

The behavioural approach in schools: a time for caution revisited

Alex Harrop* and Jeremy Swinson
Liverpool John Moores University, Liverpool, UK

This paper takes as its starting point an examination of the current status of some of the concerns that were raised in the mid-1980s about methodological problems faced by educational researchers using the behavioural approach in schools. These concerns included the measurement of agreement between observers, the interpretation of raw data extracted, the potential influences of observers and the inherent properties of research designs. Subsequently, some more wide-ranging concerns are considered, in particular the kinds of behaviour selected for treatment, the lack of analysis of what is involved in teachers' positive responses to pupils' behaviour and the relatively uninvestigated effects of teachers' negative responses. The conclusions are presented as a series of points that are listed, as far as possible, in the order in which they confront the investigator.

Keywords: Behavioural approach; Current concerns; Guidelines

More than 20 years ago an article was written (Harrop, 1985) discussing some of the methodological problems faced by researchers using the behavioural approach, with their associated research designs, in school classrooms. The areas of concern at that time were centred on the way in which data from investigations were recorded and processed. The 'human observer' was likened to a faulty 'cumulative recorder'. More specifically, problems were identified in:

1. measuring agreement between observers;
2. interpreting the raw data extracted by observers;
3. the extent to which observers influenced recorded data by their presence and by their expectations;
4. the inherent properties of available research designs.

*Corresponding author. School of Psychology, Liverpool John Moores University, Henry Cotton Campus, 15–21 Webster Street, Liverpool L3 2ET, UK. Email: A.Harrop@livjm.ac.uk

Some 20 years later, it seems appropriate to revisit these concerns, to comment on any pertinent research and developments that have taken place in British studies and to add some more recent concerns:

5. the kind of behaviour selected for treatment;
6. the lack of analysis of what is involved in teachers' positive responses to pupils' behaviour;
7. the relatively uninvestigated effects of teachers' negative responses.

The aims of the paper are:

- To refocus on the 'faulty human recorder'.
- To present some more wide-ranging concerns stemming from current British research using behavioural methodology.
- To conclude by listing key points raised in, so far as possible, the order in which they confront the investigator.

Measuring agreement between observers

It was noted in the earlier article (Harrop, 1985) that the most common way of checking recordings is to use two independently recording observers and to calculate observer agreement. Because there are a number of ways of calculating such agreement (each of which can produce a different score) (Kelly, 1977), it was suggested that investigators make use of a probability-based formula, like that of Yelton *et al.* (1977).

In subsequent years, it has become appreciated by many investigators that percentage observer agreement can be very high, through chance alone, when the behaviour concerned occurs for a very large proportion, or for a very small proportion, of the observation time. As an illustration of the former, imagine a situation in which a pupil shows 'on-task' behaviour for some 90% of the observed lesson. Two independent observers recording this 'on-task' behaviour are almost certain to produce a high level of observer agreement. Various ways of tackling that difficulty have been devised. One method has been to calculate chance agreement as well as observer agreement, and seek to demonstrate that the latter is higher than the former, or, taking that one step further, calculating observer agreement separately on occurrences and non-occurrences of observed behaviour and comparing each with chance agreement. Other investigators, for example Suen and Lee (1985), have given prominence to Cohen's kappa (Cohen, 1960). Kappa has the in-built advantage of making allowance for chance in its calculation of percentage agreement, so that a positive value of kappa indicates that percentage observer agreement is higher than that which would have been obtained by chance alone. Just how high kappa needs to be is an open question, although Suen and Lee (1985) suggested 0.60 as a lenient criterion and 0.75 as more stringent. Towstopiat (1984), however, noting that the use of kappa is limited to situations when only two observers are involved, suggested the use of multivariate agreement models.

That kappa is not necessarily the ultimate technique for representing agreement between two observers has been noted by Harrop *et al.* (1990b). They presented the results of the observations of two independent observers on one behaviour in a two-by-two contingency table. These results yielded a kappa of 0.67. When the results were distributed in a different way, preserving the number of agreements but redistributing the disagreements, the same kappa figure was obtained. Scrutiny of the two contingency tables, however, showed one table to be more convincing evidence of satisfactory observer agreement than the other. Accordingly, they argued for the presentation of the recordings made by observers in a contingency table for visual inspection, so that agreements and disagreements on occurrences and non-occurrences of behaviour recorded by the two observers could be inspected. Such presentation would allow the calculation of kappa. But when kappa is presented alone, the contingency data are not available to the reader.

