Journal of Official Statistics, Vol. 20, No. 3, 2004, pp. 467–480

The Effect of Coding Error on Time Use Surveys Estimates

Patrick Sturgis[1]

This article presents the results of a coder reliability study conducted as part of the 2000 UK Time Use Survey. Five coders coded the same 40 diaries in which respondents had recorded, in their own words, their activities for every ten minutes over the course of a particular day. Coding was done via a computerised coding system, which enabled coders to view scanned digital images of diaries and access an online coding frame. In addition to an estimate of net aggregate coder reliability, proportion of agreement coefficients are presented for each of the ten main activity codes at the highest level of the coding hierarchy. Reliabilities are also calculated for individual coders. Intra-class correlation coefficients (*Rho*) are then estimated and these are combined with the reliability estimates to produce a variance inflation factor for each of the ten higher order main activity codes (Kalton and Stowell 1979). Some illustrative examples are provided to demonstrate the true standard errors of survey estimates once this coder error has been accounted for.

Key words: Coder reliability; diary surveys; variance inflation.

1. Introduction

Time Use (TU) surveys are becoming increasingly important in addressing a broad range of sociological and policy-related research questions (Gershuny and Sullivan 1998; Fleming and Spellerberg 1999). Such studies are now experiencing their highest levels of interest and widespread implementation since their heyday in the late 1960s when 12 countries participated in the Multinational Time-Budget Research Project (Szalai 1972). Recent TU surveys have been conducted in 62 developed and developing countries (Fisher 2001). In Europe, most countries have conducted TU surveys within the last 15 years and the recent Harmonised European Time Use Surveys (HETUS) project, under the coordination of EUROSTAT, has led to the implementation of methodologically harmonised surveys in 13 member states, with eight more planned to commence fieldwork in the near future (Osterberg 2000).

TU surveys use diaries of activities conducted during the course of a particular day to provide information on what people do with their time, what proportion of time is spent on economically productive activities, leisure pursuits and personal care, including sleep. They afford a unique perspective on the shifting balance between the paid and unpaid sectors of the economy (Gershuny 2000) on how social and cultural change affects work and social practices and have been at the forefront of moves to increase the visibility of the voluntary and domestic sectors and to include unpaid production in Systems of National

[1] University of Surrey, Department of Sociology, Guildford GU2 7XH, UK. Email: P.Sturgis@surrey.ac.uk
Acknowledgment: This research was funded by the UK Office for National Statistics.

Accounts (Short 2000). This, in turn, grows out of the recognition by a range of national and international agencies, including the United Nations, the OECD and the ILO, that measuring and valuing unpaid work is an essential step in improving the status of women worldwide (Rydenstam 1996; Sturgis and Lynn 1998).

1.1. The UK 2000 Time Use Survey

This study was conducted as part of the 2000 UK Time Use Survey (UKTUS 2000). UKTUS 2000 was cofunded by the Office for National Statistics (ONS), a consortium of other government departments and the Economic and Social Research Council (ESRC). The study was conducted according to the Harmonised European Time Use Survey Programme directive and is therefore an important source of comparative time use data. Interviews were completed, with all household members aged eight and over, during 2000 and 2001 at over 6,000 households, with over 11,000 thousand individuals. Each eligible household member was requested to complete two "own words" diaries corresponding to one weekday and one weekend day during a prespecified calendar week. This yielded a total of more than 21,000 completed diaries. In accord with Eurostat HETUS guidelines, respondents completed diaries in their own words rather than selecting activities from a precoded list. Respondents were required to keep a record of all activities conducted during each of 144 ten-minute time slots comprising the diary day. Where activities did not change over contiguous ten-minute time slots, respondents were instructed to indicate this by drawing an arrow from the starting time to the finish time of the activity, rather than write in the same activity over and over again (the diary and other documentation can be downloaded from the Question Bank at the Centre for Applied Social Surveys (CASS): http://qb.soc.surrey.ac.uk/surveys/tus/tus2000.htm).

Although contextual information on secondary activity, location and who the respondent was with was also collected in the diary, the focus of this article is solely on the main activity reported by the respondent.

Activities reported in the diaries were coded to a hierarchical frame, containing 286 unique codes at the lowest (3-digit) level and ten at the highest (1-digit) level of the frame. Response rate at the household level was 61%, with 81% of eligible individuals completing individual questionnaires amongst responding households. Seventy-three per cent of eligible individuals completed at least one diary, giving an estimated net response rate of 43% for the diary component of the survey (Donmez 2002). Coding of diaries for the main survey was performed by a team of seven coders using a computerised, interactive coding system (see Lyberg and Dean (1992) for a generic description of this type of coding system). This system provided coders with an "on-line" coding menu with help and "intelligent" prompts to aid the coding of scanned digital images of the diaries.

