Clinical Child and Family Psychology Review, Vol. 3, No. 3, 2000

Methodological Issues in the Direct Observation of Parent–Child Interaction: Do Observational Findings Reflect the Natural Behavior of Participants?

Frances Gardner[1]

This review examines evidence for the utility and validity of direct observational techniques for answering particular research and clinical questions. Observational techniques often involve recording behavior in settings that are relatively unnatural for families. However, it is argued that construct validity of observational methods depends partly on whether the findings are representative of participants' typical everyday behavior. Evidence is reviewed concerning whether observational findings are affected by the presence of the observer, and by two factors which have been neglected in the literature, namely the type of task imposed by the observer (e.g., directing parent and child to play rather than observing spontaneous interaction) and the location of the observations (e.g., clinic or laboratory rather than home). The review suggests that the presence of an observer does not necessarily distort the nature of interactions. However, the small number of studies in this area suggest that interactions in structured or artificial settings are not necessarily representative of those normally taking place at home.

KEY WORDS: observational methods; consistency; validity; settings; reactivity; children.

INTRODUCTION

Observational researchers are frequently asked how confident they are that family interactions they observe reflect ''natural'' or typical behavior of the participants. Observational techniques often require researchers or clinicians to assess interaction in settings that are far from usual for the family. For example participants may be brought into a laboratory or clinic, or may be asked to engage in structured activities, tasks or family discussions upon demand by the researcher. Even under relatively more natural conditions in the home, interactions may be affected by the presence of the observer and recording equipment. For answering many research and clinical questions, it is important to know whether observational data gathered under these conditions are representative of family behavior as it occurs in typical everyday settings.

This review draws primarily on evidence from observational studies of parent–child interaction in relation to young children's behavior problems. It focuses on some important issues concerning the validity of observational techniques, examining the extent to which observational findings appear to be affected by: first the presence of the observer; secondly the type of task imposed by the observer (e.g., directing parent and child to play rather than observing spontaneous interaction); and finally the location of the observations (e.g., clinic or laboratory rather than home). Despite the very common use of structured tasks and laboratory locations in observational studies, data bearing on these two latter issues, surprisingly, are rarely reported. Of course, there are many other issues to consider in deciding whether

[1]University of Oxford Department of Social Policy and Social Work, Barnett House, Wellington Square, Oxford OX1 2ER, UK. Tel: (44) 1 865 270320. Fax: (44) 1 865 270324. E-mail: frances.gardner@socres.ox.ac.uk

1096-4037/00/0900-0185$18.00/0

an observational system has adequate reliability and validity for the purpose to which it will be put. Above all, the nature of the research or clinical question being asked will affect all decisions about a measurement instrument, including the criteria used for establishing reliability and validity. Accordingly the paper also discusses when and why it is important for observational findings to reflect typical behavior, which will depend on the purpose and theoretical basis of the observations (Margolin *et al.*, 1998). It is beyond the scope of this paper to review the entire range of practical and psychometric issues involved in designing and carrying out observational studies of parent child interaction, although the reader is referred to many useful sources (Bakeman & Gottman, 1997; Cone, 1982; Dowdney, Mrazek, Quinton, & Rutter, 1984; Hartmann & Wood, 1990; Hops, Davis, & Langoria, 1995; Johnson & Bolstad, 1973; Margolin *et al.*, 1998; Mitchell, 1979; Olson & Foster, 1991; Patterson, 1982; Sackett, 1978; Stoolmiller, Eddy, & Reid, 2000).