Interpreting the raw data extracted by observers

As regards methods for extracting data, some errors associated with time sampling were discussed, quoting the research of Powell *et al.* (1975) and of Repp *et al.* (1976). Both these groups of investigators tended to focus on two commonly used methods of time sampling: partial interval recording (PIR), also known as one-zero recording, and momentary time sampling (MTS), also known as instantaneous recording. In PIR, a recording interval is scored if the behaviour observed occurs for any portion of the interval, whereas in MTS a recording instant is scored only when the observed behaviour occurs at that instant. Repp *et al.* (1976) found PIR to be more accurate than MTS for estimating rate of responding, whilst Powell *et al.* (1975) found MTS to be superior to PIR in estimating duration of response.

There has been a subsequent refinement in our knowledge of the inherent weaknesses of the two methods of time sampling. In investigating the two methods, Harrop and Daniels (1986), using computer simulation, found that estimating duration is more accurate than estimating frequency of behaviour, that MTS is generally more accurate than PIR, but that PIR is the more sensitive of the two for detecting change; although at high rates, PIR tends to underestimate the degree of change. That these results differed from the previous work was a function of the fact that the earlier investigations did not adequately equate the conditions under which PIR and MTS were compared (see also Harrop *et al.*, 1990a).

Should an investigator wish to use one of these methods of time sampling, there are certain guidelines that can be followed. In an investigation in which the aim is to ascertain the duration of various behaviours, then MTS has to be used since PIR, by its very nature, cannot give an accurate estimate. MTS, itself, is at its most accurate when the behaviour to be observed occurs for a large proportion of the observation period, and/or when the bouts of behaviour are long by comparison with the observation intervals. For investigations in which the aim is to change the duration and/or rate of behaviours, then, unless the previously mentioned conditions occur throughout the investigation, PIR is the better method to use since it is more sensitive to

change than MTS. It must, however, be emphasized that partial interval recording, by its very nature, is not an appropriate method for obtaining accurate information on the duration of behaviour occurring.

The extent to which observers influence recorded events by their presence and by their expectations

In the earlier article (Harrop, 1985), the work of Ward (1976) was quoted in terms of the effects of the differing physical appearances of observers on pupil behaviours. It appears that there has been no subsequent work conducted on the extent to which differing appearances can influence pupil behaviour. It also seems likely that the presence of observers might also influence teachers' behaviour, either by inducing feelings of self-consciousness and/or by the teachers seeking to do what they feel observers want to see, as Sommer and Sommer (2002, p. 54) have suggested.

As regards observers' expectations, the work of O'Leary *et al.* (1975) was quoted as showing that bias can be introduced into observers' recordings of video-taped behaviour by the use of contingent feedback designed to mislead. Alternatively, the work of Harrop (1979) was quoted, which described a classroom investigation in which trained observers were instructed that the prime objective was to examine levels of observer agreement. Although the observers were given differing expectations about the course of the investigation, there was no evidence that observer expectations had affected the results, nor of effects of observer presence. Currently, however, it appears that there have been no subsequent investigations to throw further light on these potential influences.

The inherent properties of available research designs

The ABA design (baseline conditions, followed by intervention, followed by a return to baseline conditions) and the multiple baseline design were briefly discussed in 1980, where it was pointed out that the former design may not be acceptable to practitioners because of its withdrawal of an intervention which may well have been judged to have been successful. The latter was said to be more acceptable to practitioners because it does not require withdrawal of the intervention. It does, however, seem to depend upon two conflicting assumptions: first, that any confounding influence will affect more than one variable, and second, that treatment will affect only the specific behaviour for which it is introduced (Kazdin & Kopel, 1975).

