British Journal of Psychology (1994), **85**, 169–179 *Printed in Great Britain*

Observer error in the use of momentary time sampling and partial interval recording

Maurice J. Murphy* and **Alex Harrop**

Centre for Psychology, Liverpool John Moores University, Room 308, Trueman Building, Webster Street, Liverpool L3 2ET, UK

The accuracy of 60 first-year psychology degree students using either the method of momentary time sampling (MTS) or partial interval recording (PIR) was estimated by agreement measures in an independent groups study. The purpose of the study was to investigate possible method differences in a design that maximized their comparability. After training, observers were randomly allocated to MTS or PIR methods and, after further practice, recorded either one, two or three behaviours (namely, reading, writing, hand-clasping) from a constructed 10-minute videotape of human studying behaviour from which a criterion record had been derived. Results showed MTS introduced significantly less error into observers' records than PIR across all levels of complexity. However, there was no generally significant increase in error with increasing complexity and PIR showed relatively more accuracy in recording writing behaviour. Despite the apparent support offered to practitioners for the use of MTS in behavioural investigations, the experimenters qualify their findings and indicate the need for a more extensive series of experiments comparing these methods.

When behaviour is recorded non-continuously, the two commonest methods are momentary time sampling and partial interval recording, often generically referred to as time sampling methods. Momentary time sampling (MTS) is in use when 'the occurrence of particular behaviour(s) is noted at a specific moment of time' (Murphy, 1987, p. 214). Partial interval recording (PIR) is defined as that method 'in which an interval is scored as one of occurrence if a response occurs in any portion of it' (Repp, Deitz, Boles, Deitz & Repp, 1976, p. 501n). PIR can itself be employed in two distinct modes: as continuous or 'end-on' PIR in which observation intervals are consecutive and no distinct recording period is provided for; or as non-continuous PIR in which a distinct recording (non-observed) period follows each observational interval. Unless otherwise specified, references to PIR here are to the former continuous mode. This experimental study makes use of the fact that an MTS 'moment' can correspond to the final 'moment' of a PIR interval, permitting simultaneous comparison of both methods with the same material.

It follows from the definitions that MTS and PIR have differing inherent characteristics. For MTS, error is random, while, for PIR, systematic error is in-built since an interval may be scored as an occurrence when behaviour occurs for even a tiny fraction of it. Partial interval recording is only free of systematic error in the

* Requests for reprints.

event that behaviour is continuous or precisely matches the observation intervals; otherwise, PIR is bound to overestimate behaviour duration. Consider, as an example, a behaviour recorded as occurring in 60 out of 120 observation intervals of 15 seconds each in a 30-minute session. Proportionately, the duration of the behaviour occurring appears to be 0.5, i.e. 60/120, though the behaviour may only have occurred for 60 short bouts averaging three seconds each, or 0.1 of the observation session. The degree of overestimation depends on the bout durations of the behaviour observed and their distribution within the session. Furthermore, in general, it follows that the larger the proportion of the session that the observed behaviour occupies, the smaller can be the degree of overestimation (and vice versa).

Another inherent characteristic that differentiates the two methods is their sampling base. For MTS this is a sample occurring as, for example, one 'moment' every 10 seconds. For PIR the observer records the product of continual observation throughout the session (provided PIR is used in the continuous mode). Momentary time sampling is thus likely to miss some behaviour that does occur and PIR will theoretically miss none.

Given these inherent error properties, can the use of either method be justified or can any basis be found to prefer one over the other? These controversial issues cannot adequately be debated within the framework of an empirical report. Recent argument supporting and questioning the validity of time sampling methods may be found, respectively, in, for example, Bakeman & Gottman (1987) and Martin & Bateson (1988). Practically, such methods must be considered worth investigation since they are so widely used. Kelly (1977) reported that 41 per cent of studies in one journal devoted to behavioural analysis used one or other methods. Murphy (1987) offers more recent examples and interested readers will find many other examples together with discussion of issues around observer agreement measures in the wide literature on behaviour modification (e.g. Hersen, Eisler & Miller, 1980). Time sampling is often considered appropriate to multiple behaviours, intermittent behaviours and behaviours to which it is difficult to assign precise duration. It can be claimed as a valid measure of amount of behaviour. In addition, time sampling methods have one advantage that continuous (frequency or duration) methods do not normally possess, the possibility of comparing observer agreement (which is the focus of this study) point for point.