1.2. The effect of coder unreliability on statistical estimates

Because verbatim responses to open questionnaire items must be converted to nominal categories on a coding frame, an additional source of error is introduced into the data collection process, relative to using a fixed set of precoded response alternatives. The failure to apply the "correct" code (note that the notion of a correct code is somewhat problematic here, as there is no agreed-upon "gold standard" or criterion measure for the

classification of activities) for a particular verbatim diary entry can happen for a number of different reasons. Coders may be inadequately qualified, motivated or trained, or the task may be performed in an unsuitable working environment; the coding frame may not be well-defined, leading to different activities being assigned the same code or multiple codes applying to the same activity. Where coding tasks are computerised, technical problems with digital imaging and accessing online coding dictionaries will also heighten the probability of misclassification (Lessler and Kalsbeek 1992). Such errors of measurement can be both random and systematic in nature, with systematic errors leading to bias in survey estimates and random errors leading to less precise estimates and a higher probability of Type I errors in hypothesis testing (Kish 1962). Random error also attenuates estimates of the structural relationships between variables (Blalock 1963).

Coding error can be of two types, correlated and uncorrelated (Cochran 1977; Kalton and Stowell 1979; Campanelli et al. 1997). Correlated error pertains when the tendency to apply the wrong code is systematically associated with one or more coders, uncorrelated error when the application of a wrong code is randomly distributed across coders. For categorical variables, coder error is usually assessed through coder reliability studies, in which n different coders code the same open-ended responses. Uncorrelated error can then be assessed by computing the Proportion of Agreement, $\overline{P}$, which gives the percentage of all paired comparisons between coders on which the coding pair agreed. Sometimes, Kappa – a derivative of $\overline{P}$– is used to correct for the possibility of chance agreement (Cohen 1960; Fleiss 1971). Computing the intraclass correlation coefficient, *rho*, assesses correlated error. *Rho* is an estimate of within cluster homogeneity. It can be interpreted, in the context of coder reliability, as the per cent of variance in the observed variable attributable to idiosyncratic coder errors.

This distinction between correlated and uncorrelated (or "simple") coder error is important because each error type has a different effect on statistical estimates. While both correlated and uncorrelated error serve to increase standard errors, relative to measurements made without error, simple error is accounted for in basic standard error formulae, so any estimates we make will incorporate this error as standard (Cochran 1968). Thus, while clearly undesirable in that it reduces effective sample size, the usual estimator is unbiased under uncorrelated coder error (Biemer and Trewin 1997).

As with the effect of clustering, however, correlated coder error must be estimated separately and incorporated into variance estimates. This is done using the following formula (Kalton and Stowell 1979; Groves 1989; Biemer and Stokes 1991; Biemer and Trewin 1997; Campanelli et al. 1997):

$$ceff = (1 + p_c(m - 1)(1 - K_i)) \quad (1)$$

Where *ceff* is the increase in variance due to correlated coder error, p_c is the intraclass correlation for coders, m is the average coder workload and K_i is the reliability of the ith code, usually measured by either proportion of agreement or Kappa. Thus, although values of p_c may be very small (Campanelli et al. (1997) found values in the range 0–0.0058 for occupational coding in the UK), if coder workloads are high, variance inflation factors can still be of considerable magnitude.

For instance, a value of $p_c = 0.02$ in conjunction with a code reliability of 0.75 and an average workload of 1,000 questionnaires will produce a variance inflation factor of 6.

Taking the square root of the variance inflation factor gives the inflation factor for standard errors (*ceft*). In this instance standard errors would be underestimated by a factor of 2.45 – a very considerable loss of precision.

1.3. Time use survey design considerations

Because of the complexity of the information that time use studies seek to survey, there are a great many design issues, which must be taken into consideration in order to achieve the most accurate estimates for a fixed cost. This mostly involves trade-offs between what would be considered theoretically optimal and what is an acceptable burden to place on respondents. For instance, the time period usually chosen as the sampling unit would, on theoretical grounds alone, be at least a week. However, the keeping of a diary over a seven-day period in a general population survey is generally considered to be so burdensome that its negative effect on the response rate would counteract any gains that this time sampling might bring in terms of capturing population heterogeneity (Harvey 1999). Therefore, two days is generally considered the optimal design in order to achieve a balance between time coverage and reasonable response burden (Osterberg 2000).