Adding a further B (intervention) phase to the ABA design gives an additional opportunity to demonstrate the effects of a treatment and can strengthen the conclusion that the intervention is the agent of change. As an example, Murphy *et al.* (1983) described an investigation in which the effects of an intervention package involving organized games reduced the aggressive behaviour of children in a school playground. When the intervention was withdrawn, the aggressive behaviour reverted to its baseline level, and when the package was restored, the aggressive behaviour decreased once more. The ABAB design, like the ABA design, does

involve withdrawal of the intervention, but at least in this design the final phase is aimed at showing improvement.

A good, recent example of the use of an ABAB design may be seen, in the work of Panagopoulou-Stamatelatu and Merrett (2000), in an investigation in which pupils' written work was improved. What is evident from their results is that whilst large improvements occurred during the two intervention (B) phases, the written work did not decrease to baseline levels in the 'return to baseline condition'. In fact it tended to fall to a position midway between the baseline and the intervention levels, which was sufficient to show that the intervention was having an effect, whilst not returning the pupils' written work to its previously untreated level. That is an important feature of the investigation. It is dangerous to speculate from the results of one investigation, but it seems logical to conclude that the important difference between this investigation and the previously quoted example (Murphy *et al.*, 1983) is that it dealt with an academic behaviour rather than a social behaviour. As such, one might expect the learning involved to be of a more permanent nature. Taking the speculation a step further, it may well be the case that both the ABAB and the ABA design would be reasonably acceptable to both investigators and practitioners when undertaken on academic behaviours, because a return to baseline performance would not be anticipated when an intervention was to be removed.

That the multiple baseline design is not always easy to implement is seen in the work of Harrop and McCann (1984), who began an investigation with the aim of using a multiple baseline design. They were attempting to raise the creative writing performance of a class of pupils in a comprehensive school. The three 'behaviours' selected for intervention were fluency, flexibility and elaboration of writing. These were measured by examining pupils' essays for the number of ideas expressed, the changes in perspective from one idea to the next and the ways in which ideas were spelled out respectively. When the baseline was examined, it was noted that the three 'behaviours' were positively correlated. At that point the multiple baseline design was abandoned since it appeared evident that an intervention for one behaviour would affect the others, therefore invalidating the design. Nevertheless, there have been a number of investigations in which the multiple baseline design has featured—e.g. McNamara *et al.* (1986), Houghton *et al.* (1990), Bain *et al.* (1991) and Nicholls and Houghton (1995).

The kind of behaviour selected for treatment

Since 1980 there have been a number of investigations undertaken using behavioural methodology. Some have been concerned with topics other than the treatment of pupils—e.g. Wheldall *et al.* (1985) and Swinson and Harrop (2005), who were concerned with evaluating training packages for teachers, and Harrop and Swinson (2000), who were concerned with examining teachers' natural rates of approval and disapproval in classrooms. Of those investigations focused on treating pupils, the great majority have concerned themselves with pupil on-task behaviour. These include the work of McNamara *et al.* (1986), Wheldall *et al.* (1989), Houghton *et al.* (1990), Bain *et al.* (1991), Nicholls and Houghton (1995) and Swinson and Harrop

(2001). Conversely, there has been a scarcity of reported investigations concerned with achieving academic aims, the two previously cited investigations, Harrop and McCann (1984) and Panagopoulou-Stamatelatu and Merrett (2000), together with Harrop and McCann (1983) and the recent investigation of Chalk and Bizo (2004), appearing to be the only British examples.

The preponderance of work concerned with increasing pupils' on-task behaviour is likely to have been influenced by teachers' desires for well-ordered classes, and by an implicit assumption that when on-task behaviour is increased, pupil learning will also increase. That is not necessarily the case, as Klein (1979) noted many years ago. Another powerful reason for the focus on on-task behaviour lies in the difficulty of setting up an investigation using behavioural methodology that is aimed at pupil learning. Take, for example, an investigation aimed at increasing pupils' mathematical ability. It would be impossible to measure baseline and intervention levels of mathematical ability over a significant period of time, because pupils would necessarily be working at different kinds of tasks if learning was taking place. The same reasoning would apply to most school subjects. In certain specific circumstances, however, the behavioural model is appropriate, as the three examples cited previously illustrate, but generally the model is difficult to apply to pupil learning. Having said that, the worrying feature of so much research being devoted to on-task behaviour is that such research has not been accompanied by any check on the effects of increasing pupil behaviour on pupil learning. It would surely not be difficult to accompany a piece of behavioural research aimed at improving pupil on-task behaviour with a more traditional research design. It could be that those pupils whose on-task behaviour is to be increased constitute an experimental group whilst another group, not receiving the intervention, constitute a control group.