The inherent properties of these two methods of recording behaviour have been extensively empirically examined by means of simulation techniques by, for example, Repp *et al.* (1976), Powell, Martindale, Kulp & Bauman (1977), Murphy & Goodall (1980) and Brulle & Repp (1984). From their comparison of MTS and PIR conducted by computer simulation, Harrop & Daniels (1986) concluded that neither method estimated frequency accurately and MTS was more accurate than PIR where a duration estimate was wanted. However, PIR was the more sensitive and conservative indicator of change in either rate or duration. Previously, again using computer simulation, Harrop & Daniels (1985) had pointed to the need for caution in using MTS when percentage of behaviour was low and observation instants far apart. Subsequently, Harrop, Daniels & Foulkes (1990) warned that investigators needed to be wary of the use of MTS in behavioural investigations because of its low sensitivity in recording change and its inability to estimate absolute duration when

the occurrence duration of behaviour is, or has become, low. In contrast, it was maintained that, despite PIR's inability to estimate duration accurately and its conservatism, 'Using PIR, we can at least be confident that if change does occur then the method of recording will stand a good chance of detecting change' (p. 125). Importantly, Harrop *et al.* (1990) pointed out that they based their conclusions on the results of computer simulation studies and emphasized that 'there is still the open question of whether one method can be used more accurately and effectively than the other by an observer' (p. 126). It is important to note that, in using the term 'accuracy' from now on, we shall be referring only to the correctness of observers' records and not to time sampling's ability to estimate behavioural dimensions accurately.

The use of human observers is likely to increase the variability to be superimposed upon those inherent sources of error previously outlined. To date it is an open question, however, whether observer variables affect MTS or PIR differentially. Even those who see accuracy as almost wholly contingent on training and monitoring of observers (e.g. Barton & Ascione, 1984) offer no evidence that MTS or PIR might differ substantially between themselves in the degree of training and monitoring required to establish and maintain reliable levels of accuracy. Yet, it is likely that some error may occur differentially between the two methods, independent of other factors. For example, in continuous PIR the observer may look away to record and hence miss some displayed behaviour, whereas using MTS there is likely to be ample time to record between instances. Alternatively, it may be easier to see if behaviour has occurred during an interval than at an instant.

Neither logical analysis nor simulation studies can decide between these alternatives. In consequence, to take the comparison between MTS and PIR a stage further, it was decided to compare the performances of independent observers using the two recording methods. To supply comparisons with 'objective' data (see Murphy & Goodall, 1980) videotaped material for which a criterion record could be constructed was used. In addition, the authors follow the recommendation of Dorsey, Nelson & Hayes (1986) that, with limited training, accuracy is enhanced with a small number of codes. It was proposed to compare the recording accuracy of MTS and PIR observers, after limited training, coding identical samples of constructed videotaped material, simulating human behaviour at low levels of complexity. The accuracy of method use would be inferred from criterion and inter-observer agreement measures using Cohen's (1960) kappa statistic, which takes account of chance agreement.

Method

Participants

There were 60 observers, all students in their first year of modular courses with psychology as a component of their degree. The experimental study was undertaken as part of these students' normal course work and selection was based wholly on their attendance at psychology methods sessions. The age range was from 18 to 65+ years with 80 per cent in the 18-33 age group. There were 17 males and 43 females. Observers were naïve with respect to the materials and procedures used and they were not told the objectives of the experiment in advance.

