraries were coded in different general groupings in different years of the heritage studies. We have developed a harmonized method for classifying activities across the years, but the concordance file does retain the original coding of activities from each study so that users can see the earlier choices.

The 1965-66, 1975-76, and 1985 data made use of the same activity list to code both main activities and simultaneous secondary activities. The 1992-94 data does not include secondary activity. At this time, the 2003 ATUS includes main activity and secondary child and adult care only, however, the original data does include simultaneous activities when diarists reported multi-tasking. This secondary activity will not be released with the first round of the ATUS data release, but as the original information is retained, it may be possible to expand the available ATUS data in the future to include the secondary activities. Investigation of the possibility of expanding the ATUS data is worth further consideration.

At this time, we propose to use the same harmonized variable codes for main activities and for secondary activities for the earlier data sets. We code the secondary activity variables as missing for the 1992-94 data, and we will only include the child care and adult care codes for the 2003 ATUS. In the aggregated summary files, we propose to include 91 variables for total time recorded in each main activity when no secondary care information is recorded, and an additional set of 91 variables for total time recorded in each main activity which secondary care time is recorded.

Roughly one-third of episodes recorded in the 1965-66, 1975-76, and 1985 studies are multi-tasking episodes (Table 5). In the 1975-76 study, the level of reporting of secondary activities increased for both main respondents and spouses. The level of reporting of secondary activity is highest in the third wave, but generally consistent across the last three waves. This initial analysis of the reporting of secondary activities provides part of the background against which the possibility of recapturing secondary activity data for the 2003 ATUS may be judged in the future.

3.2 Summary lists of diary activity variables

Table 6 lists the time use variables and the range of years for which these variables are available in the test dataset distributed with this report.

Table 5 Recording of secondary activities					
	1965	**1975-76 main sample**	**1985**	**1992-94**	**2003**
% reporting secondary activity	96.7%	94.3%	93.2%	NA	
average simultaneous act time	5 hours 45 min	6 hours 34 min	6 hours 31 min	NA	
mean % of episodes with second acts	30.3%	37.4%	33.7%	NA	

1975-76 sample waves	**main sample**			
	wave 1	**wave 2**	**wave 3**	**wave 4**
% reporting secondary activity	89.7%	95.0%	97.0%	98.0%
average simultaneous act time	5 hours 33 minutes	6 hours 44 minutes	7 hours 34 minutes	6 hours 59 minutes
mean % of episodes with second acts	31.6%	38.4%	42.5%	39.8%

Table 6 Variable names and labels for all time use files							
		Surveys for which this range of variables is available					
name	**label**	**1965**	**1975-76 main sample**	**1975-76 spouses**	**1985**	**1992-94**	**2003**
diaryday	day of week diary kept	in dataset	in dataset	in dataset	in dataset	in dataset	
cday	calendar day		in dataset	in dataset	in dataset	in dataset	
month	month diary kept	in dataset	in dataset	in dataset	in dataset	in dataset	
year	year diary kept	in dataset	in dataset	in dataset	in dataset	in dataset	
usualday	usual or unusual day	in dataset	in dataset	in dataset			
Variable names and labels for all episode time use files							
start	minute started (of 1440 min per 24 hour day)	in dataset	in dataset	in dataset	in dataset	in dataset	
end	minute ended (of 1440 min per 24 hour day)	in dataset	in dataset	in dataset	in dataset	in dataset	
time	duration of activity in minutes	in dataset	in dataset	in dataset	in dataset	in dataset	
epnum	episode number	in dataset	in dataset	in dataset	in dataset	in dataset	

Table 6 (continued) Variable names and labels for diary and episode files							
name	**label**	**1965**	**1975-76 main sample**	**1975-76 spouses**	**1985**	**1992-94**	**2003**
main	main activity – basic codes	in dataset	in dataset	in dataset	in dataset	in dataset	
emain	main activity – extended codes						
s_car	secondary adult and child care						
sec	secondary activity – basic codes	in dataset	in dataset		in dataset		
esec	secondary activity – extended codes						
mtrav	mode of travel						
inout	activity outside, inside, or in vehicle	in dataset	in dataset	in dataset	in dataset	in dataset	
loc	location (diary location column only)						
eloc	location (diary supplemented by activities)						
with	who else was present (diary column only)						
ewith	who present (diary supplemented by activities)						
smoke	near someone smoking						
Variable names and labels for aggregated file							
t0xx to t9xx + tmiss	total minutes per day in each 1-digit code - main activity only plus missing time for complete diaries, t0pcare+ t1paid+ t2ed+ t3unpaid+ t4acvol+ t5outhm+ t6exerc+ t7inhm+ t8media+ t9trav+ tmiss =1440 min	in dataset	in dataset	in dataset	in dataset	in dataset	
tm1 to tm98	total minutes per day in each 2-digit code with no secondary care	in dataset	in dataset	in dataset	in dataset	in dataset	
tsc1 to tsc98	total minutes per day in each 2-digit code alongside secondary care the sum of tm1 to tsc98 + tmiss=1440 minutes for complete diaries	in dataset	in dataset		in dataset		

3.3 Activity context variables for the diary data

We also propose to create, in the next phase of work, a series of diary variables with harmonized contextual information. The final categories cannot be chosen until the 2003 ATUS data is released, but we can outline the following general directions for this work.

For both location and who else is present, there are two dimensions of possible harmonized codes. On one level, we can harmonize the location and who else is present variables based on the categories coded from these columns in the diaries. On a second level, we observe that there are implicit location and who else is present codes in the activity categories. For instance, the recurring category "play with child outdoors" implies that the diarist is outside. The category "look after infants aged less than 5" implies that young children are present. We thus propose to create three location variables:

- location based on diary categories
- expanded location including information implicit in activity codes
- inside, outside, in a vehicle, location unknown

and two who else is present codes:

- who else is present based on diary codes
- expanded who else is present including information implicit in activity codes.

As the ATUS data had not been released in time to complete the concordance file and test time use data sets, we have only created the inside or outside variables at this time.

The diary studies from all years except 1985 and 1992-94 include questions concerning whether respondents completed their diary on a "usual" or an "unusual" day. We will include a marker variable for usual or unusual days.

In addition to the location, who else is present, and usual or unusual day, two other dimensions of information may prove possible to construct: mode of transport and whether the diarist was in the vicinity of smokers. These two additional variables can be created in great detail for the 1992-94 data. Limited mode of transport information is available for 1985. We do not yet know if this information is available in the 2003 ATUS data. No transportation detail is available in the earlier studies. No other study directly includes around smoke information, though the presence of smoke might be imputed from some activity codes from 1975-76, 1985, and the 2003 ATUS. Nonetheless, our initial testing of the 1975-76 and 1985 data suggests that the imputed around smoke data is not consistent. We do not recommend including either of these additional contextual details unless the information is clearly distinguished in the 2003 ATUS.

4 Sampling

Appendix 1 describes the results of a wide range of quality checks on the questionnaire and diary materials in the heritage data sets. Some preliminary conclusions on the quality of the samples are set out briefly here.

It is clear that the 1965 survey varies considerably from the national population statistics, being restricted to respondents aged 19 to 65, in households with at least one member working in an urban area. This clearly affects the frequency distributions for the labor market participation variables, and also seems to affect educational level and family variables, such as marital status and number of children (probably through the age restriction). No tests have been made of how well the survey covers the target population, since it is not clear how useful this survey is for National Accounting purposes. However, the response rate is quite high, at over 70%, and the survey will still be useful for some academic research purposes. The other surveys are designed to be nationally representative (the 1985 and 1992-94 surveys are based on telephone-using households only).

Unit non-response rises over time, but seems to have consistent effects, with the Time Use Survey respondents being somewhat older, better educated and slightly less likely to be married than population statistics as estimated from the CPS. Their work hours vary somewhat from population statistics, but weighting for education and age may remove this difference. The Time Use Surveys seem to under-sample minority ethnic groups.

Attrition is quite high over the waves of the 1975-1976 survey. As is often found, attrition is highest between the first and second waves. Attrition biases the samples further towards the better-educated, employed with higher household incomes. Attrition is higher among African-Americans. We will construct special weights to compensate for attrition biases in the final three waves of this survey.