BASIC CONSIDERATIONS ABOUT SYSTEMATIC OBSERVATION

Systematic observational techniques for assessing children's social behavior were first developed in the 1930s in nursery school settings (Parten, 1932; Bakeman & Gottman, 1997). There were further advances, particularly within an ethological tradition, in the 1960s and 1970s (Barker, 1963; Blurton Jones, 1972; Hutt & Hutt, 1970). Many of these early pioneers considered important psychometric issues such as the need for adequate sampling of behavior and interobserver reliability (Wasik, 1989). In the 1970s coding systems were developed within a social learning theory framework to answer questions about the nature of parent–child interaction in families where children have conduct problems. Evidence has accumulated for the validity of many of these systems for discriminating between clinic and nonclinic groups of children, for designing interventions and for evaluating their outcome, and for answering important basic questions about parent–child interaction. As a result, many of these early systems are widely used today (Mash & Barkley, 1986; Forehand & McMahon, 1981; Robinson & Eyberg, 1981; Wahler, House, & Stambaugh, 1976). The work of the Patterson and other groups in Oregon has been seminal in this field for several decades, developing new and sophisticated psychometric evaluations of observational instruments (Hops *et al.*, 1995; Johnson & Bolstad, 1973; Patterson 1982; Stoolmiller *et al.*, 2000). More recently they have led the way in combining different measures of behavior to find the most robust and predictive indices of constructs such as child antisocial behavior or parental negative discipline style (Patterson, Reid, & Dishion, 1992).

Advantages and Disadvantages of Observational Methods

For many kinds of research and clinical questions, there are compelling reasons why clinicians and researchers should carry out direct observations of parent–child interaction. Observations can be carried out in the natural setting in which behavior occurs, most commonly in the home. Only very rarely have researchers observed parents and children in 'natural' public places such as the supermarket (Holden, 1983), presumably because observing in these settings can be impractical and because it is difficult to find a setting where all parents in a given sample typically take their children. It is not known whether observing parent–child interaction in the more unusual setting of a laboratory or clinic observation room reflects the interaction that participants would typically have in public (e.g., on buses, in the doctor's waiting room, near busy roads, in friends' houses). Given that public places are a source of great stress to many parents of children with conduct problems, it must limit our understanding of conduct problems that parent–child interaction has not been studied in these natural settings outside the home.

On the face of it, observational techniques provide a window on real behaviors of interest (e.g., shouts, hits, instructions, hugs). These can be defined consistently and reliably by the researcher, rather than by the parent. In contrast, participant reports are based on definitions that are likely to be specific to that individual. They are also more likely to be affected by systematic personal biases related to factors such as the participant's expectations, their negative attributions about the child, or their low mood (Eddy, Dishion, & Stoolmiller, 1998; Fergusson, Lynskey, & Horwood, 1993; Prescott *et al.*, 2000; Richters, 1992). Direct observations are invaluable tools for examining research questions about the mechanisms involved in social interaction, as well as for planning and evaluating intervention. They allow the researcher or clinician to view directly the overt processes within a social interaction, as they take place.

These fine details would be very hard for participants to access through self-report, as much of the behavior seen during encounters of interest (e.g., family conflict) may be automatic and fast moving. Observational techniques can summarize the relevant aspects of these complex interchanges and thus test hypotheses about how behavior unfolds over time, and how it is influenced by social conditions, including the behavioral triggers and reactions of others. Self-report measures on the other hand are invaluable for other purposes such as assessing participants' feelings, thoughts and attitudes and their perceptions of their own behavior and that of others.

As well as providing a microscopic view of how behaviors unfold in time, observational data have also been shown to be useful for providing data based on rates and proportions that represent more stable, trait-like propensities in people such as aggression, avert antisocial behavior, or poor discipline practices. Traits based on a combination of measures, settings and informants, and drawing particularly on direct observations and ratings by trained observers, avoid the problems of systematic bias that affect measures based purely on self-report. These is considerable evidence for the reliability, construct and predictive validity of traits built up in this careful way, such as child antisocial behavior or parental negative discipline style (Capaldi, Chamberlain, & Patterson, 1997; Dishion, Burraston, & Li, in press; Patterson *et al.*, 1992; Patterson & Bank, 1986).

The major drawback of observational techniques is that they are very time-consuming in terms of training observers, carrying out observations, coding interaction and carrying out inter-observer reliability checks (Margolin *et al.*, 1998). The high cost of observational techniques often limits the number of sessions of observation that can be carried out, lead to potential problems of low stability of data (Stoolmiller *et al.*, 2000), which will be discussed in a later section.

