Journal of Official Statistics
Vol. 6. No. 4, 1990, pp. 411–419

Reliability and Validity of Time Budget Data: Children's Activities Outside School

Ian Plewis, Rosemary Creeser, and Ann Mooney[1]

Abstract: Two methods of collecting time use data are compared. A time budget interview study designed to collect information about young children's educational activities outside school is described. Non-response was low, with the telephone used for three quarters of the contacts. Variance components for the different levels of the design were estimated using a maximum likelihood approach, both for continuous and for binary data. Estimates of moderate reliability were obtained for three contacts per child for the amount of time spent in an activity. The validity of the time budget estimates was probably higher than for stylized estimates obtained from the same respondents by direct questioning.

Key words: Time budgets; reliability; validity; variance components; telephone surveys; non-response.

1. Introduction

The way in which individuals allocate their time to different activities has long been of interest to social researchers, and studies of these allocations have been called time budget studies (Szalai 1972). However, "representing the expenditure of time is one of those subject matters where the reliability and validity of data are extremely sensitive to details in the manner of data collection" (Scheuch 1972). Some of these methodological issues are faced by any researcher wishing to collect retrospective data, others are specific to time budget studies.

Acknowledgements: We would like to thank Colm O'Muircheartaigh, Charles Owen, Barbara Tizard and the editor for comments on earlier drafts of this paper. The research was supported by a grant from the Economic and Social Research Council to the Thomas Coram Research Unit.

[1] Thomas Coram Research Unit, Institute of Education, University of London, 41 Brunswick Square, London WC1N 1AZ, England.

Time budget data can be collected by direct observation, or subjects can be asked to complete a diary for a predetermined period or they can be interviewed, either face-to-face or by phone. In an interview, respondents can be asked to recall all their activities for a chosen period by answering what is essentially a long, open-ended question ("recall time budget" estimates). Alternatively, they can be asked how much time they spent, say, last week or normally spend each week, in a particular activity. These estimates are referred to by Juster and Stafford (1985, ch. 1) as stylized estimates.

Direct observation of individuals over time eliminates the burden on subjects to remember or to record their activities but at the cost of possibly changing their behaviour. The method is generally both too intrusive and too expensive to be considered practicable for most questions, although it has been used, for example, to estimate how

much time children spend in various activities at school, where observers are not unusual (see Tizard, Blatchford, Burke, Farquhar, and Plewis 1988, ch. 4).

A strength of self-completed diaries is that, providing the respondent fills in the diary frequently during the day, the method can reduce recall bias by putting less strain on respondents' memories than some interview approaches do. However, diaries do present problems; response rates are often low, subjects may not provide data of sufficient detail unless previously instructed by an interviewer, they may omit potentially embarrassing activities (also a problem with interviews), and may change their activities as a result of completing a diary. Diaries can also be expensive, especially if they are not delivered and returned by mail, and if incentives are paid to subjects to complete diaries. Nevertheless, diaries are used in the analogous field of family budget studies, a well-known example being the British Family Expenditure Survey (FES). Given the similarities between expenditure and time use – or the market and non-market behaviour of households – self-completion diaries can be used to collect time use data.

When we consider different ways of collecting information by interview, then a stylized approach will usually be cheaper as only one contact per respondent is needed whereas reliable, or stable, data at the individual level will normally only be obtained for the time budget approach with more than one contact (or more than one diary) per respondent (see Kalton 1985). Also, stylized estimates can be obtained in a survey where time use is only one of several variables of interest but, because of the interview time needed to get a complete account of the chosen period, it will often be difficult to collect very much additional information using the recall time budget approach. Comparisons of time budget and stylized estimates of time use from interviews have led Robinson (1985, p. 59) to conclude that "despite the attractiveness of their much lower collection costs, stylized estimates seem unacceptable sources of data for making serious time use projections for the population."

The justifications for the conclusion that time budget estimates are acceptable whereas stylized estimates are not, are as follows. Firstly, recall time budget estimates are very similar to estimates obtained from the same sample by intensive approaches such as using electronic beepers: subjects recorded what they were doing at randomly selected moments during the day as determined by the signal from the beeper. (However, no comparisons appear to have been made with direct observation.) Secondly, stylized estimates of time use for a week tend to produce total times greater than 168 hours. Finally, aggregate comparisons of time use for particular activities such as housework, doing voluntary work and, of particular relevance here, playing with and helping children, show that stylized estimates are usually higher than recall time budget estimates. By not suggesting activities to respondents, it is likely that the time budget method lessens the bias of over or under-reporting arising from social desirability.