Item non-response is low generally. Apart from Household Income, the highest rate of item non-response seems to be for work hours in the 1992-94 survey (at 2.6%) and work hours in the second wave of the 1975 survey (3.9%). It is often found that there are quite high levels of non-response on household income, either because the informant does not know, or because of refusal. There are non-trivial numbers of cases missing on the age variable in 1985 (about 7%). Although some of these cases are children, whose demographic information was deleted, inspection of the diary data suggests that some of the cases are adults.

Appendix 1 : Data quality profiles

Since the harmonization is of extant datasets, the most important quality issues are those to do with the adequacy of the achieved sample and with consistency of measurement across surveys. These issues guided the harmonization process, and are discussed in more detail in the documentation and syntax files in the directory /History of Harmonization/questionnaires/.

A1.1 Distributions of key sample variables compared to the Current Population Survey and to Census statistics

As will be noted below, the surveys have different structures and this affects their comparability with national statistics and also their inter-comparability. The 1965-66 survey was designed to sample working households in urban labor markets. Therefore all households are selected to have at least one member in employment. The age-range is restricted to between 19 and 65. The 1975 survey is designed to be nationally representative. However, it is a panel survey and there is considerable attrition between the first and second waves (25%). Attrition thereafter is smaller, a further 8% in Wave 3 and 1% in Wave 4 approximately. Statistics on the effects of attrition between waves in 1975 are presented. The 1985 survey sampled all members of the household, while the other surveys sampled only one person per household. Therefore the age structure of the 1985 sample is younger and closer to Current Population Survey and Census statistics. In many tables statistics for only one person in the household are presented along with household statistics for 1985.

A1.2 Sample – geographic distributions

The Census variable for region was held on 1975 and on 1992-94. There were no missing values in either year. An unlabelled variable for region was available for the 1965-66 survey. The frequencies for sample points and states are given in the codebook and from these it is possible to estimate the frequencies which should be found in the dataset and thus the labels for the regions. However, the relevant dataset (n=2001) does not include the total sample (n=2044) therefore there may be some uncertainty about the labeling. As an urban sample concentrated in the Mid-West, the 1965-66 dataset cannot be compared with census statistics. However, the unweighted distribution of frequencies of the 1975 and 1992-94 data agrees well with weighted data from the Current Population Study and Census tables (http://www.census.gov/statab/www/poppart.html). Distributions for 1992-94 and for Wave 1 of 1975 are shown below.

Table A1.2a % distribution of region of respondents' home by survey (Col %)						
	Survey	1 1965-66 Szalai & national sample	2 1975-76 longitudinal survey – Wave 1	3 1985 Univ Michigan survey	4 1992-94 NHAPS survey	total
Regionc Census Region	-9.00 data not available			100		36.66
	1 Northeast	18.5	20.5		21.9	13.53
	2 Midwest	55.9	27.9		22.2	14.75
	3 South	13.0	32.5		34.3	20.08
	4 West	11.6	19.2		21.7	14.99
Total		100	100	100	100	100

In the 1975-76 study, the variable for region is given in the 1st wave only. Although there is sizeable attrition between waves on the 1975 survey, little change is found by region. No information is given in the codebook on tracking of respondents and it is likely that only respondents who did not move house remained in the survey.

Table A1.2b Percentage distribution of region of respondents' home Waves 2, 3 and 4 by Survey				
Region	Wave 2 1975	Wave 3 1975	Wave 4 1975	Wave 2 – 4
1 Northeast	19.5	19.5	19.1	19.4
2 Midwest	30.3	30.6	31.8	30.9
3 South	30.0	29.5	28.8	29.5
4 West	20.1	20.4	20.3	20.3
	100	100	100	100

A variable for State was held on the 1975-76 and the 1992-94 surveys, coded into the FIPs code frame in 1975. There were no missing values on State on either of these surveys.

Table A1.2c Percentage distribution of State for respondent's home by survey					
	1 1965-66 USA Szalai national sample	1975-76 longitudinal survey Wave 1	3 1985 Univ Michigan survey	1992-94 NHAPS survey	
-9 data not available	100		100		36.66
1 Alabama		1.39		1.49	0.93
4 Arizona		1.39		2.17	1.28
5 Arkansas		3.98		1.13	1.05
6 California		10.27		11.15	6.96
8 Colorado		1.26		1.75	1.05
9 Connecticut		1.59		1.46	0.94

10 Delaware				0.12	0.06
11 District Of Columbia		0.07		0.29	0.16
12 Florida		3.58		5.86	3.45
13 Georgia		2.19		2.61	1.60
16 Idaho				0.15	0.08
17 Illinois		4.17		5.05	3.10
18 Indiana		1.06		1.72	1.01
19 Iowa		2.72		1.04	0.86
20 Kansas				1.03	0.53
21 Kentucky		2.78		1.15	0.92
22 Louisiana		2.39		1.85	1.24
23 Maine		1.06		0.51	0.39
24 Maryland		1.26		2.08	1.22
25 Massachusetts		3.31		2.83	1.85
26 Michigan		5.30		3.48	2.42
27 Minnesota		1.52		2.05	1.24
28 Mississippi		1.13		0.57	0.43
29 Missouri		3.25		1.73	1.28
30 Montana				0.80	0.41
31 Nebraska		1.39		0.83	0.59
32 Nevada				0.68	0.35
33 New Hampshire				0.38	0.19
34 New Jersey		3.11		2.98	1.91
35 New Mexico				0.51	0.26
36 New York		6.30		7.39	4.55
37 North Carolina		3.11		3.08	1.95
38 North Dakota				0.17	0.09
39 Ohio		6.30		3.44	2.51
40 Oklahoma		1.52		1.45	0.93
41 Oregon		2.19		1.33	0.94
42 Pennsylvania		5.10		5.59	3.48
44 Rhode Island				0.62	0.32
45 South Carolina		1.19		1.24	0.78
46 South Dakota		1.52		0.18	0.27
47 Tennessee		1.99		1.94	1.24
48 Texas		2.78		5.73	3.28
49 Utah		1.26			0.15
50 Vermont				0.09	0.05
51 Virginia		2.19		3.66	2.15
53 Washington		2.78		2.34	1.53
54 West Virginia		0.93		0.69	0.47
55 Wisconsin		0.66		1.52	0.86
56 Wyoming				0.09	0.05
Total	100	100	100	100	100

A variable for the urbanicity of the respondents' home was held on all the surveys. Unweighted values are shown in Table A1.2d below. In 1965-66, the sample was designed to be an urban sample. In 1992-94 the variable urbanicity was derived from the zip codes of the respondent's home (see Directory History of Harmonization/questionnaires). Almost 14% of urbanicity values were missing on this sample. When the percentages are adjusted for missing values (rural=22%, urban=78%) the percentages agree quite well with 1990 Census estimates (rural=25%, urban=75%). Percentages for the 1985 survey also agree well with 1980 Census estimates (rural=26.3%, urban=73.7% - see http://www.census.gov/statab/hist/HS-02.pdf).

However, the rate for rural dwellers in 1975 is high. Weighting the data by the demographic weight supplied (DEMOWGT) reduces the percentage of rural dwellers by 2%. Juster et al, (2001) in the codebook for this dataset remark that urbanicity does not agree with census estimates and that a separate weight for urbanicity is necessary.

Table A1.2d: Urbanicity by survey

Urban Urbanicity of Resp's Home		1 1965-66 USA Szalai national sample	2 1975-76 longitudinal survey – Wave 1	3 1985 Univ Michigan survey	4 1992-94 NHAPS survey
	-8 Missing			0.04	14.5
	0 Rural		36.9	27.1	18.5
	1 Urban	100	63.1	72.9	67.0
Total		100	100	100	100

A1.3 Gender

The unweighted distributions for sex are shown in Table A1.3. The proportions of women in the samples are between 8% to 11% higher than of men. Statistics broken down by age from the US 2000 Census show in the age-group sampled, slightly over 48% of people enumerated in the census are men, compared to nearly 52% of women. This suggests that the bias by gender is between 2% and 3% and reflects the familiar finding that women are more likely to respond to sample surveys than men. Gender distributions vary little by survey.