Another issue which often vexes designers of TU surveys is whether to collect diary data in respondents' own words or whether to provide a precoded list of activities from which respondents can select in order to describe their activities over the course of the day. Using own words diaries allows far more detail and heterogeneity in activity patterns to be captured (Lingsom 1979; Gershuny 1992; Harvey 1993). It also affords a great deal of flexibility in that different coding frames may be applied to the raw data, even retrospectively from some future vantage point (Gershuny and Sullivan 1998).

However, because of the burden that own words diaries place on the average respondent, in terms of time and effort as well as information processing capacity, others have argued for the potential benefits of the precoded approach (Sudman and Ferber 1979; Zmud 2001). Using a precoded list of activities generally imposes less of a burden on respondents and may, therefore, have a beneficial influence on response rates and data quality. Precoded diaries may also be less likely to disadvantage less literate respondents relative to well-educated counterparts, again leading to more valid and reliable estimates and comparisons between population subgroups. However, debate concerning the relative costs and benefits of each approach has usually been based on their intuitive appeal rather than any sort of empirical comparison. The aim of this study is to take steps towards filling this gap by quantifying the effect of the coding of own words diaries on the accuracy of statistical estimates. Specifically, the study focuses on how correlated and uncorrelated coder error affect the *precision* of activity estimates from own words time use diaries.

2. Study Design

Five members of the coding team of seven who worked on the main survey coded the same 40 diaries using the on-line system employed for the main stage coding. Because, for each of the 40 diaries, one of the five coders had already coded the diary in question as part of the main stage survey, the recoding sample for this study actually comprised 160 rather than 200 diaries.

The other two coders in the team of seven had left the data collection agency by the time this study was conducted. The five coders were asked not to discuss the coding of these diaries with one another, either before or during the period of study. No steps were taken to verify whether this instruction was followed strictly. However, it is safe to assume that coders will generally consult and discuss aspects of the coding task with one another in most real coding situations. Because our general interest is in making inferences to error rates in the main stage of surveys, coder interaction should not be a particular concern, so long as the amount of discussion (and, therefore, nonindependence of ratings) is not greater than it would be in main stage coding environments.

Unfortunately, the design for this coder reliability study was not drawn up with this particular analysis in mind. Rather, the data were collected with the aim of producing proportion of agreement statistics for a prespecified set of codes to meet contractual obligations between the data collection agency and the survey sponsor. This meant that the diaries that form the basis of this study were not selected randomly but purposively, with the aim of ensuring that satisfactory reliability estimates could be obtained for all of the ten main activity codes at the one-digit level and for a range of key activity codes at lower levels of the coding frame. Analysing the data file of coded diaries did this, and iteratively selecting diaries until a total of 40 had been selected provided a good coverage of all ten main activity codes. Comparing the distribution of main activities from the completed survey with that from the subset of diaries used for this study, as shown in Figure 1, reveals a very similar pattern. However, it is important to note that this aspect of the design limits the generality of the findings of this study, as inferences to the full survey can only be reliably made if the analysis is performed on a random sample of all completed diaries (Kalton and Stowell 1979).

Many different estimates can be produced from the activity variables, each of which, due to differing levels of aggregation, is likely to have a different error structure.