The lack of analysis of what is involved in teachers' responses to pupils' behaviour

A discussion of 'praise' figures prominently in educational psychology texts (e.g. Pressley & McCormick, 1995; Gage & Berliner, 1998; Woolfolk, 2004). The work of Brophy (1981) is often quoted, complete with what are described as 'critical attributes of effective praise'. Gage and Berliner (1998), for example, list 12 such aspects of 'effective praise'. When we turn to the reports of investigations using behavioural techniques, there is, however, considerable variation between investigations in both terminology and definitions of behaviour. Bain *et al.* (1991), for example, used the term 'encouragement', with a checklist of six items, which included 'teachers' positive verbal comments', and 'building on pupils' ideas'. Wheldall *et al.* (1985) wrote of teachers' positive responses, both verbal and non-verbal, and were careful to point out that these were 'contingent with reference to what had been done'. Other researchers have tended to use the terms approval and disapproval without any standardization of definitions.

Wyatt and Hawkins (1987) gave detailed definitions of 'approval' and 'disapproval' (pp. 33, 34), and went a stage further than usual by differentiating between descriptive

and non-descriptive approval and disapproval. Harrop and Swinson (2000) followed on from that work examining teachers' natural rates of approval and disapproval, once again giving detailed definitions, differentiating between descriptive and non-descriptive approval and disapproval and adding 'disapproval with redirection', defined as the teachers's response following disapproval which describes an approved behaviour (p. 477).

More recently, Chalk and Bizo (2004) examined the effects of what they called 'specific praise' and 'positive praise'. In the specific praise condition, teachers were asked to link praise statements to pupils (individuals and groups), and for social and academic behaviours, to a rule, strategy or effort put in by the pupil. Thus making the praise more informational and specific. In the positive praise condition, 'teachers were asked to praise individuals and groups for social and academic behaviours but were given no instruction on the content of this praise' (p. 340). Two primary school teachers operated under each condition during numeracy lessons. An AB design was employed for each condition. The results showed both conditions resulting in increased pupil on-task behaviour but no significant differences between the two conditions. On a measure of academic self-concept mean pupils' scores increased significantly for the specific praise condition but not for the positive praise condition. On a measure of numeracy enjoyment there was no significant difference. Agreement between observers was calculated using Cohen's kappa, a high level of agreement being demonstrated.

The terminology used in that investigation is somewhat confusing, the notion of positive praise suggesting that there might be something known as 'negative praise'. That conjures up all sorts of strange scenarios. Nevertheless, there is an important distinction here made in terms of the information accompanying praise. Moreover, this investigation is a good example of the use of multiple indices to examine the effects of an intervention.

In more general terms, there are undoubtedly semantic problems involved in using terms like praise, encouragement, approval, etc. It is fairly evident that all these words are not synonymous. Harrop and Swinson (2000), for example, defined approval as: 'Any teacher response which indicated praise or satisfaction with the behaviour of one or more pupils'. That included such comments as 'Excellent', 'Well done', 'Good boy/girl', 'Yes'. It also included the rather less positive statement, 'That's right/correct', and 'the teacher's repeating of a pupil's answer in a positive, neutral but non-querulous manner' (p. 447). The definition includes a variety of kinds of teacher response and it would be impossible to justify them all as praise.

Despite the lack of analysis of what is involved in teachers' positive responses to pupils' behaviour, there is a wealth of research evidence showing the effectiveness of what might best be called teachers' positive responses towards their pupils. At the same time, it appears to be important to begin to be more analytic, following the lead of research such as that of Wyatt and Hawkins (1987), who added descriptive and non-descriptive approval; Harrop and Swinson (2000), who introduced teachers' disapproval with redirection; and Chalk and Bizo (2004), who used the notion of level of information accompanying praise.