TableA1.3a : Gender distributions by survey

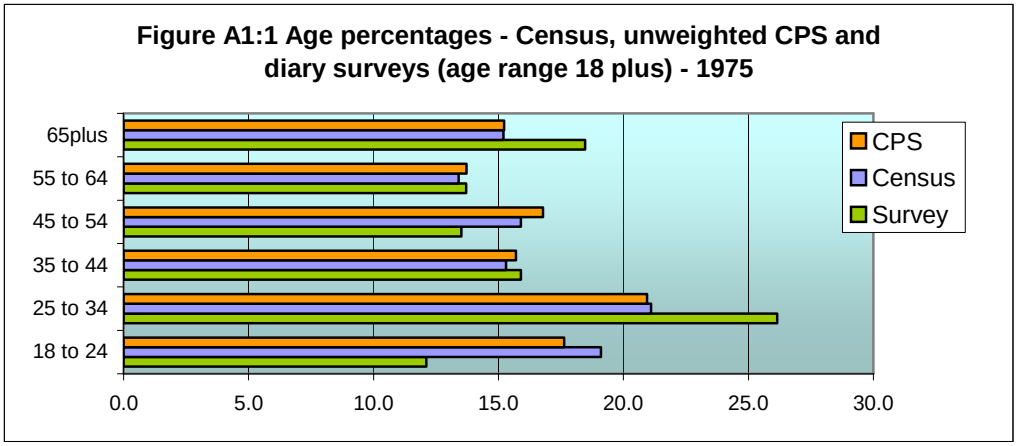
		1965-66 USA Szalai national sample	1975-76 longitudinal survey Wave 1	1985 Univ Michigan survey	1992-94 NHAPS survey	
SEX Sex of Respondent	1 Male	44.9	44.1	46.0	44.2	44.7
	2 Female	55.1	55.9	54.0	55.8	55.3
Total		100	100	100	100	100

Table A4a: Gender distributions of 1975 survey by wave				
	Wave			
SEX Sex of Respondent	1	2	3	4
1 Male	44.1	43.0	42.9	41.9
2 Female	55.9	57.0	57.1	58.1
	100	100	100	100

Table 4a shows that attrition is slightly greater among men than among women in the 1975 survey.

A1.4 Age

It was noted in the first report that younger people are under-sampled. The same finding is shown in the graphs below, which compare Census and Current Population Survey statistics with the slightly altered samples used for the final dataset (excluding 1965-66). Population weights were not available on all the CPS extracts held at Essex, however, the CPS is self-weighting within states and for most variable, the population weights make only small adjustments. It can be seen that the 18 to 24 age-group is under-sampled in both 1975 and 1992-94, while the difference is less marked in 1985. In 1975, it seems clear that the age-groups who are most likely to be living in small households, those aged 25 to 34 and those aged 65 plus, are slightly over-represented. However, this effect is not so marked in 1992-94. As will be seen later there are more households which are likely to be smaller in 1992-94.



The 1965-66 survey does not include 18 year-olds or anyone aged over 65 (except two cases coded 67 and 69). On 1992-94 respondents aged 92 plus are coded into one category. Table A1.4a below shows the distribution of age, with and without selection for one person per household in 1985. This table confirms that more young people are found in the household survey. By construction, no missing values on age are shown on the 1985 data, since respondents were selected to be aged 18 or over. Since other information suggests that many of these respondents were adults a later attempt may be made to impute age for the missing cases. However, in the 1992-94 survey, respondents could be selected by interview type (child or adult) and 2% of age values are missing. In 1965-66 1.3% of age values are missing.

Figure A1:2 Age percentages - Census, unweighted CPS and diary surveys (age range 18 plus) - 1985

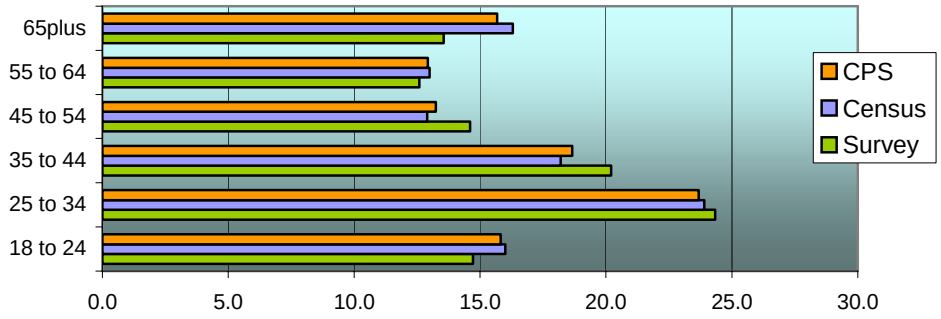


Figure A1:3 Age Percentages - Census, unweighted CPS and diary surveys (age range 18 plus) - 1993

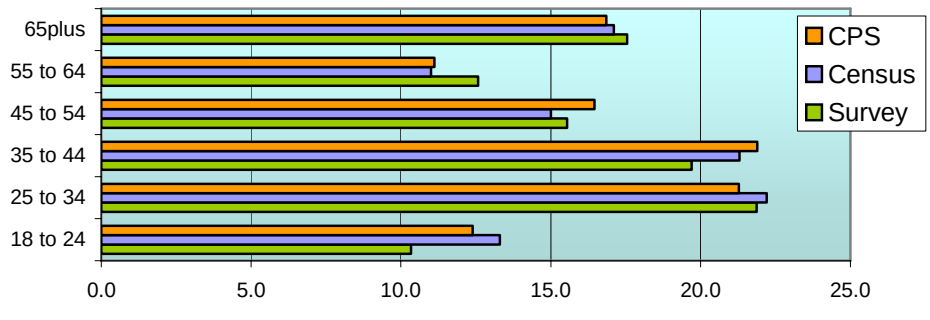


Table A1.4a : Age distributions by survey						
Survey Age (categorized)	1965-66 USA Szalai national sample	1975-76 longitudinal survey	1985 Univ Michigan survey	1985 Survey Person 1 only	1992-94 NHAPS survey	Total
-8.00 missing	1.3	0.2			2.0	1.2
1.00 18 to 24	14.2	12.1	14.7	10.0	10.3	12.1
2.00 25 to 34	24.1	26.2	24.3	25.5	21.9	23.8
3.00 35 to 44	24.0	15.9	20.2	20.8	19.7	19.2
4.00 45 to 54	21.6	13.5	14.6	14.1	15.5	15.6
5.00 55 to 64	13.5	13.7	12.6	13.6	12.6	13.0
6.00 65plus	1.3	18.5	13.6	16.0	17.5	15.1
	100	100	100	100	100	100

Table A1.4b: Waves of 1975 survey						
	Wave	1	2	3	4	Total
AGE	-8 missing	0.3	0.1	0.1	0.1	0.2
	1 18 to 24	12.7	11.7	11.8	12.0	12.1
	2 25 to 34	24.8	26.3	27.2	27.1	26.2
	3 35 to 44	16.0	15.7	16.2	15.7	15.9
	4 45 to 54	13.4	13.6	13.3	13.8	13.5
	5 55 to 64	13.8	14.0	13.5	13.5	13.7
	6 65 plus	19.1	18.6	18.0	17.8	18.5
Total		100	100	100	100	100

Table A1.4a shows that there is little variation in the age structure by wave of the 1975 survey, the largest variation being 1.3% for those aged 65 plus.

Table A1.4b shows the survey distributions of age by sex. Table 5c shows the differences between the proportions found in the CPS and in the TU Surveys. The second column compares the 1965-66 data with the CPS statistics for the age-range 19 to 64. Ignoring the comparison between 1965 and the CPS full age-range, the largest disparity to be found is that between the CPS and the 1975 TU survey statistics, with the TU survey being 6.4% lower for men in the youngest age group. However, disparities between the TU Surveys and the CPS are broadly similar by gender, suggesting that there is no gender bias to the sampling of age groups which differs between the respondents to these two surveys.