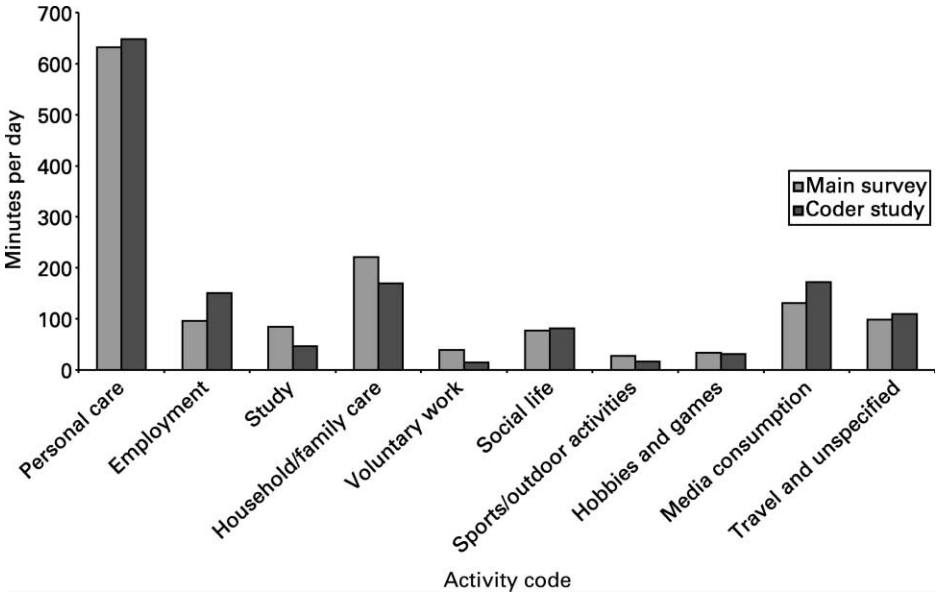

Fig. 1. Distribution of main activities for main survey and coder study

For instance, it is possible to produce estimates from the UKTUS concerning, *inter alia*, the proportion of people conducting particular activities during the course of a week, a day, or at particular times during particular days. Additionally we might be interested in the mean amount of time spent on particular activities, or in including contextual variables to estimate rates and proportions of activities conducted with other individuals, alone, or in particular locations.

The unit of analysis in this study is the ten-minute time slot, 144 of which comprise a diary day. This unit of analysis was selected, despite the fact that it is undoubtedly one of the less common estimates analysts will be interested in producing from the survey, because it was the primary coding unit. That is to say, respondents were requested to record what they were doing for every ten minutes of the day and coders were required to enter a code accordingly. This means that comparing reliabilities for codes applied to larger blocks of time was not feasible, as this would confound the activity and activity duration parameters of the coding task. Technically, it would be possible to compare the estimates of mean time spent on particular activities – certainly a parameter of greater analytical interest. However, as each coder in this study coded only 40 diaries in total, the sample size was too small to make sufficiently precise and representative comparisons at this aggregated level.

3. Analysis

As there were 40 diaries, each comprising 144 main activity codes, and 5 coders participated in the study, the total number of activity codes was 28,800 and the total number of paired comparisons was 57,600. The measure of reliability employed in this study is the proportion of agreement, $\overline{P}$, which gives the percentage of all paired comparisons that were given the same code for the same ten-minute time slot across all five coders. This was preferred to Kappa (Cohen 1960; Fleiss 1971), which makes an adjustment for chance agreement, because $\overline{P}$ is conceptually clearer and additionally, when the number of codes on the frame is large, the probability of random agreement is low and K and P become statistically equivalent (Elliot 1983; Campanelli et al. 1997). Furthermore, Kappa's correction for chance agreement is relevant only under conditions of statistical independence of raters. Since coder ratings in most coding tasks are clearly not independent, the appropriateness of this correction to actual agreement levels is very questionable (Maclure and Willett 1987).

The proportion of agreement for the coding frame as a whole may, through aggregation, mask considerable variation in reliability across codes and coders. $\overline{P}$ can, however, be broken down into the various constituent indices of which it is a weighted average, namely the reliability of individual coders and of individual codes (Kalton and Stowell 1979). These additional measures provide insight into the nature and extent of variation in the reliability of the coding and may be of use for improving the reliability of the frame in future applications.

3.1. Overall and individual coder reliabilities

Table 1 shows the overall proportion of agreement for the whole coding frame and for the five individual coders at the one-digit level for all activities and for "waking time" only.

Table 1. Proportion of agreement ($\overline{P}$) for whole coding frame and individual coders

Coder	$\overline{P}$–1-digit level Including sleep	$\overline{P}$–1-digit level Excluding sleep
Coder 1	95%	92%
Coder 2	95%	93%
Coder 3	94%	92%
Coder 4	94%	91%
Coder 5	94%	91%
Overall	94%	92%

These are presented separately because sleep takes up approximately one third of all diary entries and has a high inter-coder reliability. Including sleep, therefore, inflates the aggregate reliability of the code frame.

The $\overline{P}$ coefficients in Table 1 show very high aggregate levels of reliability, with 94% of paired comparisons matching when sleeping time is included and 92% matching when sleep is excluded. There is also very little variation in reliability across each of the five coders, the largest gap being the 2% between Coder 2 and Coders 4 and 5 when sleep is excluded from the analysis.