The relatively uninvestigated effects of teachers' negative responses

Teachers' negative responses, in definition terms, are the opposite of positive responses. Usually, in behavioural research, there is an explicit or implicit assumption that negative responses should be reduced in frequency in order to effect an improvement in the behaviour being treated. That probably stems from research dating back to the 1970s that showed teachers giving more disapproval than approval in the classroom (e.g. White, 1975; Rutter *et al.*, 1979). Such research painted a rather negative picture of teachers. Later work, however, has found a change, with more approval being given than disapproval (e.g. Merrett & Wheldall, 1987; Harrop & Swinson, 2000). Given that disapproval is still used in classrooms, it is unwise to leave its potential effects unexamined. In a recent investigation, Swinson and Harrop (2001) found evidence which suggested that teachers' disapproval had a curvilinear effect on the behaviour of infant school pupils, but not on those in secondary schools, both too little and too much seeming to reduce pupil on-task behaviour. Data such as those suggest that it is inadvisable to assume that the effects of teachers' negative responses are merely the opposite of the effects of teachers' positive responses on their pupils. Moreover, the research of Houghton *et al.* (1990), which reported increased pupil on-task behaviour when teachers switched from public reprimands to private reprimands, demonstrates the necessity for a more detailed analysis of what is involved in teachers' negative responses to their pupils.

Further comments

There has been so far no reference in this paper to the teaching curriculum and that lack of reference is a reflection of what is to be found in much of the research previously quoted. When a pupil shows behaviour that is not appropriate, it is essential to consider the curriculum first, unless the pupils have not been placed in the appropriate class. There are a number of possibilities: it may be, for example, that the level of work demanded is too high, or it may be that the material is presented in a monotonous, unimaginative manner. Bearing in mind that the teacher has some 30 or so pupils to consider, it may or may not be possible to differentiate the curriculum to accommodate the pupil. If it is not possible, then techniques extraneous to the curriculum will need to be used. If such techniques are used without a careful examination of the curriculum, however, there is the possibility that they are being used to prop up an inappropriate curriculum. That is a most undesirable outcome. Yet it is very difficult to find a research article which reports that attention has been paid to the curriculum before techniques unrelated to the curriculum are put into practice. Our own experience tells us that one reason for not reporting that attention has been paid to the curriculum is that it does not seem to be part of the research design employed. In retrospect, however, not reporting an examination of the curriculum is a dangerous practice since it gives the impression that such an examination did not take place. The behavioural approach can therefore portray itself as a technology to be applied irrespective of the classroom curriculum.

So far, behavioural work has had its focus on verbal approval and disapproval, although some investigators—e.g. Bain *et al.* (1991)—have included the measurement of non-verbal behaviours. These are usually defined as facial expressions, head nods, etc. Yet when much of the teachers' time is spent on marking pupils' work, it is surprising not to find research on the effects of teachers' written comments, ticks, etc. There is surely considerable scope here for influencing pupils' behaviour, the nearest approach having been the use of 'a letter home saying how well the pupil has done', as reported by Harrop and McCann (1983, 1984). The marks and the writing that teachers put on pupils' books can be interpreted as conveying approval or disapproval certainly as easily as can teachers' comments, and since they are not transient, the likelihood is that they can be more accurately interpreted than comments. Moreover, they have the potential for being witnessed by a different population—i.e. the parents rather than the other pupils in the classroom. The work of Burns (1978) demonstrated the popularity of 'a favourable report home', so that it appears likely that homework books containing favourable comments are likely to be shown to parents with a consequent impetus being given to school work.

Conclusion

Despite the risk of oversimplification, the simplest way of drawing together the points raised in the foregoing discussion is to present some of the key considerations in a summarized fashion. In an attempt at clarity, these points are presented, as far as possible, in a linear manner, in the order in which they confront the investigator, although in practice many of these considerations are interdependent.