Table A1.4c : Comparison of TU surveys with Current Population survey									
	TU					CPS			
	1960s	1970s	1980s	1990s		1965	1975	1985	1995
Men									
1 18 to 24	12.4	11.8	15.0	12.3		14.7	18.1	16.3	12.7
2 25 to 34	24.7	26.7	25.1	24.0		18.1	21.5	24.5	21.7
3 35 to 44	23.8	17.8	20.8	20.2		20.5	16.0	19.2	22.6
4 45 to 54	23.3	12.4	14.5	15.9		19.0	17.1	13.4	16.9
5 55 to 64	14.4	13.5	12.4	13.0		13.9	13.6	13.0	11.2
6 65 plus	1.5	17.8	12.3	14.5		13.8	13.6	13.6	14.8
	100.0	100.0	100.0	100.0		100.0	100.0	100.0	100.0
Women									
1 18 to 24	15.9	12.4	14.5	9.3		14.2	17.2	15.4	12.1
2 25 to 34	24.3	25.8	23.7	20.6		17.8	20.4	23.0	20.9
3 35 to 44	24.9	14.5	19.7	19.7		19.9	15.4	18.1	21.3
4 45 to 54	20.7	14.4	14.7	16.3		18.2	16.5	13.1	16.0
5 55 to 64	13.0	13.9	12.7	12.9		13.8	13.9	12.9	11.0
6 65 plus	1.2	19.0	14.6	21.2		16.2	16.6	17.5	18.6
	100	100	100	100		100.0	100.0	100.0	100.0

Table A14d : Differences in % in age-groups between the CPS and the TU Surveys					
	1960s	1960s age 19-65	1970s	1980s	1990s
Men					
18 to 24	-2.3	-1.8	-6.4	-1.3	-0.4
25 to 34	6.6	3.0	5.2	0.6	2.3
35 to 44	3.3	-0.9	1.7	1.5	-2.3
45 to 54	4.3	0.5	-4.7	1.0	-1.0
55 to 64	0.5	-2.2	-0.1	-0.5	1.8
65 plus	-12.4		4.2	-1.3	-0.4
Women					
18 to 24	1.7	1.5	-4.8	-0.9	-2.8
25 to 34	6.5	2.5	5.4	0.7	-0.3
35 to 44	4.9	0.4	-0.9	1.6	-1.6
45 to 54	2.5	-1.6	-2.1	1.6	0.3
55 to 64	-0.7	-3.9	0.1	-0.2	1.9
65 plus	-15.0		2.4	-2.9	2.5

A1.5 Ethnicity

Ethnicity data was held on the 1970s and the 1990s TU surveys. Race data was held on the 1960s survey. In both the two earlier surveys the data was based on interviewer's observation, with only 'white' or 'Negro' or 'other' given as choices in the 1965 data,

while Asian and Hispanic were also given as choices in the 1975 data. The 1990s surveys used self-reported ethnicity.

Table A1.5a: Condensed ethnicity by survey							
	TUS				CPS		
Ethnic2	1960s	1970s	1990s		1965	1975	1995
1 white	87.7	88.9	83.1		89.8	89.3	84.6
2 black	6.6	8.6	9.8		9.4	9.5	9.2
3 other	5.7	2.5	7.1		0.9	1.3	6.2
	100	100	100		100	100	100
Missing	0.1	0.5	1.5				

Missing values in the TU surveys are shown in the final row of Table A1.5a and it can be seen that they are lower in the two interviewer observation variables. However, the 1992-94 surveys, though yielding a much larger sample, also tend to have more item missing values over all variables. Selecting out the missing values, the percentages can be compared with the CPS proportions. It can be seen that the 1960 survey has a lower proportion of African Americans and a higher proportion of 'other' races. Possibly this reflects the urban, mid-western bias of the sample. For other years the proportions of 'Black' and 'Other' are similar between the CPS and the TUS. However, small overall percentage differences may make a sizeable difference where the group is a small proportion of the sample. No real indication of differences in categorization by question type is seen here, although the 1990s TU surveys (with self-report ethnicity) seem to be slightly higher on minority ethnicity than the CPS, while the 1970s (interviewer observation) seem to be slightly lower. Both sets of unweighted surveys slightly under sample the Black population, with census statistics for 1970 reporting the Black population as 11.1 of the population, and for 1990 12.3%.

Table A1.5b shows the distributions of the more detailed ethnicity variable by gender. It can be seen that African-American women are more likely to be sampled than African-American men, taking into account the higher response rate among women.

Table A1.5b: Ethnicity by survey by gender					
	1970s			1990s	
	Men	Women		Men	Women
Ethnic					
1 White	90.6	87.5		84.1	82.3
2 Black	6.8	10.0		8.5	10.8
3 Asian	0.2	0.4		1.9	1.4
4 Some Other Race	0.3	0.2		1.7	1.7
5 Hispanic	2.1	1.9		3.8	3.8
	100.0	100		100.0	100

Table A1.5c shows attrition over waves by ethnicity. Attrition seems to be slightly higher among black respondents, comprising around 30% of black respondents.

Table A15c: Attrition in the 1975 survey by ethnicity					
	Wave of 1975 survey				
	1	2	3	4	All
Ethnic Ethnic Group Of Respondent					
-8.00 Missing	0.5	0.4	0.3	0.2	0.4
1.00 White	88.4	90.1	91.2	91.9	90.1
2.00 Black	8.5	7.1	6.2	5.5	7.1
3.00 Asian	0.3	0.2	0.2	0.2	0.2
4.00 Some Other Race	0.3	0.3	0.3	0.3	0.3
5.00 Hispanic	2.0	1.9	1.8	1.8	1.9
	100	100	100.0	100.0	100.0

Given the low percentage of African-Americans in the sample, it is not clear that there is much difference in attrition by gender (Table A1.5d below).

Table A1.5d: Attrition in the 1975 survey among African Americans by gender				
	Wave 1	Wave 2	Wave 3	Wave 4
	1	2	3	4
Men	6.8	5.7	4.9	3.8
Women	10.0	8.2	7.2	6.8

A Hispanic variable is held in 1992-94. It refers to Spanish/Hispanic descent rather than current ethnicity.

Table A1.5e: Hispanic in 1992-94 TUS		
HISP Hispanic origin or descent	Frequency	Percent
-8 Missing	124	1.9
0 No	6053	91.2
1 Yes	457	6.9
Total	6634	100

Comparing this statistic with the weighted CPS 1995 for those aged over 17, the percentage is slightly low, with the CPS reporting 9.2% of respondents having Hispanic origins.

A1.6 Educational Level

Table A1.6a shows education level by TU survey. It can be seen that there is a clear gradient towards higher levels of education with succeeding decades. The levels of

respondents in 1965-66 with no or primary education only is lower than in 1975, because of the exclusion of older people from the sample. The bottom row of the table shows the proportions in the lowest educational level for the sample aged 18 to 65, confirming that the lower rate in 1965 was due to the age restriction. The first column for 1985 shows the distribution when young adults are included in the sample and there are fewer respondents with high levels of education, as many of these young people may not have completed their education.

Table A1.6a: Education level by Time Use Survey

Educational Level	1 1965-66 USA Szalai national sample	2 1975-76 longitudinal survey Wave 1	3 1985 Univ Michigan survey	1985 Univ Michigan survey 1st person	4 1992-94 NHAPS survey
-8 Missing Or Dirty On The Case Record	1.6	0.5	1.0	1.0	1.8
1 0 - 8th Grade	12.6	14.9	6.3	6.9	3.1
2 9 - 11th Grade	19.8	14.7	9.6	9.5	7.9
3 High School Graduate	39.3	37.3	42.8	40.1	34.5
4 Some College	14.3	15.3	17.7	16.8	23.9
5 College Graduate	10.1	9.1	15.3	16.7	16.5
6 Post College	2.3	8.1	7.3	9.1	12.3
	100	100	100	100	100
Aged < 66					
0 – 8th Grade	12.6	10.0		4.3	1.6

Table A1.6b: Education level by Year of Current Population Survey

	1965	1975	1985	1995
Education Level				
1.00 0 to 8th grade	27.2	17.3	11.5	7.6
2.00 9 to 11th grade	18.4	15.1	11.9	11.2
3.00 High School Graduate	33.0	36.9	37.9	33.5
4.00 Some College	11.8	16.8	20.1	26.6
5.00 College Graduate	6.1	8.2	10.7	14.2
6.00 Post College	3.5	5.7	7.8	6.8
	100	100	100	100.0

Table A1.6b shows the distributions of education level found in the CPS. As noted in the first report, the Time Use samples are better qualified than the population. This may partly denote the literacy requirement for filling in the diary. Table A1.6c shows attrition patterns in the 1975 survey. More of the better educated remain in the survey, with the biggest decrease among the less educated between Wave 1 and Wave 2.