To form an idea of the substantive meaning of these figures, it is helpful to know that four coders applying code *X* and one coder applying a different code, *Y*, results in a $\overline{P}$ of 60% (i.e., 60% of paired comparisons are the same), while three coders applying code *X* and two coders applying a different code, *Y*, results in a $\overline{P}$ of 40% and so on. Aggregate agreement of 94%, then, (from the estimate including sleeping time) represents a very close correspondence between the codes applied by the five coders. By this estimate, the increase in the variance of the sample mean relative to a measure with 100% reliability is only around 6% for codes at this level of the hierarchy.

This overall level of reliability should be considered as more than satisfactory by traditional standards. Landis and Koch (1977), for instance, propose that values larger than 75% represent "excellent" rates of agreement between coders. Values falling between 40% and 75% represent "fair to good" levels of agreement and only reliabilities less than 40% do they rate as "poor" inter-coder agreement. In similar studies, Kalton and Stowell (1979) found overall code reliabilities ranging between .60 and .88 for the questions they examined, while Campanelli et al. (1997) found an overall reliability of .78 for occupational coding in the UK. On these comparative criteria too, then, the reliability of the Time Use Survey coding would appear to be extremely good. There are, of course, many differences in the coding frames used in these studies, so such headline comparisons should be interpreted with care.

3.2. Reliabilities of individual codes

As was noted earlier, the aggregate reliabilities presented in Table 1 are a weighted average of the reliabilities of the individual codes, which together constitute the frame. It is possible, therefore, that with a large number of codes, these aggregate figures might mask a good deal of variation across individual codes, with heavily used, reliable codes

obscuring the poor reliabilities of less frequently used ones. We can obtain the reliability of individual codes by taking the conditional probability that the second of two randomly selected coders will apply the code in question, given that the first coder has already applied that code (Fleiss 1971; Kalton and Stowell 1979). Let Q_j be an estimate of the conditional probability that the second of two randomly selected coders applies code j, given that the first coder has already applied that code, and q_j denote the overall proportion of codings that applied code j. Then:

$$\overline{P} = \sum q_j Q_j \tag{2}$$

Table 2 shows the proportions of agreement for all ten codes at the highest, one-digit, level of the coding frame.

Table 2. Proportion of agreement ($\overline{P}$) of individual codes at 1-digit level

Code	% of all codes	$\overline{P}$
Personal care (V0)	43.6	97.2
Employment (V1)	6.4	95.7
Study (V2)	5.8	96.0
Household/family care (V3)	15.4	91.2
Voluntary work (V4)	2.7	87.4
Social life (V5)	5.5	78.6
Sports/outdoor activities (V6)	1.8	82.3
Hobbies and games (V7)	2.5	89.9
Media consumption (V8)	9.0	94.4
Travel and unspecified (V9)	6.7	86.3

The reliabilities range in value from a low of 78.6% for code 5 (social life) to a high of 97.2% for code 0 (personal care). Adopting Landis and Koch's range criteria, all ten codes at this level can be described as having "excellent" reliability (Landis and Koch 1977). Nonetheless, it would indeed seem to be the case that there is quite considerable variation in the reliabilities with which different codes in the frame are applied. The highly reliable codes 0 for personal care and 1 for employment, which amount to 50% of all codes applied, are clearly having a major effect on the overall reliability of 94%, masking the poorer reliabilities of other, less frequently applied, codes.

One factor that may underlie this variation in code reliability is the manner in which respondents recorded different types of activity in the diary on the UKTUS. As was noted previously, respondents were directed, for both "sleep" (reliability = 99.1%) and "working in main job" (reliability = 92.4%), to simply draw a continuous line through contiguous diary time periods to indicate time spent on these activities. This means that there is little room for disagreement between coders as to the appropriate code to apply for these periods of activity, which serves, in some senses, to artificially inflate the reliability of the frame as a whole. Concentrating solely on the overall reliability of the frame in this

type of study may well, therefore, present an unrealistically optimistic view of the loss of efficiency of estimators as a result of simple coder error on key survey estimates.