Comparing recall time budget and diary estimates, Juster (1985, p. 88) concludes that although the latter may give slightly more valid data, "the difference does not appear to justify a cost difference that may be of the order of three- or four-to-one." Lyberg (1989) reaches a similar conclusion, finding few differences between the two modes of data collection in a large Swedish study.

The balance of the research evidence suggests that if a recall period of 24 hours is used for weekdays and 48 hours for weekends, then valid data can be obtained from time budget interviews. However,

there have been very few studies which have compared stylized and time budget estimates for the *same* respondents. One of the aims of the study described here was to discover how best to obtain accurate data at the individual as well as at the aggregate level on the extent to which six-year old pupils are helped with reading, writing, and maths (the 3Rs) by their parents and others outside school. The design of the study enabled comparisons of stylized and time budget estimates to be made for a series of more detailed variables than most of those discussed in Juster and Stafford (1985) and Lyberg (1989). It also enabled estimates of intra-individual variation to be obtained. The study therefore provides a further contribution to research in the measurement of time use, and the way certain types of questions in surveys are answered. Substantive results from the study are given in Plewis, Mooney, and Creeser (1990).

2. Study Design

In the light of the research discussed above, recall time budget data were collected on the way the six-year old children in the sample had spent the previous day, using the phone as much as possible. It was thought that six-year olds were too young to be interviewed directly in this way and so one of their parents or guardians, usually their mother, responded for them. The interviewers went through the 24 hours of the previous day (or, on Mondays, the previous weekend) and asked the parent to give a sequential account, with timings, of what the child had been doing during that time. The respondents were not asked direct questions about the child's activities but the interviewers probed to get as much detail as possible, particularly at those times of the day when 3R activities might have taken place, if only for a short time. The explanation given to the respondents about the purpose of the study was that we were interested in learning more about children's activities outside school to complement our knowledge about their activities within school.

The children were chosen from a sample of 20 inner London infant schools, 18 of which had been in the earlier longitudinal study described in Tizard et al. (1988). From 19 of these schools (one school did not want to participate in this project), 230 children were selected. As in the earlier project, the sample was restricted to two ethnic groups – white children whose parents were born in the U.K. and black British children of Afro-Caribbean origin. An attempt was made to select at least one white boy, one white girl, one black boy and one black girl from each school and to have approximately equal numbers from these four groups in the sample as a whole. It should be noted that the sample is not a probability sample of children attending inner London infant schools.

In a small pilot study of 40 children, four sets of data had been collected for each child, of which one was for a weekend and the others for three different days of the week. As a result of this pilot work, it was clear that none of the reliabilities were likely to be high (intra-class correlations were generally less than 0.3), and that the increment in reliability from having four rather than three days would not compensate for the reduction in sample size which would be necessary. Thus, it was decided to collect data for just three "days" in the main study (a weekend counting as a day here). (Not all respondents provided data for a weekend in the main study.) Also, because it was found to be particularly difficult to collect data on Saturdays for Friday's activities, Fridays were sampled at half the rate of the other five days. Each respondent was assigned a pattern of contact days (not dates) and the

three interviewers (all of whom were white and female) could only collect data for these days, although the order was not important. Interviewers were also expected, as far as possible, to keep two weeks between each interview. All contacts with a respondent were by the same interviewer. It was clear from the pilot study that telephone interviewing was feasible, and in the event 77% of all the contacts were by phone. (All respondents were sent an advance letter from their children's schools.) If the initial contact was by phone then, in nearly all cases, all subsequent contacts were by phone. However, over half the respondents who were initially interviewed at home were later contacted by phone. The main explanation for this was that the original phone numbers were supplied by the schools and were sometimes either unknown or incorrect. Thus, the proportion of phone contacts rose from 70% at the first contact to 82% at the third contact (Table 1).

At the end of the first contact, basic demographic information was collected. At the end of the final contact, stylized questions were asked about whether the child had read in the previous 24 hours, how often, in general, parents heard the child read in a week, etc. Most of these questions were the same as those used by Tizard et al. (1988) and provided data on frequency but not on duration. Clearly, these questions could not have been asked earlier for fear of influencing the respondents' reports of their children's activities.