Table A1.6c: Education level by wave of 1975 survey				
Wave of 1975 Survey	1	2	3	4
Qualification				
1 0 - 8th Grade	16.8	14.9	13.9	13.4
2 9 - 11th Grade	16.7	14.8	13.5	13.2
3 High School Graduate	36.6	37.0	37.9	39.0
4 Some College	15.1	15.7	15.6	15.2
5 College Graduate	7.8	9.4	10.0	10.2
6 Post College	7.0	8.2	9.0	9.1
	100	100	100	100

Table A1.6d breaks down the Time Use Survey qualification data by gender. It can be see that women are concentrated more than men in the middle of the distribution. Fewer women have very poor and fewer women have very high qualifications. Similar though less marked effects are found for the Current Population Survey.

Table A1.6d: Qualifications by survey and gender									
	TUS					CPS			
	1960s	1970s	1980s	1990s		1965	1975	1985	1995
Men									
1 0 - 8th Grade	14.4	15.1	6.8	2.9		28.7	17.8	11.7	8.5
2 9 - 11th Grade	20.6	14.8	9.6	6.9		17.6	14.0	11.3	11.2
3 High School Graduate	33.9	30.0	40.0	33.4		29.2	33.0	35.2	31.9
4 Some College	15.1	15.4	17.3	24.6		12.4	17.9	20.0	25.0
5 College Graduate	12.7	11.7	17.0	18.1		7.0	9.4	11.9	15.0
6 Post College	3.2	13.0	9.4	14.1		5.1	8.0	10.0	8.5
	100.0	100.0	100.0	100.0		100.0	100.0	100.0	100.0
Women									
1 0 - 8th Grade	11.5	14.9	6.0	3.4		26.0	16.9	11.4	8.4
2 9 - 11th Grade	19.7	14.8	9.8	8.9		19.1	16.0	12.5	11.2
3 High School Graduate	44.9	43.2	46.0	36.6		36.4	40.3	40.3	34.7
4 Some College	14.0	15.4	18.4	24.1		11.2	15.9	20.2	27.0
5 College Graduate	8.4	7.2	14.1	15.8		5.2	7.1	9.7	13.3
6 Post College	1.6	4.5	5.6	11.2		2.1	3.8	5.9	5.4
	100	100	100	100		100.0	100.0	100.0	100.0

Table A1.6e below shows the percentage differences between the Time Use Surveys and the CPS by gender. Differences overall are greater for the 1960s surveys due to the restricted age range, and perhaps due also to the more urban sample. In the 1960s surveys, the difference in disparities between men and women on high school graduation is quite high, about 4%. The differences between the CPS and 1975 TUS is also higher for men than for women for post-college qualifications, by about 5%. However, most of the other differences seem to be in the same direction and of similar magnitude.

Table A1.6e: Percentage differences between the TU surveys and CPS by gender				
	1960s	1970s	1980s	1990s
Men				
1 0 - 8th Grade	-14.2	-2.7	-4.9	-5.6
2 9 - 11th Grade	3.0	0.8	-1.7	-4.3
3 High School Graduate	4.7	-3.0	4.8	1.6
4 Some College	2.7	-2.5	-2.8	-0.4
5 College Graduate	5.7	2.3	5.1	3.1
6 Post College	-1.9	5.0	-0.6	5.6
Women				
1 0 - 8th Grade	-14.5	-2.0	-5.4	-5.0
2 9 – 11th Grade	0.6	-1.2	-2.7	-2.2
3 High School Graduate	8.5	2.9	5.8	2.0
4 Some College	2.8	-0.5	-1.8	-2.9
5 College Graduate	3.2	0.1	4.5	2.5
6 Post College	-0.6	0.7	-0.3	5.7

A1.7 Marital status

Table A1.7a below shows marital status by Time Use Survey, excluding 1992-94 where the question was not asked. Table 8b shows the distributions from the Current Population Survey.

Table A1.7a: Marital status by time use survey				
CIVSTAT RESPONDENT'S MARITAL STATUS	1965-66 USA Szalai national sample	1975-76 longitudinal survey Wave 1	1985 Univ Michigan survey	1985 Univ Michigan 1st person only
-8 missing or dirty on the case record	0.0		0.6	0.3
1 MARRIED	81.1	64.4	64.5	59.7
2 SEPARATED,DIVORCED	5.5	11.4	8.1	11.2
3 WIDOWED	3.9	11.7	6.5	10.4
4 NEVER MARRIED	9.4	12.5	20.4	18.4
	100.0	100.0	100	100.0

Table A1.7b: Distributions of marital status from the Current Population Survey			
Year	1965	1975	1985
Married	72.7	68.1	62.1
Separated/divorced	4.7	6.8	9.7
Widowed	8.9	8.5	7.8
Never Married	13.7	16.6	20.4
	100	100	100

It is noticeable that there are fewer widowed respondent in the 1965-66 TUS, presumably due to the exclusion of those aged over 65. Comparing the Time Use Surveys with the CPS, it is noticeable that there are fewer 'never married', presumably because there are fewer young people in most of these surveys. This affects all other percentages. However, the total household statistics for 1985 are similar to those from the CPS, with slightly more married, and fewer separated, widowed or divorced in the CPS. Both sets of surveys show a downward trend in married couples and an upward trend in people who have never married.

Table A1.7c: Changes in marital status over waves of the 1975 survey				
	Wave			
Marital Status	1	2	3	4
1 Married	64.4	65.1	67.4	68.8
2 Separated, Divorced	11.4	10.3	9.5	8.9
3 Widowed	11.7	12.1	11.7	11.4
4 Never Married	12.5	12.5	11.3	10.9
	100	100	100	100

Table A1.7c shows minor changes in marital status over waves of the 1975 survey. More respondents (4.4%) are married and fewer are separated/divorced (2.5%) or 'never married' (1.6%). Perhaps the latter two categories are most likely to move or change address and therefore leave the survey. However, it should be noted that while it was possible to identify respondents who had probably married between waves (0.5%), it was not possible to identify respondents who had separated with certainty, although married respondents with no spouse living in the household increased by about 0.8%. However, this could include spouses living away, for instance, in the Armed Forces. Generally the effects are relatively small.

A1.8 Number of children living in the household

Table A1.8a shows the number of children aged under 18 living in the household. It can be seen that there is an upward trend in households with no children, although the relatively low percentage in 1965 is partly due to the age restriction.

Table A1.8a: Distribution of number of children living in the household over TUS				
Under18 number of children under 18	1965-66 USA Szalai national sample	1975-76 longitudinal survey	1985 Univ Michigan survey	1992-94 NHAPS survey
-8 missing or dirty on the case record	0.3	0.1		1.8
0	39.7	55.5	66.3	74.9
1	17.8	14.8	16.0	9.6
2	19.2	15.0	12.1	9.2
3	12.4	8.2	3.7	3.2
4	5.5	4.3	1.3	1.0
5	2.9	1.5	0.4	0.2
6	1.3	0.4	0.1	0.1
7	0.5	0.2		0.0
8	0.2	0.1		
9	0.1			0.0
10	0.1			
11 more than 10	0.0			
	100	100	100	100

Table A1.8b: Distribution of number of children living in the household over CPS				
	Year			
Number of children aged under 18	1965	1975	1985	1995
None 0	50.7	No data	60.2	No data
1	15.8		17.1	
2	14.5		14.5	
3	9.4		5.6	
4	4.9		1.8	
5	2.5		0.5	
6	1.1		0.1	
7	1.1		0.1	
	100		100	

For the CPS, data was only found for 1965 and 1985 and the question was about the respondent's own children, i.e. stepchildren, grandchildren, etc are excluded. However, a similar trend towards fewer households with children can be seen.

A1.9 Paid work and other economic activity

As noted in the first report, there are inconsistencies both in sampling and in question form between the Time Use surveys. These are described in detail in Directory History of Harmonization/questionnaires. The 1965-66 survey sampled households with at least one member in employment. Turning to the questions, in the earlier surveys, respondents were asked to select one of 5 options (employed, student, housewife, retired, disabled) as their main economic activity, while in 1992-94, respondents given the following options.