1. Whatever the aim of the investigation, an examination of the classroom curriculum should be made to determine whether the curriculum can be delivered in a manner which meets the aim.
2. If that examination does not yield a potential solution, the investigator should consider carefully what behavioural measures could be used to meet the teacher's aims. The temptation to seek to increase on-task behaviour before considering whether one or more indices of classroom learning can be measured should be resisted.
3. If it is decided to use 'on-task' behaviour for a whole class (or classes), the investigator should examine the situation to see whether a control class (or classes) could be part of the research design. If that is possible, a decision should be made about what achievement needs to be measured 'before' and 'after' for both classes.
4. Categories for observation should be carefully defined, with due consideration given to how much information can be extracted, particularly in terms of the teacher's behaviour.
5. The most appropriate method of systematic observation to be used needs to be decided upon, bearing in mind the projected results of the investigation.

6. Two independent observers should try out the categories and observer agreement calculated by the most stringent method allowed for by the method of observation—e.g. kappa, if time-sampling is used. If agreement is low, the categories should be redefined and the process repeated until a satisfactory level of observer agreement is obtained.
7. A simple AB (baseline followed by treatment) design could be used if there is a control group. If not, it seems best to start with a multiple baseline design, because unlike the ABA design, it does not have the ethically dubious possibility of removing a successful intervention. A limitation is the fact that the multiple baseline design does require the measurement of at least three behaviours, or individuals or settings.
8. When using a multiple baseline design, it can be ascertained very rapidly whether or not the baseline measures are independent. If they are not independent, the design would have to revert to ABA or better to an ABAB design.
9. In ideal circumstances observers should not know the course of the investigation. Whilst a primary observer records throughout the investigation, a second one is needed from time to time to check the primary observer's recordings. Ideally, the second observer's recording should be random at points unknown to the primary observer, although this is not possible in a normal classroom when transient behaviour is being recorded.

It is hoped that considerations such as those discussed above will go some way to renovating and updating the 'faulty human recorder', whilst at the same time enabling investigators to take into account some features and implications of more recent research.

References

- Bain, A., Houghton, S. & Williams, S. (1991) The effects of a school-wide behaviour management programme on teachers' use of encouragement in the classroom, *Educational Studies*, 17(3), 249–260.
- Brophy, J. (1981) Teacher praise: a functional analysis, *Review of Educational Research*, 51, 5–32.
- Burns, R. B. (1978) The relative effectiveness of various incentives and deterrents as judged by pupils and teachers, *Educational Studies*, 4, 229–243.
- Chalk, K. & Bizo, K. (2004) Specific praise improves on-task behaviour and numeracy enjoyment: A study of year four pupils engaged in the numeracy hour, *Educational Psychology in Practice*, 20(4), 335–351.
- Cohen, J. (1960) A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 20, 37–46.
- Gage, N. L. & Berliner, D. C. (1998) *Educational psychology* (New York, Houghton Mifflin).
- Harrop, A. (1979) A behavioural workshop for classroom problems, *British Journal of In-Service Education*, 1(1), 47–50.
- Harrop, A. (1985) Behaviour modification in schools: a time for caution, *Bulletin of the British Psychological Society*, 33, 158–160.
- Harrop, A. & Daniels, D. (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. (1990a) The use of momentary time-sampling and partial interval recording in behavioural research, *British Journal of Psychology*, 80, 181–189.