3.3. Correlated coder error

Next we turn to the correlated component of coder error, that is, the proportion of the variance in activities attributable to the idiosyncratic application of codes across individual coders. Table 3 shows estimates of the intra-class correlation coefficient (*rho*) for codes applied at the one-digit level of the frame. Also shown in Table 3 is the 95% confidence interval of *rho*, the reliability of the code ($\overline{P}$) and the associated variance inflation factor (*ceff*), assuming average coder workloads of 3,000 (this is the mean number of diaries that was coded by each coder on the main stage survey (21,000/7)). *Rho* is calculated from a one-way analysis of variance, using coders as the factor variable,

$$Rho = S_a^2/\left(S_a^2 + S_b^2\right) \tag{3}$$

where S_a^2 is the between interviewer variance and S_b^2 is the within interviewer variance (Kish 1962; Groves and Magilavy 1986). (*Rho* was calculated using the "loneway" procedure in Stata 7.0, with standard errors and confidence intervals estimated using Gleason's formula (Gleason 1997). Bootstrapped estimates of the variance of *Rho* were almost identical to those presented in Table 3, contact author for details of these analyses.) The final column in Table 3 (*ceft*) is the squared root of *ceff* and indicates the proportional increase in standard errors due to correlated coder error. Estimates of *ceff* and *ceft* were obtained using Equation 1 on page 3 (Kalton and Stowell 1979; Campanelli et al. 1997).

Table 3. Correlated coder error and variance inflation estimates for one-digit codes

Activity code	*rho*	95% confidence interval		$\overline{P}$	*ceff*	*ceft*
		Lo	Hi			
Personal care	0.004	0.0014	0.0381	97.2	1.34	1.16
Employment	0.002	0.0000*	0.0368	95.7	1.26	1.12
Study	0.006	0.0004	0.0639	96.0	1.72	1.31
Household/family care	0.006	0.0016	0.0559	91.2	2.58	1.61
Voluntary work	0.000	0.0000*	0.0206	87.4	1.00	1.00
Social life	0.016	0.0043	0.1285	78.6	11.26	3.36
Sports/outdoor activities	0.019	0.0016	0.1706	82.3	11.08	3.33
Hobbies and games	0.034	0.0089	0.2484	89.9	11.29	3.36
Media consumption	0.013	0.0036	0.1071	94.4	3.17	1.78
Travel and unspecified	0.016	0.0045	0.1296	86.3	3.00	1.73

Ceff and *ceft* assume average coder workloads of 3,000; * = truncated at zero.

All but two of the estimates of *rho* (for employment and voluntary work) are significantly different from zero at the 95% level of confidence. There is, then, clear evidence of a significant effect of correlated coder error on the variance of activity estimates at the one-digit level of the coding frame. While these effects are, in the main,

significantly different from zero, there is still considerable variation in their order of magnitude, ranging from zero for "voluntary work" to nearly 3.5 % for "hobbies and games."

Given the high average coder workload of 3,000 diaries on the UKTUS 2000, even such (comparatively) small percentages of the total variance attributable to coder idiosyncrasy have major effects on the precision of any estimates we might make from the main survey data. For example, with a *rho* of 0.034, a code reliability of 0.9 and a workload of 3,000 diaries, the variance inflation factor for "hobbies and games" is 11.3. Translating this into the more intuitively meaningful measure of percentage increase in standard error (*ceft*), we see that standard errors for univariate point estimates of this activity should be 3.4 times larger than would be estimated by conventional formulae, assuming a simple random sample.

Table 4 shows how these estimates of correlated and simple coder error affect a range of actual survey estimates. For each of the ten codes at the one-digit level of the code frame, Table 4 presents – from the full UKTUS data – the estimated proportion of the UK population, aged eight or above, participating in the activity in question, at the time in question, on an average day. Also presented in Table 4 is the standard error assuming a simple random sample and zero correlated coder error and the standard error incorporating this error component as estimated in this study. Incorporating these estimates of coder error results in substantial increases in standard errors and concomitant widening of confidence intervals. For several of these estimates, the loss of precision due to correlated coder error yields an effective sample size of less than 2,000 (effective sample size, *neff*, can be calculated by taking the ratio of the full sample size to the design effect (*deff*) (Groves 1989)).

Table 4. Point estimates and standard errors for % UK population doing different activities with and without correction for correlated coder error

Activity	Time of day	Point estimate	S.E. without coder error	S.E. with coder error
Personal care	08:20–08:30	52.4%	0.34	0.42
Employment	14:40–14:50	18.7%	0.27	0.40
Study	11:20–11:30	7.3%	0.18	0.30
Household/family care	16:50–17:00	19.5%	0.27	0.24
Voluntary work	15:30–15:40	2.1%	0.10	0.44
Social life	22:30–22:40	3.5%	0.13	0.10
Sports/outdoor activities	13:20–13:30	12.0%	0.22	0.75
Hobbies and games	16:20–16:30	5.2%	0.15	0.52
Media consumption	21:20–21:30	43.4%	0.34	0.61
Travel and unspecified	16:00–16:10	16.6%	0.26	0.71

As the full data contain over 21,000 diary records, the true cost of coder error is, clearly, very high. Most, if not all estimates produced from own words diary surveys do not incorporate coder error into variance estimation. Yet, failing to take the correlated component of coder error into account in statistical analyses of this form of data will result

in a substantial increase in the rate of Type I errors and a general overconfidence in the reliability of population inferences.