EMP: Are you currently:

1 = Employed full time
2 = Employed part time
3 = or not employed at all?
9 = Ref/Not asked

Only respondents filtered through code 3 were asked if they were students, retired, homemakers, etc. and only respondents who were employed fulltime or part-time were asked their work hours. Work hours were asked of all respondents in 1975 and 1985, whatever their main economic activity. Sets of dummy variables were created for the various economic activity statuses and the fulltime and part-time work dummy variables included respondents who had given a non-paid-work activity as their main economic activity. Additionally, respondents in 1992-94 were asked for their work hours in the previous week, while respondents in the earlier surveys were asked for their usual work hours. Respondents working less than 10 hours a week were excluded in 1965-66, possibly accounting for the very low rate of part-time work reported in 1965-66. Given this the statistics reported in Table A1.9a below might be considered artefactual.

Table A1.9a: Distribution of full and part-time work in the time use surveys

EMPSTAT employment status of respondent	1965-66 Szalai & national sample	1975-76 longitudinal survey	1985 Univ Michigan Survey -all	1985 Univ Michigan survey 1st person	1992-94 NHAPS survey
-8 missing	0.8	0.5	2.0	2.1	1.1
1 employed full-time	66.4	50.1	53.4	55.8	53.9
2 employed part-time	2.7	5.6	8.8	8.4	10.4
3 not employed	30.1	43.9	35.7	33.7	34.5
	100.0	100.0	100	100.0	100.0

In particular, while there are trends between 1975 and 1985 towards more paid work, there are actually more people not in paid work in 1992-94 than in 1985. However, similar trends are found in Current Population Survey data, categorizing respondents according to their work hours, as the variable most consistent with the Time Use Surveys. Perhaps demographic ageing is implicated in these trends. Comparisons between the TUS and CPS by gender are shown below.

Table A1.9b: Distribution of paid work in the TUS and CPS by gender										
	TUS					CPS				
	1960s	1970s	1980s	1990s		1965	1965 aged 19 to 65	1975	1985	1995
Men										
Employed full-time	94.6	70.3	67.4	65.9		70.5	80.2	62.9	63.1	63.3
Employed part-time	1.3	4.1	6.8	8.2		4.3	3.5	5.2	5.0	5.1
Not employed	4.1	25.6	25.8	25.9		25.2	16.3	31.9	31.9	31.7
	100	100	100	100		100.0	100	100.0	100.0	100.0
Women										
Employed full-time	44.4	34.6	43.7	45.8		30.6	35.4	33.5	40.8	44.8
Employed part-time	3.8	6.8	10.9	12.3		5.5	5.6	7.2	8.3	8.6
Not employed	51.8	58.6	45.5	41.9		63.9	58.9	59.4	50.9	46.6
	100	100	100	100		100		100	100	100.0

Table A1.9c shows changes in the distribution of paid work through waves of the 1975 survey. It can be seen that for both men and women, levels of employment increase over waves, by about 4% for men and about 7% for women. Quite a large percentage of women who entered employment entered from fulltime family care, while men tended to enter fulltime employment from unemployment or study, and part-time work from retirement/disability. Different employment patterns for men and women are at work here. However, among both men and women, attriters were more likely not to have been in paid work in Wave 1 (men 45%; women 38%).

Table A1.9c: Distribution of paid work in the 1975 TUS by wave and gender				
	Wave			
Employment Status	1	2	3	4
Men				
1.00 employed full-time	70.3	75.3	76.5	74.6
2.00 employed part-time	4.1	3.1	3.1	3.6
3.00 not employed	25.6	21.7	20.3	21.8
	100.0	100.0	100.0	100.0
Women				
1.00 employed full-time	34.6	39.7	39.6	42.3
2.00 employed part-time	6.8	8.8	8.9	8.0
3.00 not employed	58.6	51.5	51.5	49.7
	100.0	100.0	100.0	100.0

Table A1.9d shows the distributions for dummy variables for economic activity for survey. Not all variables could be constructed for the Current Population Survey from the CPS extract held at Essex. Since data for the 1960s is not easily comparable, it is excluded.

Table A1.9d: Dummy variables for economic activity by gender and survey

Dummy variable		TUS				CPS		
		1970s	1980s All	1990s		1975	1985	1995
Unemployed								
Men	YES	4.7	2.9	3.2		6.6	5.6	4.3
Women	YES	5.0	3.9	3.7		4.2	3.8	2.9
Homemaker								
Men	YES		0.7	0.2		0.4	0.5	No data
Women	YES	44.3	20.8	10.4		45.5	32.9	
Student								
Men	YES	5.3	6.0	3.0		3.6	3.4	5.1
Women	YES	3.6	5.8	2.6		3.0	2.9	5.0
Retired or Disabled								
Men	YES	21.1	15.5	18.7		16.9	18.9	No data
Women	YES	10.7	14.1	23.3		4.9	9.3	

Although unemployment rates are within the same range over both surveys, the CPS shows slightly higher rates than the TUS. Possibly this reflects the better educational level of the Time Use Survey respondents. Fewer women describe themselves as fulltime housewives in the 1985 TUS, while more describe themselves as retired. Again this may reflect a greater labor market orientation of the better educated respondents in the TUS. However, it is possible that different methodologies produced this result, or there may be data error. While one quarter of respondents who classified themselves as homemakers also classified themselves as part-time workers, only 9 (about 3%) of these housewives reported any work hours. (We are concerned that there may be an as yet undetected data error here.)

A1.10 Work hours

As noted earlier, the questions and filters were different over different Time Use Surveys. Table A1.10a below shows the percentages on work hours variables affected by these filters.

Table A1.10a: Distribution of work hours by Time Use Survey				
Work hours condensed	1960s	1970s	19880s	1990s
-8.00 missing	0.8	0.5	2.0	2.6
-7.00 NA		43.9	35.7	
-4.00 zero to 10 hours 1965	29.7			
-3.00 routed out in the 1992-94 survey				34.2
1.00 0 to 20	3.3	5.5	8.8	10.4
2.00 21 to 30	3.7	3.1	4.5	4.4
3.00 31 to 40	29.2	26.6	28.0	25.1
4.00 41 to 50	19.3	13.2	12.0	12.4
5.00 51 thru hi	14.1	7.3	8.8	10.9
	100	100	100	100

Table A1.10b confirms the findings from the economic activity variables, with more people in paid work remaining in the survey.

Table A1.10b: Change in work hours over the 1975 survey				
Wave of 1975 survey	1	2	3	4
Work hours				
-8.00 missing	0.5	3.9	3.0	2.6
-7.00 NA	43.9	38.7	37.0	36.8
1.00 0 to 20	5.5	6.2	6.2	6.0
2.00 21 to 30	3.1	3.8	4.4	4.7
3.00 31 to 40	26.6	25.2	27.0	28.2
4.00 41 to 50	13.2	14.1	14.9	14.5
5.00 51 thru hi	7.3	8.3	7.4	7.2
	100	100	100	100

Table A1.10c below breaks down work hours by gender and compares them with CPS statistics. Table A1.10d shows the percentage difference between CPS and TUS statistics.