Materials and apparatus

Three distinct observational materials were used, two for practice and the third for the experimental observation. Both practice and experimental videotape material were drawn arbitrarily from 10-minute sequences of constructed videotape made by a research assistant. In each sequence the female research assistant simulated aspects of 'studying' behaviour. There was no verbal behaviour. A single pinging audio tone was superimposed on the taped sequences every 10 seconds, lasting 0.5 second, and a double tone every 60 seconds. Each sequence featured very similar behaviours and differed only in the alternation of different behaviours and bout lengths.

Behaviour displayed can be broadly categorized as being on-task or off-task. While on-task, i.e. studying, the subject displays reading, writing and infrequent page-turning behaviour. While off-task, the subject displays a variety of behaviours that include drinking from a cup, pencil-sharpening, gazing about and various forms of self-attention, such as arranging hair, adjusting glasses and hand-clasping.

From the first 10-minute sequence, a computer analogue was constructed by one of the authors as follows. A criterion duration record was prepared, categorizing only on- and off-task behaviours. These were then represented by the presence or absence of symbolic screen objects displayed by a BASIC computer program. As with the original material, audio tones were programmed every 10 and 60 seconds. The computer analogue was intended to familiarize observers with the use of both MTS and PIR and a criterion occurrence record was prepared from it for the observation as it was to be recorded with MTS used for the first 30 instants and PIR for the second 30 intervals. A two-minute demonstration exemplifying events and audio cues was also programmed. This complete program is referred to as the computer analogue in the succeeding text.

The two 10-minute videotaped sequences were labelled practice sequence A and experimental sequence B. Reading and writing were chosen as typical study behaviours to observe exhibiting what were believed to be medium-high and medium-low levels of inference with equal bout occurrence in the observation. In contrast, an off-task behaviour of fairly low frequency and low inference, hand-clasping, was also chosen.

A criterion duration record was constructed for the experimental sequence B. There were 22 bouts of reading totalling 196 seconds (32.67 per cent), range 2-50 seconds, median 6 seconds; 22 bouts of writing totalling 122 seconds (20.33 per cent), range of 1-20 seconds, median 5 seconds; and eight bouts of hand-clasping totalling 58 seconds (9.67 per cent), range 2-18 seconds, median 5 seconds. From the duration record a criterion occurrence record was derived from sequence B as it might be recorded by both MTS and PIR for all three behaviours of interest. In addition, there was a two-minute demonstration tape made by the research assistant *in situ*, explaining and exemplifying various behaviours including those of interest and the system of audio cues.

Procedure

The 60 observers were divided into six independent groups of 10. Five of each group always used PIR and five MTS for video A and video B. These subgroups were randomly assigned to either PIR or MTS. Each group of 10 recorded at one of three levels of complexity:

1. Reading behaviour
2. Reading and writing behaviours
3. Reading and writing and hand-clasping behaviours.

Thus, for example, the PIR1 group used PIR to code reading behaviour while the MTS2 group used MTS to code both reading and writing behaviours. This design and nomenclature are shown in Table 1.

Sessions lasted approximately 2.25 hours, including task instruction and familiarization with the basic processes of time sampling prior to practice and experimental sequences. Observers were seated in a semi-circle in a university laboratory some 2-3 metres from a large-screen TV/monitor and one of the authors, acting as experimenter, checked with observers that they had an unimpeded view of the screen.

Observers were next given practice and objective feedback on the computer analogue task. After the two-minute demonstration they coded event and non-event using MTS for the first 30 instants and

Table 1. Experimental design and nomenclature for six independent groups assigned according to method type and complexity ($N = 60$)

Behaviours	1	2	3
Group	PIR1 Reading	PIR2 Reading & writing	PIR3 Reading & writing & clasping
Observers	10	10	10
Group	MTS1 Reading	MTS2 Reading & writing	MTS3 Reading & writing & clasping
Observers	10	10	10

changing to PIR for the second 30 intervals. Observers used their records to immediately calculate an agreement percentage with a criterion record displayed to them. The experimenter recorded individual agreement levels but no evaluative feedback was given in accordance with recommendations of O'Leary, Kent & Kanowitz (1975). Typical agreement levels were over 80 per cent for this relatively simple task, indicating observers had acquired at least basic skill in method use.