- Harrop, A., Foulkes, C. & Daniels, M. (1990b) Observer agreement calculations: the role of primary data in reducing obfuscation, *British Journal of Psychology*, 89, 181–189.
- Harrop, A. & McCann, J. (1983) Behaviour modification and reading attainment in the comprehensive school, *Educational Research*, 25, 191–195.
- Harrop, A. & McCann, J. (1984) Modifying ‘creative writing’ in the classroom, *British Journal of Educational Psychology*, 54, 62–72.
- Harrop, A. & Swinson, J. (2000) Natural rates of approval and disapproval in British infant, junior and secondary classrooms, *British Journal of Educational Psychology*, 70, 473–484.
- Houghton, S., Wheldall, K., Jukes, R. & Sharpe, A. (1990) The effects of limited private reprimands and increased private praise on classroom behaviour in four British secondary school classes, *British Journal of Educational Psychology*, 60, 255–265.
- Kazdin, A. & Kopel, S. (1975) On resolving ambiguities of the multiple baseline design: problems and recommendations, *Behaviour Therapy*, 6, 601–608.
- Kelly, M. (1977) A review of the observational data collection and reliability procedures, *Journal of Applied Behavior Analysis*, 10, 97–101.
- Klein, R. (1979) Modifying academic performance in the grade school classroom, in: M. Hersen, R. Eisler and P. Miller (Eds) *Progress in behavior modification* (London, Academic Press).
- McNamara, E., Evans, M. & Hill, W. (1986) The reduction of disruptive behaviour in two secondary school classes, *British Journal of Educational Psychology*, 50, 209–215.
- Merrett, F. & Wheldall, K. (1987) Natural rates of teacher approval and disapproval in British primary and middle school classrooms, *British Journal of Educational Psychology*, 57, 95–103.
- Murphy, W., Hutchinson, M. & Bailey, C. (1983) Behavioral school psychology goes outdoors: the effects of organized games on playground aggression, *Journal of Applied Behavior Analysis*, 16, 29–36.
- Nicholls, D. & Houghton, S. (1995) The effect of Canter’s assertive discipline program on teacher and student behaviour, *British Journal of Educational Psychology*, 65, 197–210.
- O’Leary, K., Kent, R. & Kanowitz, J. (1975) Shaping data collection congruent with experimental hypotheses, *Journal of Applied Behavior Analysis*, 8, 43–51.
- Panagopoulou-Stamatelatos, A. & Merrett, F. (2000) Promoting independence through behavioural self-management, *British Journal of Educational Psychology*, 70, 603–622.
- Powell, J., Martindale, A. & Kulpe, S. (1975) An evaluation of time-sample measures of behaviour, *Journal of Applied Behavior Analysis*, 8, 463–469.
- Pressley, M. & McCormick, C. (1995) *Advanced educational psychology* (New York, Harper Collins).
- Repp, A., Roberts, D., Slack, D., Repp, C. & Berkler, M. (1976) A comparison of frequency, interval and time-sampling methods of data collection, *Journal of Applied Behavior Analysis*, 9, 501–508.
- Rutter, M., Maughan, B., Mortimore, P. & Ouston, J. (1979) *Fifteen thousand hours* (Shepton Mallet, Open Books).
- Sommer, B. & Sommer, R. (2002) *A practical guide to behavioral research* (New York, Oxford University Press).
- Suen, H. & Lee, P. (1985) Effects of the use of percentage agreement on behavioural observation reliabilities, *Journal of Psychopathology and Behavioral Assessment*, 7, 221–234.
- Swinson, J. & Harrop, A. (2001) The differential effects of teacher approval and disapproval in junior and infant classrooms, *Educational Psychology in Practice*, 17, 157–166.
- Swinson, J. & Harrop, A. (2005) An examination of the effects of a short course aimed at enabling teachers in infant, junior and secondary schools to alter the verbal feedback given to their pupils, *Educational Studies*, 31, 115–129.
- Towstapiat, O. (1984) A review of reliability procedures for measuring observer agreement, *Contemporary Educational Psychology*, 9, 333–352.
- Ward, J. (1976) Modification of deviant classroom behaviour, *Bulletin of the British Psychological Society*, 29, 257–267.

- Wheldall, K., Houghton, S., Merrett, F. & Baddeley, A. (1989) The behavioural approach to teaching secondary aged children (BATSAC): two behavioural evaluations of a training package for secondary school teachers in classroom behaviour management, *Educational Psychology*, 9, 185–197.
- Wheldall, K., Merrett, F. & Borg, M. (1985) The behavioural approach to teaching package (BATPACK): an experimental evaluation, *British Journal of Educational Psychology*, 55, 65–75.
- White, M. (1975) Natural rates of teacher approval and disapproval in the classroom, *Journal of Applied Behavior Analysis*, 8, 367–372.
- Woolfolk, A. (2004) *Educational psychology* (New York, Pearson).
- Wyatt, J. & Hawkins, R. (1987) Rates of teachers' verbal approval and disapproval, *Behavior Modification*, 11, 27–52.
- Yelton, A., Wildman, B. & Erikson, M. (1977) A probability-based formula for calculating interobserver agreement, *Journal of Applied Behavior Analysis*, 10, 127–143.