Table A1.10c: Work hours by decade : TUS and CPS											
		TUS						CPS			
	Work Hours	1960s	1970s	1980s	1980s 1st person	1990s		1965	1975	1985	1995
Men	1.00 1 to 20	2.4	5.5	9.2	8.5	12.5		5.8	7.6	7.3	7.4
	2.00 21 to 30	2.8	3.7	4.1	3.4	5.0		3.7	4.9	5.3	5.9
	3.00 31 to 40	35.7	41.4	42.3	41.8	35.5		47.6	52.1	48.9	45.4
	4.00 41 to 50	31.1	30.8	24.0	24.2	22.4		23.8	20.0	21.9	22.6
	5.00 51 thru hi	28.0	18.7	20.4	22.1	24.5		19.1	15.4	16.6	18.6
		100.0	100.0	100.0	100.0	100.0		100.0	100.0	100.0	100.0
Women	1.00 1 to 20	8.4	16.1	19.9	18.9	20.5		15.2	17.6	16.9	16.1
	2.00 21 to 30	9.4	8.4	10.6	9.1	8.9		9.5	11.7	11.1	11.9
	3.00 31 to 40	52.2	56.8	48.2	50.1	43.9		56.9	57.7	54.7	50.5
	4.00 41 to 50	22.5	13.5	14.1	14.3	16.7		13.0	9.3	12.3	14.5
	5.00 51 thru hi	7.5	5.2	7.2	7.6	10.0		5.4	3.7	5.1	7.0
		100	100	100	100.0	100		100.0	100.0	100.0	100

Table A1.10d: Differences between TUS and CPS by gender and decade				
Workhours	Difference 1960s	Difference 1970s	Difference 1980s (all)	Difference 1990s
1 to 20	3.4	2.1	-1.9	-5.1
21 to 30	1.0	1.3	1.1	0.9
31 to 40	11.9	10.7	6.6	9.9
41 to 50	-7.3	-10.8	-2.1	0.2
51 thru hi	-8.9	-3.3	-3.8	-5.9
	0.0	0.0	0.0	0.0
1 to 20	6.7	1.5	-3.0	-4.4
21 to 30	0.1	3.4	0.4	3.0
31 to 40	4.7	0.9	6.5	6.6
41 to 50	-9.5	-4.3	-1.8	-2.2
51 thru hi	-2.0	-1.5	-2.1	-3.0
	0.0	0.0	0.0	0.0

It can be seen that men are concentrated more in the 31 to 40 hours band per week in the TUS surveys. With the rise in fulltime labor force participation, women in the TUS surveys are also more likely to be in this band when compared to the CPS. Perhaps this is also associated with the higher educational level of the TUS respondents, who may be more likely to be in bureaucratic jobs and may do less overtime.

A1.11 Household Income

Household income data was not collected in 1992-94 and was banded into quartiles in the 1985 dataset, therefore the household income data for the other surveys had to be recoded into approximations to quartiles. Household income was banded into 10 groups in 1965-66 and into 18 groups in 1975. Because of clumping in the higher income bands with three of these containing over half of the cases, the lowest quartile in 1965 contains only 16% of cases. Table A1.11a shows the percentages found in each survey. Since 1965-66 contained only working households, the lower percentage coded into the lowest income group may perhaps be justifiable on empirical as well as practical grounds.

Table A1.11a: Approximate quartiles of household income distribution by survey

Survey	1960s	1970s	1980s 1st person
Approximate quartiles of household income			
Lowest quartile	16.3	20.7	21.6
2nd lowest quartile	27.6	27.6	26.5
2nd highest quartile	24.7	24.6	23.8
Highest quartile	31.4	27.0	28.0
	100	100	100

As noted earlier, an unidentified value 10 was also found in the 1965 dataset, containing 2.2% of cases. Quite large percentages of missing values are found in both 1975 and 1985 (see Table 12b below). In 1975 it was possible to identify the respondent and DK responses were higher where the respondent was not the Head of Household or wife of Head of Household. However, even among heads and wives of heads non-response was over 10%. As noted earlier in the report the variable used in 1975 was Household income in the previous year (1974) rather than current household income.

Selecting out multiple responses from households in 1985 shows a slight increase in households in the poorer half of the income distribution. This is clearly reasonable for household income, since the larger households contain more adults of working age.

Table A1.11b : Distributions of household income by survey

Survey	1960s	1970s	1980s 1st person	1980s
Household Income				
-8.00 missing	2.2	11.8	13.1	12.6
Lowest quartile	15.9	18.3	18.8	16.0
2nd lowest quartile	27.0	24.3	23.0	22.5
2nd highest quartile	24.2	21.7	20.7	21.4
Highest quartile	30.7	23.9	24.3	27.6
	100	100	100	100

Table A1.11c shows changes over waves in the composition of the sample by household income. It can be seen that there is a trend for the lower income households to attrit, with

about 5% fewer in the lowest income band, and increases in the two higher income bands, particularly the highest band. This is consistent with results reported above on the education and employment status of those who remain in the survey over the year Fall 1975 to Summer 1976.

Table A1.11c: Distribution of household income by wave of the 1975 survey						
Wave	1	2	3	4		All
Lowest quartile	20.7	18.1	16.7	16.0		18.2
2nd lowest quartile	27.6	25.4	25.5	25.1		26.1
2nd highest quartile	24.6	26.2	26.9	26.5		25.9
Highest quartile	27.0	30.3	30.9	32.4		29.8
	100	100	100	100		100

A1.12 Data quality in the time diaries

Three broad dimensions of data quality arise in the creation of the heritage files. The first dimension relates to the recording of the diary information. The second dimension arises from the distribution of diaries across the days and seasons of the year. The third dimension relates to the information contained in the diaries.

We encountered a large number of errors in the entry of the diaries. These covered a range of problems, from lines from one household members diary recorded within the diary of another household member, to single episodes being recorded multiple times in the same diary, but with the time of the episode assigned to only one line, to finish times not properly matched with starting times. We have corrected all these problems, and the full documentation of this data cleaning appears in the syntax files for the diaries.

The diaries reflect roughly even distribution across the seasons, but more significant imbalances across the days of the week. The 1965-66 and 1985 studies are most balanced in terms of distribution by day, while the other heritage studies show a strong bias in favor of weekend days. The imbalances in distributions across season and day of the week, however, can easily be corrected by weighting. Unlike other sampling errors, in the case of days of the week and months of the year, we know exactly how many months from each season and how many of each day of the week was present in the sample period, so we can correct this distribution accordingly. While the distribution is uneven across the days, the sample size of the least represented days remains relatively large for all studies. In the 1965-66 and 1975-76 studies, roughly two-thirds of diary days are recorded as normal days, and roughly one-third recorded as unusual days. We await the 2003 ATUS data before we can comment on whether the usual day variable will be comparable with the latest data.

The more significant problem of data quality arises in the content of the diaries, summarized in the tables below. We detail three measures of diary quality. First, we consider the mean and median number of episodes recorded in the diaries. For all surveys and for all years, women recorded more episodes of activity that men. Whether this suggests that women have busier lives than men or that women are better diarists than men we cannot say. The main respondents in the 1975-76 survey also have slightly higher numbers of episodes than the spouses, but this difference is slight. More worrying, though, is the apparent significant drop in the number of reported episodes in the 1992-1994 study. It is unlikely that people became less busy in the USA in the 1990s and more likely that given the depth of other information people were asked to supply in this study compared with the others, that diarists put less effort into completing the daily activity record in the NHAPS study than in the earlier studies.

Low episode diaries raise particular concerns. When diarists give insufficient detail of their day, their diary serves the twin purpose of allowing the over-estimation of time spent in the activities which the diarist did record, and an under-estimation of the activities which the diarist did not record. We propose the benchmark of classifying diaries as of reasonable quality if they have at least 15 recorded activities (unless other information suggests that the diarist had a good reason for having a low activity day, such as being ill in bed). At the very least, we recommend the removal of diaries with particularly low activity counts, or the production of weights that remove these diaries from weighted analysis.

Relatively modest numbers of diaries have low activity counts in the 1965-66 and 1985 data, and just over 15% of diaries have low activities in the 1975-76 data. Nearly half of the diaries in the 1992-94 study, however, have low activity counts (see table A1.12a), which raises significant concerns about the data from this study. It would be difficult for the researcher to determine if statistically significant differences between the 1992-94 data and other surveys reflected survey time use information collection techniques or genuine change over time.

The second measure of diary quality arises in relation to the total number of minutes for which no activity is recorded in the diary. No diaries appear to have missing time in the 1965-66 or 1992-94 studies. In the case of the former study, we suspect the data from diaries with missing time may have been lost. In the case of the more recent data, we suspect that missing time was coded as an extension of the previously recorded activity. The total amount of missing time is highest in the 1985 data, where 5.5% of the average diary is missing.

The third measure of diary quality arises from the presence or absence of basic activities. On any given day, the majority of people engage in some form of sleep or rest, some degree of using the toilet, bathing, and related personal care, eat or drink something, and move from at least one place to another. While on rare days people may not engage in one of these activities at all, the number of days where people actually perform none of two or more of these broad categories is activity is rare, and on such days, people's circumstances are likely to be sufficiently strained that they are unlikely to have kept a

time diary. Around 5% of diaries in the 1975-76 and 1992-94 studies lack two or more basic activities, and fewer than 3% of diaries in the 1965-66 and 1985 studies show such omissions.