The strength of direct observational techniques can be seen in the contribution they have made to understanding how parent–child interaction influences child psychopathology, and how to treat it (for reviews see Campbell, 1995; Dishion, French, & Patterson, 1995; Loeber & Hay, 1994; Patterson *et al.*, 1992; Reid, 1993). A wide range of coding systems have been developed, each designed to investigate a particular set of research questions about parenting and child behavior. There are a number of relatively straightforward coding systems, which sample a manageable number of child and parent behaviors during structured tasks such as joint play and clear up. As well as being useful for testing hypotheses about the relationship between parenting and child behavior problems, behaviors from these coding systems are frequently used as key outcome measures in randomized controlled trials. For example, Robinson and Eyberg's (1981) DPICS and Forehand and McMahon's (1981) system have been used extensively to answer questions such as whether the frequency of child and parent negative behavior observed in the home or clinic decreases following parenting intervention (Webster-Stratton, Kolpacoff, & Hollinsworth, 1989; Webster-Stratton, 1994; 1998).

There are also many complex coding systems sampling large numbers of behavioral categories, which have been used to test questions about subtle processes, including sequential analyses of interaction. For example, the LIFE coding system developed by Hops *et al.* (1995) has been used to examine how maternal behavior toward the child varies with changes in her depressive symptoms (Hops *et al.*, 1987). The Interact system developed by Dumas (1987) has been used to examine patterns of elicitation, response and control in parent–child interaction sequences (Dumas, LaFreniere, & Serketich, 1995). These two systems are expensive in training time but have the advantage of using direct computer entry of coded data, which retains real time and sequence information. Other complex systems have been used to measure the contribution to child outcome of many developmentally important aspects of parenting such as warmth, responsiveness, playfulness, conflict (Dowdney *et al.*, 1984; Dunn & Kendrick, 1982, Gardner, 1987, 1994; Shaw, Keenan, & Vondra, 1994; Zahn-Waxler, Iannotti, Cummings, & Denham, 1990), how parents prevent misbehavior (Gardner, Sonuga-Barke, & Sayal, 1999a, Gardner, Burton, & Wilson, 1999b) and their inconsistency in handling conflict (Gardner, 1989).

Validity of Observational Coding Systems

The studies referred to above provide evidence for the predictive and construct validity of observational systems for answering the kinds of theoretical and clinical questions outlined. Many of these studies are longitudinal, and demonstrate the utility of these observational measures in understanding causal processes, predicting child outcomes and evaluating the outcome of intervention. Direct observations provide unique information about complex and fast-moving

processes, but where this kind of information is not needed, some researchers have asked whether cheaper and easier self-report measures have the same predictive value for answering some questions (Arnold, O'Leary, Wolff, & Acker, 1993).

Although a comprehensive discussion of reliability is outside the scope of this review, before further considering issues of validity, it is important to note that reliability of a measure sets limits on its validity. Adequate levels of interobserver reliability are normally established and reported in observational studies. However, researchers do not always examine test–retest reliability of observational or self-report measures. The particular problem in observational data is that researchers may not have the resources to sample sufficient sessions of observation. If there are also high levels of day-to-day variability in behavior, this can lead to problems such as low occurrence of some behaviors, or findings that do not represent stable estimates of the behaviors of interest. It is difficult to assess the extent of this problem, as many researchers are not able to repeat their observations on two or more occasions close together in time, so are unable to provide test–retest reliability. Where researchers do carry out multiple observation sessions under the same conditions, test-retest reliability data are not always reported. Stoolmiller *et al.* (2000) discuss these problems in depth, and describe a method for correcting statistically for both day-to-day variability and for the problem of nonoccurrence of certain behavior due to short observation periods, known as ''censoring.'' They applied these corrections to data collected as part of the evaluation of a school-based prevention program. The outcome measure in question was antisocial behavior, observed directly in the playground for three 10-min sessions per child, at weekly intervals. Their evidence suggested that effect sizes for antisocial outcomes were substantially underestimated when they used conventional techniques that do not correct for low reliability across sessions and for censoring. The authors suggest that it is important to address these reliability issues, as otherwise there may be a serious danger of underestimating the effects of intervention programs, and indeed underestimating the effects of theoretically important observational variables in general. It is important to note that in these analyses, the assumption is made that low reliability in observational data is due mainly to random error. The same corrections cannot be applied to the more systematic biases that have been found in self-report data due to effects such as mood of the reporter (Eddy *et al.*, 1998; Fergusson *et al.*, 1993; Richters, 1992).