Observers were then randomly allocated half to MTS and half to PIR observation method subgroups for the remainder of the session and seated so that no two observers using the same method sat next to each other. A complexity level for the group was also assigned (one, two or three behaviours) and the categories of behaviour they were to record were defined for them and presented to them (see Appendix). The two-minute demonstration by the research assistant was then shown and discussion was allowed until all observers assented to an understanding of their task.

Observers were shown the 10-minute practice videotape A of studying behaviour and they recorded the assigned behaviour(s) on standardized printed record sheets. After further clarificatory discussion and a short break, observers were reminded of their category definitions, the 10-minute experimental videotape B of studying behaviour was shown and they recorded the assigned behaviour(s) as before.

There was no verbal behaviour during the observe/record periods and observers were not made aware that any of the three recordings they had made was a source of experimental data. All 60 observers completed records for the experimental material B and a brief questionnaire at the end of the session, allowing them to comment on aspects of the observation.

Only observer records for videotape B were analysed. Occurrence records were arbitrarily paired within method/complexity groups for inter-observer comparisons, the 'pair-score' thus derived being assigned to both individuals. Observers were also compared individually against the criterion record for those behaviours they had coded. Mean kappa scores were then calculated for each group for each behaviour.

Results

Overall, MTS introduced less error into the recordings than PIR across all levels of complexity whether measured between observers or against the criterion.

Table 2 shows the difference in recording accuracy obtained for each group compared by criterion and inter-observer agreement. Inspection suggests MTS is producing consistently acceptable levels of objectively measured accuracy (criterion agreement), while PIR does so chiefly only for writing behaviour. Partial interval recording means are lower than MTS means throughout. Values of mean kappa less than .6 have been asterisked. This value is sometimes taken as a 'low stringency'

Table 2. Group mean criterion and inter-observer agreement for PIR and MTS recording one to three behaviours (kappa values)

	Total number of behaviours recorded			
	1	2	3	Behaviour
Criterion				
PIR	.47*	.60	.52*	Reading
MTS	.75	.81	.75	Reading
	PIR	.76	.73	Writing
	MTS	.85	.83	Writing
		PIR	.55*	Clasping
		MTS	.81	Clasping
Inter-observer				
PIR	.43*	.58*	.41*	Reading
MTS	.70	.80	.80	Reading
	PIR	.69	.65	Writing
	MTS	.80	.83	Writing
		PIR	.50*	Clasping
		MTS	.76	Clasping

* = kappa values < .6.

level of acceptable observer agreement (Gelfand & Hartman, 1975; Suen & Lee, 1985).

The results in Table 2 for inter-observer agreement are generally consistent in direction with those obtained from criterion agreement though they are lower overall. This is almost certainly because criterion comparisons are individual while in inter-observer comparisons an 'inaccurate' partner inevitably reduces their co-observer's 'score' as well as their own—a reminder of the essential difference between the two correspondence measures: many to one, one to one.

Since averaging can hide wide disparities in data values it is worth noting that these group means are an accurate reflection of the shape of the data from which they derive. Only 34 out of 60 PIR 'scores' (56.67 per cent) exceeded the 'lenient' .6 level of kappa, while 58 out of 60 MTS 'scores' (96.67 per cent) did so. Only 15 out of 60 PIR 'scores' would have met the 'stringent' .75 level of kappa (25 per cent), recommended by Landis & Koch (1977) and Suen & Lee (1985), compared with 45 out of 60 MTS 'scores' (75 per cent).

Table 3 shows that between-methods differences are preponderantly statistically significant or highly significant, showing greater accuracy for MTS in recording reading and clasping. However, differences between methods miss statistical significance for most recordings of writing behaviour ($p < .10$), a point to be discussed later.