The problems of missing time and missing basic activities can be partially addressed through imputation where diaries have missing time gaps of shorter periods (of an hour or less). In contrast with the demographic data and most other forms of survey data, there is an extent to which shorter gaps in diary data reliably may be imputed. The reliability of such imputation arises from the constraints which the activities and location of the diarist on either side of a short missing time period necessarily impose of the range of possibilities of what might have happened in the intervening gap. For instance, if a person is sleeping at home from the early hours, then has a short missing time gap, then is apparently in the car driving to work, we can reasonably suppose that during at least part of the intervening time, this person engaged in some form of unspecified personal care.

We recommend that some limited imputation be undertaken to reduce the degree of problems in the diary data. Imputed activity properly should be marked as such, and we have written four codes into the 91 harmonized activity code frame for this purpose:

(2) imputed unknown personal care
(4) imputed sleep
(71) imputed social activity
(90) imputed travel.

We could add a fifth imputed activity code: (58) imputed out of home activity if you would deem this of value. We have not undertaken an imputation process yet, and Table A1.12b below displays the degree of each of these problems across each dataset, as well as the degree of loss if the bad diaries were removed (before any imputation). In the present test data, the imputation codes are not used, with one exception. In the 1975-76 and 1985 studies, the original codes include a category for time when the diarist was with another person and reported the other person's activity rather than their own activity. We have coded this time from these two studies as (71) imputed social time.

Using imputation can reduce the level of problems in the diaries by up to half. The next question which arises is that of what to do with the diaries which remain of poor quality after imputation. We do not recommend putting such diaries to the USA research community as though the diaries were of the same quality as the rest of the sample. One option for addressing such diaries would be to exclude them entirely from the dataset. Exclusion reduces the risk of use of the dataset for the creation of poor quality research. While such an approach has merit, we do note that some researchers may wish to examine the people who produce poor quality diaries. We therefore recommend a second approach of retaining these poor quality diaries, but weighting these diaries with 0 weights, so that they are effectively excluded from weighted analysis (which is the only form of analysis of time use which should be published using this data), and carefully documenting the nature and degree of the problem diaries in the user guide.

We have not created weights for this test data as the decision about the treatment of the low quality diaries needs to be resolved before we can generate reasonable weights. We

will need to account for the day and season distribution, and the likelihood of actually completing a good diary once a respondent agreed to participate in the study after the decision is taken about what the final definition of the poor quality diaries. We will construct this dimension of the weights ourselves, but will make use of the sample weights provided in the original data to correct for sample selection and participation issues.

Table A1.12a Activity Totals: Means and Median Scores								
Study date	**Mean episodes**			**Median episodes**			**Most frequent diary day**	**Least frequent diary day**
	all	women	men	all	women	men		
1965-1966	27.6	30.4	24.1	26	29	23	all - Tuesday (15.7%) women – Tuesday (16.2%) men – Monday (15.7%)	all - Saturday (13.3%) women – Monday (13.0%) men – Saturday (12.3%)
1975-1976 main sample	22.8	24.6	20.3	21	23	19	all – Sunday (25.3%) women – Sunday (25.3%) men – Sunday (25.4%)	all – Wednesday (8.9%) Thursday (8.7%) women–Wednesday (8.7%) men – Thursday (8.0%)
1975-1976 spouse sample	21.5	23.5	19.1	20	23	19	all – Sunday (25.4%) women – Sunday (25.4%) men – Sunday (25.3%)	all – Wednesday (9.1%) Thursday (9.0%) women–Thursday (8.1%) men – Wednesday (8.8%)
1985	24.1	25.9	22.6	23	25	22	all – Saturday (16.6%) women – Wednesday (16.3%) men – Saturday (16.2%)	all – Sunday (12.7%) Thursday (12.8%) women – Sunday (12.0%) men – Thursday (12.7%)
1992-1994	16.7	17.6	15.6	16	17	15	all – Sunday (19.2%) women – Sunday (19.7%) men – Sunday (18.6%)	all – Friday (8.6%) women – Friday (8.4%) men – Friday (8.9%)
2003							all - women – men –	all - women – men –

Table A1.12b Bad diary counts.				
Study date	**Number & % of diaries missing >90 minutes (pre-imputation)**	**Number & % of diaries including <15 activities (pre-imputation)**	**Number & % of diaries missing 2 or more basic activities (pre-imputation)**	**Total number & % of diaries that would be excluded on any of these criteria (pre-imputation)**
1965	0 diaries 0.0% of diaries	62 diaries 3.1% of diaries	18 diaries 0.9% of diaries	74 diaries 3.7% of diaries
1975-1976 main sample	139 diaries 3.0% of diaries	697 diaries 15.2% of diaries	238 diaries 5.2% of diaries	863 diaries 18.8% of diaries
1975-1976 spouse sample	54 diaries 2.2% of diaries	448 diaries 17.9% of diaries	141 diaries 5.6% of diaries	526 diaries 21.0% of diaries
1985	155 diaries 5.3% of diaries	289 diaries 9.9% of diaries	61 diaries 2.0% of diaries	417 diaries 14.2% of diaries
1992-1994	0 diaries 0.0% of diaries	3098 diaries 41.2% of diaries 7165 diaries 9.5% have <10 episodes	389 diaries 5.2 % of diaries	3144 diaries 41.8% of diaries using <10 acts definition 899 diaries 12.0% of diaries
2003	diaries % of diaries	diaries % of diaries	diaries % of diaries	diaries % of diaries

The Excel file “initial trends”, located is located on the data CD under /History of harmonization/diary/, displays the initial distribution of mean time spent by men, women, and the whole sample on week days and on weekend days across each of the 1-digit activity codes. These 1-digit codes reflect time spent in main activity only. The file contains two worksheets, one showing mean minutes, and the second showing mean minutes converted into a more user-friendly hours and minutes format. These figures are not weighted, and also include the bad diaries without imputation. Consequently, the reader should not place any importance on interpreting this initial set of figures. Nonetheless, the following general issues of note emerge at this point.

- time in personal care, in-home leisure and travel (overall) appear stable over time
- hours of work appear to have fallen for men but increased for women, though men spend more hours in paid work than women
- hours of unpaid work have risen modestly for men and fallen markedly for women, though women still spend more hours in unpaid work than men
- time in education appears to have risen slightly for men
- both men and women spend more time in voluntary activities on weekends than on weekdays, but voluntary time has remained relatively stable, and women spend more time volunteering than men
- men engage in more physical exercise and out-of-home free time activities than women
- the level of use of the media (particularly TV) and computers has increased markedly over time

The “initial trends” file also displays mean time spent in main activity child care, adult care, care of pets, and total care for men and women on week days and weekend days. These initial results suggests that men and women perform similar amounts of pet care and that care of pets has remained stable. The figures also indicate that women perform more child and adult care than men, though total main activity time in care appears to have declined. These figures do not account for who else is present or secondary activity care at this time, and we will return to this topic in greater detail for the final report.

Finally, the “initial trends” file shows the mean time spent inside, outside, in a vehicle, and in an unknown location across the surveys. The first matter of note here is that time recorded being in a vehicle appears to have remained stable across the decades. Time in an unknown location varies considerably. The NHAPS study of 1992-1994 placed great emphasis on exposure to hazards such as sunlight, and the 1975-76 study included a detailed coding frame and a relatively detailed diary. For these two studies, the mean time in an unknown locations is under 25 minutes. Unknown location time is over an hour for the 1965-66 data, and between 4 and 6.5 hours in 1985.

Time spent inside and outside raises more concern. The figures suggest that men spend more time outside than women across the whole period, and that both women and men spend more time outside on weekend days than on week days. Neither of these findings is surprising. Time outside is roughly consistent between 1965-66 and 1975-76m then again between 1985 and 1992-94. What does require explanation is the apparent jump in the reporting of time outside between 1975-76 and 1985. Most likely this jump is an artifact

reflecting the growing importance of concern among social policy researchers with exposure to sunlight, leading to a higher priority for a activity code-frame that accurately records time spent out-of-doors. This issue will need to be carefully documented for users of the data. Nevertheless, the inside/outside variable is not presently generally available to users in other time use datasets, and this variable—in at least the later years—does offer a potentially valuable innovation for future users of this data.

Figure A1.3 Average minutes per day spent out-of-doors by men and women on week and weekend days.