Few studies have directly compared the predictive validity of observational and self-report measures of child problem behavior. One exception is the work of Patterson and Forgatch (1995) who found that observational rather than parent self-report (CBCL) measures of child antisocial behavior were better predictors of key long-term outcomes such as arrest rates and being placed 'in care' or detention. Other studies have examined convergence rather than comparing the predictive validity of two measures. In a clinic based study, Stormshak, Speltz, DeKlyen, and Greenberg (1997) found very few significant correlations between parent-reported child behavior problems on the CBCL, and observational measures of child negative behavior. Robinson and Eyberg (1981) however, using structured tasks in the laboratory, found high correlations between observed and parent reported child deviant behavior. It is interesting that these two studies employing similar samples of young children referred for treatment of conduct problems in university clinics found very different results, and this highlights some of the problems in attempting to draw together disparate findings in this field.

There are also a few studies examining the degree of convergence between observational and self-report measures of parenting constructs, with mixed findings. Not surprisingly, given the arguments in favor of using observational techniques, some studies find modest (Deater-Deckard, Dodge, Bates, & Pettit, 1996) or low levels of convergence. Patterson *et al.* (1992; p. 68), for example, attempted to produce a robust construct of parental discipline by using a range of measures, including direct observation, daily telephone interviews, systematic observer ratings and parent self-report. However, the self-report measure correlated so weakly with other measures of discipline, that it was excluded thereafter. It is not the purpose of this paper to review the growing literature on questionnaire and interview measures of parenting. However, it should be noted that although some studies find significant relationships between observational and self-report measures of the same constructs (Arnold *et al.*, 1993; Kochanska, Kuczynski, & Radke-Yarrow, 1989; Webster-Stratton & Spitzer, 1991; Webster-Stratton, 1998), in many cases correlations are of the order of $r = .3$ or so, suggesting a good deal of unique information may be provided by both sources. Moreover, the Arnold *et al.* (1993) study was very small ($n = 6$), and the Kochanska

et al. (1989) study used observations by mothers of their child, rather than by researchers, which might contribute to the higher correlations they obtained. There is a need for further studies and reviews in this area before we can draw clear conclusions about which aspects of parent and child behavior can be adequately measured in which settings using self-report, and which techniques are most reliable, valid and useful for different purposes.

Importance of Observational Findings Reflecting Natural Behavior

Despite the evidence for the predictive validity of observational methods, it is important to consider under what conditions they produce data that reflect real life interaction. Many researchers have commented on the paucity of studies examining generalizability of observational findings across settings (Hartmann & Wood, 1990; Kazdin, 1982; Pett, Wampold, Vaughan-Cole, & East, 1992). Yet there are important reasons for needing to know whether observational findings, gathered under conditions that are not entirely natural, reflect the typical interactions of participants. Indeed, from many theoretical perspectives, it is essential for the construct validity of observational measures that this is the case. Firstly, observational measures, for example, of child and parent negative behavior, are often used as primary outcome measures in well-controlled trials of interventions for conduct problems (Stoolmiller *et al.*, 2000; Webster-Stratton, 1994; 1998). Observations are used because there is good evidence that they are less subject to systematic biases than parent and teacher report measures (Eddy *et al.*, 1998; Patterson, 1982; Patterson *et al.*, 1992; Patterson & Forgatch, 1995). Since these behaviors are themselves the target of interest, it is important that they reflect behavior in typical problem settings. Clinic observations that indicated behavior had improved from pre- to post-treatment would be of limited value if these changes bore little relation to changes in behavior that were taking place in the home. Secondly, if we want to translate the findings from longitudinal research on parenting into new interventions, as has been most fruitful in this field, then it is essential that the parental behavior observed is closely similar to that which the child normally experiences. Finally, many clinicians use systematic observations of parent–child interaction in the clinic to analyze the functional relations between parent and child behavior, to determine individual targets for intervention, and to assess change. Such behaviors seen in the clinic would not make sense as targets for intervention if they bore little relation to behavior seen at home, or in other settings in which the family typically spend time.