It had been expected that there would be differences across complexity levels, with increased codings showing decreased accuracy for one or both methods. To examine this possibility, different groups that had coded the same behaviours by the same

Table 3. Comparisons based on mean differences for PIR and MTS groups recording the same behaviour

Groups	Behaviour	I/O t	Sig. <	Crit. t	Sig. <
PIR1 × MTS1	Reading	4.00	.001	4.77	.001
PIR2 × MTS2	Reading	2.91	.01	3.22	.01
PIR2 × MTS2	Writing	1.92	n.s.	1.83	n.s.
PIR3 × MTS3	Reading	10.46	.001	4.38	.001
PIR3 × MTS3	Writing	2.95	.01	1.72	n.s.
PIR3 × MTS3	Clasping	4.39	.001	4.92	.001

Note. Student $t(18)$, two-tailed. I/O = inter-observer; Crit. = criterion. No. in group = 10; N in sample = 60.

method were compared. This expectation was not generally borne out. There were no significant differences at all against criterion. For inter-observer agreement with MTS, the single significant difference (MTS1 × MTS3, $t(18) = 2.90$, $p < .01$, two-tailed) is in the direction of *decreasing* error with *increasing* complexity; in the single PIR case, as expected, there is increased error with increasing complexity (PIR2 × PIR3, $t(18) = 2.35$, $p < .05$, two-tailed). Both these differences occurred with reading. However, they are really too slight to establish any convincing trend.

The experimenters also examined the results of the training computer analogue records in which all observers had practised MTS and PIR. The consistently high inter-observer and criterion agreement that was found here, though we have not allotted further space to tabling it, does support an inference that subsequent differences in the performance of MTS and PIR are unlikely to be the result of the allocation of the sample between methods or a failure on the part of observers to grasp at least the rudiments of observational technique during their, admittedly brief, training period.

We also looked at any significant differences within groups for coding different behaviours (only applicable to groups coding more than one behaviour). Two significant results emerged, both for PIR groups. Against the criterion, PIR2 and PIR3 groups showed a statistically significant difference in coding reading as against writing ($t(9) = 7.20$, $p < .001$; $t(9) = 4.44$, $p < .01$, respectively). The PIR3 group also showed a significant difference in coding writing as against clasping ($t(9) = 3.30$, $p < .01$). These results underscore the relatively greater accuracy achieved by PIR in coding writing behaviour.

Finally, the questionnaire was analysed. It had been thought that self-report might correlate with the more objective measures of accuracy obtained from observer records. However, the responses to questions asked proved generally unable to discriminate between the two method groups and we therefore do not present the questionnaire or its detailed analysis here. However, there was one significant association that emerged, that between method use and claimed efficiency in further method use: MTS observers were significantly more likely to claim the ability to observe longer and more efficiently by their method than were PIR observers

($\chi^2(2) = 5.71, p < .05$). In fact, there were over four times as many observers in the MTS groups who felt able to sustain their method in use for 30 minutes or more than in the PIR groups; MTS does seem to have been the user-preferred method.

Discussion

The results show MTS introducing significantly less error into observer recordings than PIR across all complexity levels. Momentary time sampling indices mostly exceed the 'stringent' level and PIR indices mostly fall below the 'lenient' level of acceptability. Additionally, observers showed some preference for MTS. The size of the sample should increase confidence in the reliability of these results. However, any clear-cut support for the automatic choice of MTS from a single experiment must be qualified.

There are several possible explanations for the generally lower accuracy of PIR here, though this design can exclude inadequacies of the category definitions. Observers may not have understood and applied PIR correctly or diligently. The use of continuous PIR may have imposed a strain. The observers' training may have been inadequate. We consider these possibilities briefly, while bearing in mind that MTS performed under near-identical conditions.