The frequency and duration of the following 3R activities were coded from the time budgets: reading aloud, reading on own, being read to, writing on own, writing with other, maths on own and maths with other. However, in this paper, only reading aloud, being read to and combined measures of writing and of maths are discussed. It was necessary to make some essentially arbitrary decisions about coding. For example, flipping through books was coded as reading on own. And on those occasions when the respondent reported that the child was involved in more than one activity – e.g., watching TV and looking at a book – then a fraction of the period was allocated to the

Table 1. Sample details

	Total	% of net sample	Ethnic Group		Mode of contact	
			Black	White	Home visit	Telephone
Addresses issued	230					
Out of scope	12					
Net sample	218	100	112	106	75	143
Respondents						
First contact	196	90	95	101	59	157
Second contact	187	86	88	99	38	149
Third contact	158	72	71	87	29	129
Non-response						
First contact*	22	10	17	5	16	6
Later contacts	38	17	24	14	n.a.	n.a.
Total non-response (%)	28	–	37	18	n.a.	n.a.

*of which 5 refused and 17 were never contacted

3R activity. There were some activities, particularly those children did on their own, when it was only possible to code that it happened and not the duration, and there were occasions when children were out of the home with relatives or friends when it was not possible to find out how they spent their time. The coding was done by the interviewers soon after the contact.

We see from Table 1 that the response rate for the first contact (90%) was high for an inner city survey. Also, 81% of those receiving a first contact had all three contacts. Non-response was higher for blacks than it was for whites throughout; the total response rate was 63% for blacks and 82% for whites. It is not possible to give data on response by mode of contact for the reason discussed earlier. Some of the marked decline in response between contacts two and three can be accounted for by the fact that the interviewers had to achieve a contact for a particular day at that stage, rather than having the choice of two or three days which they had earlier.

3. Reliability

The study design generated a three-level data set: schools, children within schools and contacts (i.e., days) within children. Alternatively, we can say the data were produced by a three-stage sampling process. We can regard variation between schools and variation between children within schools as "true" variation, with variation between days within children as "error" variance (although see below). Methods for estimating variance components, and hence reliability (or stability) coefficients for designs of this type are discussed by Plewis (1988) in the context of psychometric generalizability theory. Note that the focus here is on estimating the reliability of our measures for a (random) child. There is, in addition, sampling error for the child and school means which can be estimated in the usual way. The only error variance that we were able to estimate was that induced by sampling days within children. We were unable to estimate measurement error variances such as interviewer variance because, for cost reasons, respondents could not be randomly assigned to interviewers. Hence, all references to reliability in this paper are restricted to this one component of error, and the estimates should be regarded as upper bounds for generalizability as a whole.

Before estimating between day variation for the amount of time spent in the four activities, it was necessary to establish whether there was any trend across the school term, for if there were then some of the between day variation would have been better regarded as true rather than as error variance. However, there was no evidence of trend for these variables (and no a priori reason for considering other functions). For being read to and writing, but not for reading aloud and maths, there was evidence that more time in these activities was coded at weekends than on weekdays. Consequently, because not all respondents provided data for a weekend, the weekend data were weighted to take account of this. The estimated variance components (between children and between days within children), intra-class correlations (or one-day reliabilities), and reliabilities for the mean amount per contact (or three-day reliabilities) for the four variables are given in Table 2. (There was no significant variation between schools for amount of time.) The estimates were obtained using an iterative maximum likelihood approach incorporated into the program VARCL (Longford 1988a). This deals with unbalanced designs and so it was possible to use information from all respondents, not just those with all three contacts.

Table 2. Variance components, intra-class correlations and reliabilities (amount)

Variable	Variance component			Intra-class correlation	Reliability
	Mean* (mins/week)	Child (c)	Day (d)	$c/(c + d)$	$c/(c + d/3)$
Reading aloud	42	50.0	108.9	0.31	0.58
Being read to	45	48.0	85.5	0.36	0.63
Writing	38	59.9	147.4	0.29	0.55
Maths	15	1.51	74.0	0.02	0.06

*The medians are considerably lower than the means.

The estimates should be viewed with caution as the variables' distributions are certainly not normal; they are very skewed with many scores of zero and are also "lumpy" with non-zero scores usually being multiples of 5. However, it is reassuring to note that the variance components estimated from the more traditional expected mean squares approach (using the NESTED procedure in SAS (1985)) are very similar. If four contacts had been made in the main study, the corresponding reliabilites would have been 0.64, 0.69, 0.62, and 0.08. The estimated reliability for maths is very much lower than for the other three variables, a point we return to later. Note that the estimated intra-class correlations from the main study are somewhat higher than the (imprecise) estimates from the small pilot study.