4. Discussion

This study has demonstrated the existence of considerable unreliability in the coding of activities from own words time use diaries. This was shown to be a result of both correlated and simple coder error. Although little variation was found between individual coders, apparently high reliability for the frame as a whole was shown to mask considerable variation across individual codes, with frequently used reliable codes masking the unreliability of less frequently used ones. Despite the fact that, at the overall level, the loss of effective sample size due to uncorrelated coder error was only 6%, for particular codes at the one-digit level the corresponding figure rose as high as 22%. This demonstrates the importance of disaggregating overall reliabilities into their constituent elements when making an assessment of the reliability of a coding frame in this type of study.

Despite the clear undesirability of and need to minimise the uncorrelated component of coder error, conventional variance estimators include this component as part of their estimate. The same, however, is not true of correlated coder error. These results indicate not only that there is a significant component of correlated coder error on the overall variance of main activities but also that most estimates of precision produced from the 2000 UKTUS (and other TU surveys with similar designs) will likely be underestimated and that, consequently, the strength and reliability of structural relationships as well as differences between groups will be overestimated.

A clear implication of these findings for future time use surveys is the need to integrate monitoring programmes of the coding task within the main stage programme of fieldwork (Morganstein and Marker 1997). These should be implemented early on in the fieldwork period to locate and minimise any weaknesses in the code frame or amongst individual coders. Periodic monitoring on a smaller scale should then be sufficient to ensure the continued effective implementation of the coding task. These findings suggest that it would also, undoubtedly, be cost-effective to increase the number of coders working on this type of study in the future and for commissioning agencies to write this into contracts with data collection agencies.

For instance, doubling the number of coders working on the 2000 UKTUS from seven to fourteen would have reduced the variance inflation factors presented in Table 3 by approximately 25% to 30% (i.e., the formula for estimating *ceff* in Table 3 now assumes average workloads of 1,500 rather than 3,000). Gains made by increasing the number of coders must, of course, be balanced against the increased cost of employing and training a larger number of coders to a sufficient level of proficiency (Martin et al. 1995). However, given that this and previous research has shown that it is not uncommon to find values of *rho* of 2% and above, keeping the average coder workload to 500 or below (and assuming an average code reliability of 0.8) would ensure an *upper bound* variance inflation factor of three.

Finally, what does this study tell us about the relative merits of "own words" versus pre-coded diaries for future TU surveys? On the one hand, the high variance inflation factors

suggest that the own words design may not be the preferred option. However, it is important to recognise that evidence of coder unreliability in own words diary designs does not mean that there is no unreliability in measurements obtained using a precoded diary schedule. Indeed, the difficulty of obtaining estimates of response error in precoded diaries may itself be a reason for preferring the own words approach in the first place. At least with the latter option it is possible and practically feasible to detect problems in the frame using studies like the one reported in this article and to implement improvements, in the form of increased training, amendments to the frame and modifications in the design and management of the coding procedure. This study, of course, has focused on the effect of coding error on *variance estimation*, without touching on the equally important issue of response *bias*. In order for Time Use researchers to make fully informed choices regarding the relative merits of own words and precoded diaries, future research will need to address this, perhaps less tractable, issue.

5. References

Biemer, P.P. and Stokes, S. (1991). Approaches to the Modeling of Measurement Errors. In Measurement Errors in Survey, P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds). New York: Wiley, 487–516.

Biemer, P.P. and Trewin, D. (1997). A Review of Measurement Error Effects on the Analysis of Survey Data. In Survey Measurement and Process Quality, L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds). New York: Wiley, 603–632.

Blalock, H. (1963). Making Causal Inferences for Unmeasured Variables from Correlations Among Indicators. American Journal of Sociology, 69, 53–62.

Campanelli, P., Thomson, K., Moon, N., and Staples, T. (1997). The Quality of Occupational Coding in the United Kingdom. In Survey Measurement and Process Quality, L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds). New York: Wiley.