The evidence of good performance on the computer analogue task does not suggest that observers failed to grasp the basic technique of MTS or PIR. Of greater importance is the issue of comparability between methods in this design. Partial interval recording was used continuously rather than non-continuously, yet Hersen & Barlow (1978) have noted that, 'This technique of "non-continuous" observation controls for observers' possible inattention while recording a particular event' (p. 124). Many of PIR's errors here might be accounted for by inattention. Loss of place in the record was also reported only by PIR observers. Momentary time sampling observers had nine seconds available to code their record and PIR observers no time at all. It is thus in the nature of the methods that comparability in their use cannot be precisely achieved. In this connection, though, we note that increasing complexity levels brought no significant decrease in accuracy for PIR (or MTS). This may be explicable as a floor effect, in that coding one to three behaviours did not affect observers' accuracy either way, but it also leads us to question whether PIR's performance is well explained by lack of an observation time in which to record. If this were the case, it would be anticipated that the more behaviours there were to code the more error there would be for PIR and this was not generally found.

With respect to training, the experimental videotape was recorded for only 10 minutes from over two hours' observer attendance, the balance of which was spent in instruction, practice and discussion. How much training is adequate remains an unresolved question. In the present case, the authors would argue that their experimental approach is a sound one and facilitates comparisons with similar research whenever this also limits itself to 'minimal training'. Nevertheless, more research is needed to probe the extent and importance of any relationship between training for time sampling and recording accuracy. It may be, for example, that extensive training would reduce or even reverse differences found here between methods.

It has been suggested previously that the behaviours recorded differed in the level of inference observers needed to apply. In particular, writing (which occupied as many bouts as reading) was recorded with relatively higher accuracy by PIR than either of the other behaviours. This does suggest that level of inference may itself be a variable producing differences within and between methods and does deserve further study, though definitional problems remain over 'level of inference'. It may equally be that the very motility of writing made it more salient to observers than the activity of reading, which lacked any gross motoric movements. Such issues are typical of the difficulty attending even the identification of the important variables in observation and they remain to be resolved. The high accuracy level achieved with hand-clasping by MTS also deserves note, since it appears to contradict earlier findings of one of the authors in relation to MTS's inability to detect infrequent behaviours of low duration (Harrop *et al.*, 1990). Clasping occupied less than 10 per cent (58 seconds) of the observation in eight bouts. To some extent, MTS's superior accuracy here can be explained by the nature of the material, with six out of eight bouts falling across an MTS observe 'moment'. The odds against such a coincidence are about 6 to 1. This illustrates quite well one of the problems inherent in using singular samples derived from videotape, that one can hardly avoid importing idiosyncrasies of the material which may in turn limit generalizability of the results.

Finally, observers' expressed preference for a method, though needing to be treated cautiously, may be worthy of more practical and research consideration than it has so far received.

In summary, the superiority of MTS to PIR as estimated by criterion and inter-observer agreement accuracy has been shown for relatively naïve observers in a single observation of a constructed videotape with its own inherent idiosyncrasies. Admittedly, this was not a 'live' investigation; however, there is no compelling evidence that observers will observe 'live' or 'taped' material with different degrees of fidelity. Meanwhile, controlled conditions enable subtle manipulations and some of the above limitations can be overcome with varied samples of diverse behaviours. It is also important to stress that this study looked only at objective accuracy with respect to the criterion and not at accuracy with respect to overall level of behaviour(s) occurring on the videotape.

This single experiment did not attempt to test either the hypothesized sensitivity of PIR to changing levels of behaviour or the hypothesized increasing inaccuracy of MTS at low levels of behaviour. This may be accomplished using a variety of constructed videotapes in which these variables are systematically altered. The present experiment must be considered a first stage in such a systematic programme and the authors consider that definite conclusions would be premature. There is still some way to go before either MTS or PIR can be unequivocally recommended for use in general contexts on the basis of research with human observers and the differences in recording error that they produce adequately explained. This first stage, however, shows MTS performing better than PIR within the particular circumstances used.

References

Bakeman, R. & Gottman, J. M. (1987). *Observing Interaction: An Introduction to Sequential Analysis.* Cambridge: Cambridge University Press.