We also estimated the variance components for the probability that an activity took place on a particular day. The basic data are then binary – either an activity did or did not take place – and the variance components were estimated using a quasi-likelihood approach with a logit link as described by Longford (1988b). The intra-class correlations are somewhat higher for the probabilities than they are for the mean amounts of time but the rank order is the same. The estimated reliabilities – 0.71 for reading alound, 0.72 for being read to, 0.64 for writing and 0.52 for maths – should be treated with caution in that there are only four possible observed values and the error variance for true probabilities of 0 and 1 must be zero so that the error variance cannot be independent of the true values.

4. Comparing Data Collection Methods

Here we look at the differences between stylized and time budget estimates for three of the four variables of interest (stylized estimates for maths were not collected). Table 3 gives the results, based on the same sample of respondents in each case and including, for greater precision, respondents from the pilot study. For comparability, the time budget data have been weighted up to produce weekly estimates so that, for example, 5-7 times a week and 3 occurrences of the activity are treated as equivalent frequencies. The stylized estimates come from questions asked at the final contact (essentially, how often does the activity take place), the time budget estimates are based on all three contacts.

There are large differences between the two types of estimates for each variable. It is very likely that the stylized estimates are a good deal too high (although we cannot be sure for we do not know the "true" values) and possible explanations for this are discussed below. It is also possible that the time budget estimates are too low. Some evidence on this, for children reading, comes

Table 3. Comparisons of stylized and time budget estimates (Percent)

Frequency	Reading aloud (n = 198)		Read to (n = 198)		Writing (n = 198)	
	Stylized	Time budget	Stylized	Time budget	Stylized	Time budget
5–7x/week (3 occurrences)	43	12	35	10	43	6
3–4x/week (2 occurrences)	27	24	32	33	20	21
1–2x/week (1 occurrence)	24	25	19	15	29	28
< 1x/week (0 occurrences)	6	39	14	42	7	45

from responses at the final contact to further questions about whether the child had read aloud or read on its own during the period covered by the diary, again including the pilot study respondents. Seven per cent (9 out of the 128 who did not report reading aloud in the diary question) claimed that the child had read aloud; 15% (19 out of 129) claimed that the child had read on its own. Even if all these respondents had genuinely forgotten – which is perhaps unlikely – this could not account for such large differences between the two sets of estimates.

Although the distributions for the stylized and time budget estimates are different, there is nevertheless some agreement between them. Using weighted kappa to measure agreement, the estimates are 0.30 for reading aloud, 0.44 for read to and 0.20 for writing. The associations (measured by Kendall's τ_b) are 0.40, 0.50, and 0.29 respectively. However, these associations are not high and it is interesting to note that, whereas the association between the time budget estimate of reading aloud and mother's education is positive (τ_b = 0.25) as one might expect, the association is essentially zero when the stylized estimate is used. This goes against Robinson's view (1985) that stylized and time budget estimates produce similar patterns of demographic correlates.

5. Discussion

An important goal of social science is accurate measurement. The evidence presented in this paper reinforces the view previously expressed that stylized estimates of time use will often be inaccurate. However, the reliabilities, or stabilities, of the time budget estimates are not high, about 0.6 for amount of time spent on reading and writing, based on information from three diaries each covering a weekday or weekend. And it needs to be borne in mind that these stabilities provide upper bounds to what psychometricians call generalizability – there are other sources of error variance such as interviewer variance which are not accounted for. Nevertheless, it is also true that the median three-day reliability for the 21 activities listed by Kalton (1985) was only 0.49 and so our values are not unusually low. It is interesting to note that to attain a reliability of 0.8 for the mean amount of reading aloud in this study – the sort of

figure one would expect for an attainment test, for example – would have required nine contacts. Such intensive data collection is unrealistic for most studies. Nevertheless, it is possible to estimate, as Plewis et al. (1990) do, the parameters of statistical models which include relatively unreliable time use variables by using the correction techniques described by Fuller (1987).

The estimated reliability for amount of maths (see Table 2) is very low and not significantly different from zero, although the estimate for the probability of doing maths is higher. Very little maths activity was reported for these children – none at all for two thirds of the sample and a mean of only two minutes per day. Clearly, the time budget method used here is unsuitable for estimating individuals' time in activities which occur both rarely and fleetingly. (Kalton gives very low reliabilities for time spent on medical care, on home improvements, and at spectator events.)