Cochran, W. (1968). Errors of Measurement in Statistics. Technometrics, 10, 637–666.

Cochran, W. (1977). Sampling Techniques. New York: Wiley.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20, 37–46.

Donmez, C. (2002). Technical Report of the UK 2000 Time Use Survey. Office for National Statistics, London.

Elliot, D. (1983). A Study of Variability in Occupational and Social Class Coding – Summary of Results. OPCS Survey Methodology Bulletin, 144, 48–49.

Fisher, K. (2001). General Notes on the MTUS. Institute for Social and Economic Research, http://www.iser.essex.ac.uk/mtus/technical.php

Fleiss, J. (1971). Measuring Nominal Scale Agreement Among Many Raters. Psychological Bulletin, 76, 378–382.

Fleming, R. and Spellerberg, A. (1999). Using Time Use Data: A History of Time Use Surveys and Uses of Time Use Data. Statistics New Zealand, Wellington: New Zealand.

Gershuny, J. (1992). Time Budget Research in Europe. Eurostat, Brussels.

Gershuny, J. (2000). Changing Times: Work and Leisure in Post-industrial Societies. Oxford University Press, Oxford.

Gershuny, J. and Sullivan, O. (1998). Sociological Uses of Time-Use Diary Data. European Sociological Review, 14, 69–85.

Gleason, J. (1997). SG65: Computing Intra-class Correlation and Large Anovas in Stata. Stata Technical Bulletin, 35, 25–31.

Groves, R. (1989). Survey Errors and Survey Costs. New York: Wiley.

Groves, R. and Magilavy, L. (1986). Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys. Public Opinion Quarterly, 50, 251–266.

Harvey, A. (1993). Guidelines for Time Use Data Collection. Social Indicators Research, 30, 197–228.

Harvey, A. (1999). Guidelines for Time Use Data Collection and Analysis. In Time Use Research in the Social Sciences, W. Pentland, M. Powell Lawton, A. Harvey, and M.A. McColl (eds). Kluwer, Amsterdam.

Kalton, G. and Stowell, R. (1979). A Study of Coder Variability. The Journal of the Royal Statistical Society, Series C, 28, 276–289.

Kish, L. (1962). Studies of Interviewer Variance for Attitudinal Variables. Journal of the American Statistical Association, 57, 92–115.

Landis, J. and Koch, G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, 33, 159–174.

Lessler, F. and Kalsbeek, W. (1992). Nonsampling Error in Surveys. Wiley, New York.

Lingsom, S. (1979). Advantages and Disadvantages of Alternative Time Diary Techniques. Statistisk Sentralbyrå, Oslo, Norway.

Lyberg, L. and Dean, P. (1992). Automated Coding of Survey Responses: An International Review. Statistics Sweden.

Maclure, M. and Willett, W. (1987). Misinterpretation and Misuse of the Kappa Statistic. American Journal of Epidemiology, 126, 161–169.

Martin, J., Bushnell, D., Campanelli, P., and Thomas, R. (1995). A Comparison of Interviewer and Office Coding of Occupations. Joint Proceedings of American Statistical Association and American Association of Public Opinion Research, Survey Research Methods. American Statistical Association. Washington, DC, 1122–1134.

Morganstein, D. and Marker, D. (1997). Continuous Quality Improvement in Statistical Agencies. In Survey Measurement and Process Quality, L Lyberg, P Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds). Wiley, New York, 475–501.

Osterberg, C. (2000). Methodological Guidelines of Harmonised European Time Use Surveys - With Reference to Experiences of the European Time Use Pilot Surveys, Eurostat.

Rydenstam, K. (1996). The European Time Use Survey: Methods, Properties of and Potential Uses of Data. In Seminar on the European Time Use Survey. Office for National Statistics, London.

Short, S. (2000). Time Use Data in the Household Satellite Account. Economic Trends, October.

Sturgis, P. and Lynn, P. (1998). The 1997 UK Pilot of the Eurostat Time Use Survey. GSS Methodology Series 11.

Sudman, S. and Ferber, R. (1979). Consumer Panels. American Marketing Association, Chicago.

Szalai, A. (1972). The Use of Time: Daily Activities of Urban and Suburban Populations in Twelve Countries. Mouton, The Hague.

Zmud, J. (2001). Designing Instruments to Improve Response: Keeping the Horse Before the Cart. Proceedings of the International Conference on Transport Survey Quality and Innovation, Kruger National Park, South Africa.

Received September 2002
Revised July 2003