Barton, E. J. & Ascione, F. R. (1984). Direct observation. In T. H. Ollendick & M. Hersen (Eds), *Child Behavioural Assessment*, pp. 166–194. New York: Pergamon.

Brulle, A. R. & Repp, A. C. (1984). An investigation of the accuracy of momentary time sampling procedures with time series data. *British Journal of Psychology*, **75**, 481–485.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

Dorsey, B. L., Nelson, R. O. & Hayes, S. C. (1986). The effects of complexity and behavioral frequency on observer accuracy and interobserver agreement. *Behavioral Assessment*, **10**, 349–363.

Gelfand, D. M. & Hartman, D. E. (1975). *Child Behavior Analysis and Therapy.* New York: Pergamon Press.

Harrop, A. & Daniels, M. (1985). Momentary time sampling with time series data: A commentary on the paper by Brulle & Repp. *British Journal of Psychology*, **76**, 533–537.

Harrop, A. & Daniels, M. (1986). Methods of time sampling: A reappraisal of momentary time sampling and partial interval recording. *Journal of Applied Behavior Analysis*, **19**, 73–77.

Harrop, A., Daniels, M. & Foulkes, C. (1990). The use of momentary time sampling and partial interval recording in behavioural research. *Behaviour Psychotherapy*, **18**, 121–127.

Hersen, M. & Barlow, D. H. (1978). *Single Case Experimental Designs.* Oxford: Pergamon.

Hersen, M., Eisler, R. M. & Miller, P. M. (Eds) (1980). *Progress in Behavior Modification*, vol. 10. London: Academic Press.

Kelly, M. B. (1977). A review of the observational data collection and reliability procedures reported in the Journal of Applied Behavior Analysis. *Journal of Applied Behavior Analysis*, **10**, 97–101.

Landis, R. J. & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, **33**, 363–374.

Martin, P. & Bateson, P. (1988). *Measuring Behaviour: An Introductory Guide.* Cambridge: Cambridge University Press.

Murphy, G. (1987). Direct observation as an assessment tool in functional analysis and treatment. In J. Hogg & N. V. Raynes (Eds), *Assessment in Mental Handicap*, pp. 190–238. Beckenham: Croom-Helm.

Murphy, G. & Goodall, E. (1990). Measurement error in direct observations: A comparison of common recording methods. *Behaviour Research & Therapy* **18**, 147–150.

O'Leary, K. D., Kent, R. N. & Kanowitz, J. (1975). Shaping data collection congruent with experimental hypotheses. *Journal of Applied Behaviour Analysis*, **8**, 43–51.

Powell, J., Martindale, B., Kulp, S. & Bauman, R. (1977). Taking a closer look: Time sampling and measurement error. *Journal of Applied Behavior Analysis*, **10**, 325–332.

Repp, A. C., Deitz, E. D., Boles, S. M., Deitz, S. M. & Repp, C. F. (1976). Differences among common methods of calculating inter-observer agreement. *Journal of Applied Behaviour Analysis*, **9**, 109–113.

Suen, H. K. & Lee, P. S. C. (1985). Effects of the use of percentage agreement on behavioural observation reliabilities: A reassessment. *Journal of Psychopathology and Behavioural Assessment*, **7**, 221–234.

Received 25 November 1992; revised version received 18 May 1993

Appendix

Behaviour category definitions

The following category definitions were available to observers before and during the observations:

Reading behaviour

Subject is reading if she is fixating or scanning text. Text is a photostat document on the desk.

Writing behaviour

Subject is writing if she is holding a writing implement and moving it across text or notes in contact with the paper.

Hand-clasping behaviour

Subject is hand-clasping if:
—the fingers of both hands are intertwined, bringing the palms together, or
—the palms are brought together obliquely with the tips of the fingers folded over the backs of the other hand, or
—one hand grasps all or part of the other hand below the wrist.

Behaviours are not necessarily exclusive