Furthermore, it is important to consider the theoretical assumptions underlying the observational research in question (Cone, 1982; Margolin *et al.*, 1998). Much of the research in the field of parenting and conduct problems has been carried out broadly from a social learning theory perspective, which posits that one of the main influences on child behavior is the overt behavior of other family members. Particularly important are the moment-to-moment patterns and processes (for example negative reinforcement) within family interactions (Eddy, Leve, & Fagot, 2000; Gardner, 1989; Patterson, 1982). From this perspective, construct validity of observational measures depends on the data reflecting these actual interactions, as well as showing good predictions to other, also imperfect, measures. This contrasts with the theory underlying, for example, assessment of personality or attachment relationships. In attachment research, observed behavior (for example, a toddler avoiding his mother when reunited with her during the ''strange situation'') is used to make an inference about a covert, underlying construct, namely, insecure attachment. In the case of the strange situation it could be argued that the important question is not whether the behavior is representative of normal life, but only whether it predicts outcome in accordance with the theory.

ECOLOGICAL VALIDITY OF OBSERVATIONS

The main part of this review will focus on studies that shed light on the question of whether and to what extent observational data reflect typical interaction. It will examine how much we know about the extent to which behavior is affected by the presence of the observer or their equipment, by the imposition of tasks, such as asking the participants to play together or clearup toys, and by the location of observation, which is often a clinic or laboratory rather than home.

Reactivity to the Observer's Presence

Observational researchers are frequently asked to what extent families are affected by the presence of

an observer, and whether their behavior is therefore representative of routine interactions. Clearly it is not possible, without deceiving the participants, to conduct a pure comparison between behavior when observed and that when not observed, but a number of researchers have developed useful approaches for assessing the extent of observer reactivity. The first approach is to compare differing levels of observer intrusiveness, on the assumption that if there is significant reactivity to being observed, then participants' behavior will vary under different conditions.

Bernal *et al.* (1971) and Johnson and Bolstad (1975) compared negative family interactions in the home under two conditions, audiotape alone, and audiotape plus observer present. Neither study found any difference between the two conditions in the rates of parent or child negative behavior, nor in the rate of parental commands. These findings suggest that either observer effects are not very strong, or, that these effects do not vary with the level of intrusion. The latter seems an unlikely explanation for these findings, since both studies employed methods which contrasted two very different levels of obtrusiveness, and both studies analyzed many hours of data. Bernal *et al.* used a particularly unobtrusive recording method, where the tape recorder was hidden and activated automatically over a period of 6 weeks. Bernal *et al.*'s study results should be treated with some caution as the subjects were drawn from only one family, whereas Johnson and Bolstad (1975) observed 12 families. These findings were replicated in a more recent and well-designed study by Jacob, Tennenbaum, Selhamer, and Bargiel (1994), who observed distressed and non-distressed families interacting at mealtimes. They found a range of measures of family interaction to be largely unaffected by varying levels of intrusiveness of the recording procedure.

A second approach to assessing observer reactivity is based on the assumption that reactivity will tend to decline over time, as participants habituate to being observed. Thus if there are reactivity effects, then the first part of an observation session, or the first session of a series, is hypothesized to show stronger reactive effects, and therefore be more atypical, than later ones. A number of studies have tested whether observational data collected on the first session differs from that collected on later sessions, whilst keeping the task demands and setting the same in each session. The drawback of this approach is that, despite the efforts of the researcher to ensure comparability across sessions, behavior may change across sessions for reasons other than changes in reactivity. For example, there may be systematic differences across sessions as a result of changes in the child's or the mother's mood, or as a result of differences in activities or other setting events.