The stylized estimates of time spent in 3R activities are probably too high for a number of reasons. It is likely that replies to questions of this kind – which either explicitly or implicitly require respondents to provide an average figure – are influenced by parents' wish to present themselves in a good light. In recent years in London, parents have been strongly encouraged to hear their young children read, for example. This social desirability effect could work in two ways: firstly, by raising the overall level of reporting and secondly by leading respondents to focus on those weeks when a lot of 3R activities took place rather than on the average level. It is also possible that particular sub-groups of the population have a greater tendency to over-estimate than others. However, there was no evidence from these data to suggest that the agreement between the stylized and time budget estimates varied by mother's education, ethnic group, and child's sex; for all these groups, the stylized estimates were too high by a similar amount. Nor was there any variation by mode of interview. Also, the stylized estimates in Table 3 are very close to the corresponding stylized estimates obtained by Tizard et al. (1988); for example, for reading aloud, their distribution was 44%, 20%, 24%, 12%, compared with 43%, 27%, 24%, 6% in Table 3. Thus, there is no evidence that the different contexts of the two studies influenced the stylized estimates.

It is now widely believed that telephone interviewing, combined with face-to-face interviews for households without phones (a dual mode approach), can give good data for a variety of questions (see, for example, Sykes and Collins (1988)). Certainly the cost advantage that interviews have over self-completion diaries applies much more to telephone interviews than it does to face-to-face interviews. In this study, the initial response rate for phone contacts was higher than for home visits, although this is, of course, a comparison of different sub-populations. The face-to-face interviews were longer (the means were 43 minutes and 25 minutes) but more 3R activities were reported in the phone contacts; for example, seven minutes per day for reading aloud compared with four minutes for home visits. However, as both reading aloud and the probability of a phone contact were related to mother's education and as one would not expect longer interviews to produce less complete information, there are no strong grounds for supposing that the mode of interviewing has affected the data obtained.

Although the methodological issues in this paper have been discussed in the context of a particular time budget study, they have more general ramifications. This is especially true of within subject variation over time which is essentially random. This occurs in

a number of guises but is not often measured. Consequently, within subject variation is often wrongly amalgamated with between subject variation which can, in turn, lead to both biased and inefficient estimates of parameters in statistical models. And, although we have no "true" values against which to judge the validity of our two approaches, there are strong grounds for supposing that the time budget estimates are more valid than the stylized estimates. Researchers intending to use a stylized approach for estimating time use should often be advised to consider alternative approaches which employ time budget interviews, self-completion diaries or both.

6. References

Fuller, W.A. (1987). Measurement Error Models. New York: John Wiley.

Juster, F.T. (1985). The Validity and Quality of Time Use Estimates Obtained From Recall Diaries. In Time, Goods and Well-Being, edited by F.T. Juster and F.P. Stafford, Ann Arbor, MI: Institute for Social Research, p. 88.

Juster, F.T. and Stafford, F.P. (Eds.) (1985). Time, Goods and Well-Being. Ann Arbor, MI: Institute for Social Research.

Kalton, G. (1985). Sample Design Issues in Time Diary Studies. In Time, Goods and Well-Being, edited by F.T. Juster and F.P. Stafford, Ann Arbor, MI: Institute for Social Research.

Longford, N.T. (1988a). VARCL Manual. Princeton, NJ: Educational Testing Service.

Longford, N.T. (1988b). A Quasilikelihood Adaptation for Variance Component Analysis. Proceedings of the Statistical Computing Section, American Statistical Association.

Lyberg, I. (1989). Sampling, Nonresponse and Measurement Issues in the 1984/85 Swedish Time Budget Survey. Paper presented to U.S. Bureau of the Census Fifth Annual Research Conference, Washington DC.

Plewis, I. (1988). Estimating Generalizability in Systematic Observation Studies. British Journal of Mathematical and Statistical Psychology, 41, 53–62.

Plewis, I., Mooney, A., and Creeser, R. (1990). Time on Educational Activities at Home and Educational Progress in Infant School. British Journal of Educational Psychology (forthcoming).

Robinson, J.P. (1985). The Validity and Reliability of Diaries Versus Alternative Time Use Measures. In Time, Goods and Well Being, edited by F.T. Juster and F.P. Stafford, Ann Arbor, MI: Institute for Social Research, p. 59.

SAS (1985). SAS Users' Guide. Cary, NC: SAS Institute.

Scheuch, E.K. (1972). The Time-Budget Interview. In The Use of Time, edited by A. Szalai, The Hague: Mouton.

Sykes, W. and Collins, M. (1988). Effects of Mode of Interview: Experiments in the U.K. In Telephone Survey Methodology, edited by R. Groves et al., New York: John Wiley.

Szalai, A. (Ed.) (1972). The Use of Time. The Hague: Mouton.

Tizard, B., Blatchford, P., Burke, J., Farquhar, C., and Plewis, I. (1988). Young Children at School in the Inner City. Brighton: Lawrence Erlbaum.

Received March 1989
Revised July 1990