Hughes *et al.* (1979) provided a careful example of this approach. They observed normal 4-year-olds' conversations with their mothers in the home on four occasions. They found that the first session was atypical in that there was more talk to the observer. Despite this, there were no differences between the first and later sessions in the frequency and nature of conversations between mother and child, which were the variables under investigation. Similarly, Johnson and Bolstad (1975), observing in the home, found no evidence of systematic changes within or across sessions in the frequency of negative or positive behaviors. Their study was particularly thorough, comparing across six 45-min sessions of observation, as well as within sessions, which involved comparing the first, second and third 15 min of each session. Kier (1996) found no differences in pre-school siblings' interactions on the first and second of two-hr-long home observations. However, unlike Johnson and Bolstad she did find systematic differences between the first and last 10 min of each session. In the first 10 min, siblings played on their own more, yet stayed closer to their sibling, compared to the last 10 min. This suggests that there may be a possible habituation effect between the beginning and the end of the session. It would have been useful if the author had provided data on whether the first or the last 10 min are more representative of the whole session. Instead, the plausible but untested assumption was made that the first 10 min would be likely to show more reactive effects, and that researchers should consider discarding data from this period.

Many of these studies have investigated observer effects on the behavior of mothers and their young children, mainly between the ages of 3 and 8. However, the extent of observer reactivity is unlikely to be uniform across participants. Factors such as the child's gender or age, the gender of the parent, the familiarity of participants with the observer and the observation setting could all potentially influence reactivity. Harris and Lahey (1982) and Hartmann and Wood (1990) reviewed studies of the factors influencing reactivity, although many of the studies are not very relevant here as they observed the behavior of participants other than parents and children. Hartmann and Wood (1990) suggested that older children and those who are more sensitive or anxious may react more to being observed, but could not find

systematic studies of age or personality effects on children's behavior. Two more recent studies have found effects of gender of the parent on reactivity, with fathers' behavior appearing to show more influence of being observed than mothers' (Lewis *et al.*, 1996; Russell, Russell, & Midwinter, 1992). These studies point to the need to be cautious about assuming that reactivity effects will necessarily be similar in all samples.

Most of the studies reviewed in this section examine reactivity to observers with audio or paper-and-pencil equipment. Increasingly, though, video recording is used for observing complex interactions, because a complete record can be preserved for later coding and reliability checks. These records can also re-analyzed at a later date in the light of new research questions or coding techniques. However, little is known about the reactive effects of video. In laboratory settings, mounting the camera unobtrusively may minimize reactivity, but, as discussed in a later section, observing in the somewhat artificial setting of the laboratory may bring its own problems of poor cross-setting generalizability. There do not appear to be any studies designed specifically to examine reactivity to home video recording, or to compare it with other techniques. Pett *et al.* (1992) used video recording in the home, and found few differences in mother and child controlling and positive behaviors between two sessions. This is suggestive of a lack of habituation effects, but for a stronger test comparisons are needed across a larger number of sessions. Under certain conditions video may be less reactive than live observers, for instance in Pepler and Craig's (1995) pioneering method for studying aggression in the playground using microphones and a camera with telephoto lens.

Although we should always be alert to the possible effects of being observed, these findings suggest that observer reactivity does not pose a substantial threat to the validity of observational data. However, until there have been further studies of the reactive effects of video recording in the home, it is not entirely clear if this optimistic conclusion can be applied to video techniques. Since video techniques are now very widely used, and have considerable advantages, there is a need for researchers to incorporate evaluation of reactivity effects into their studies, for example by comparing video with other methods.

There is also very little evidence to help inform researchers about the effectiveness of different recommendations for reducing reactivity, as these have not generally been subjected to comparative tests of their effects on participants. Nevertheless, many researchers make recommendations about how to reduce reactivity. Kier's (1996) study suggests that not observing for the first 10 min may help toward reducing observer effects, a practice also adopted by Dunn and Kendrick (1980). Some authors recommend not observing at all on the first visit, something that at first sight is not supported by the literature reviewed here. However, it should be noted that some researchers routinely carry out an introductory visit that may serve to habituate participants to the observer's presence, even if formal observations do not take place. Whether habituation effects are minimal even without an introductory nonobservation visit is not clear from the literature, as authors do not always state whether initial discussions with participants were made during a visit, or by other means. As a general rule, it would seem unwise to discard the first session of observations, given that there is little evidence of reactivity effects. On the contrary, it is important to maximize the number of sessions of observation, in order to increase reliability of the data, especially given Stoolmiller *et al.*'s (2000) evidence suggesting that inadequate sampling of behavior may lead to serious underestimation of predicted effects.

Other recommendations include ensuring the same observer visits each time; using unobtrusive equipment; familiarizing families with the recording procedures, which may be particularly important with video (see Dunn & Kendrick, 1982; Gardner, 1987). Some recommend avoiding interaction with participants during recording, drawing on evidence that familiarity with the observer can influence behavior during observations (Hartmann & Wood, 1990; Kazdin, 1982), while others argue that reactivity is reduced if observers act naturally by responding to overtures from family members (Dunn & Kendrick, 1982). Clearly further studies are needed to clarify the effectiveness of these different strategies in reducing reactivity. Harris and Lahey (1982) recommend an empirical approach, that of periodically checking for reactivity effects during data collection by introducing less obtrusive methods (e.g., audio-recording instead of live observer).

The Nature of the Task Imposed by the Observer

Parent–child interaction is frequently assessed during a somewhat artificial task imposed by the observer. The purpose of introducing a task rather than

watching 'natural' interaction may be to elicit efficiently the behaviors of interest and to introduce a degree of comparability between subjects (Hughes & Haynes, 1978). Introducing a constrained task also tends to increase the reliability of the findings, by decreasing the range of possible situational influences on the behavior, although this does not necessarily mean the data are more valid. Typically parent and child are asked to take part in a task that is not uncommon for parents and children of that age to engage in, such as joint play or clear up. Nevertheless, for many families the task will not be very typical for them, or it may not feel very natural for them to be asked to carry out the tasks at the time or in the particular sequence imposed by the researcher. Thus it seems reasonable to assume that for most families imposed tasks will feel more unnatural than being asked to carry on with their normal activities at home, as far as possible as they would if the observer were not there. On the other hand, if parents are asked to 'act natural' in the lab, as in the study of Webster-Stratton (1985), that might well feel more unnatural than being instructed to engage in a clearly defined task. It is likely that perceived unnaturalness will depend to some extent on the context and type of explanations and instructions given to families, and these may be deserving of further empirical study.

Examples of tasks include Forehand and McMahon's (1981) 'Child's Game', where parents are instructed to engage in joint play but to minimize their demands and instead follow the child's lead. In contrast, Roberts and Powers' (1988) 'Compliance Test' encourages negative interactions, by asking parents to get the child to clear up toys. Barkley (1989) in his studies of hyperactive children interacting with their mothers in the laboratory, uses a combination of imposed tasks. These include free play between mother and child, and a set of five structured tasks, such as clear up, copying figures, and etch-a-sketch. Of course the researcher's choice of task will depend on the clinical or research question being asked, but in most cases, an assumption is made that behavior during the task bears some relationship to the participants' style of interacting in more natural settings. This appears to be a largely untested assumption, because although many researchers use defined tasks in the home, few have compared the data with those from natural situations. Others have compared laboratory tasks with naturally occurring situations, such as mealtime, in the home (e.g., Crockenberg & Litman, 1990), so that the effects of imposing activities and potentially different levels of structure are confounded with differences in location.

It is important to note that different tasks would not be expected to elicit comparable rates of behavior. For example we might expect mothers to issue more commands, or children to be less compliant, in a clear-up task than in free play, and this would show up as differences in mean rates of behavior across settings. However, it would be reasonable to expect that parents who are more controlling in one task might tend to be more controlling in other tasks, or that children who are more noncompliant in one task might tend to be more noncompliant in other tasks. This would show up in correlational analyses across settings, which indicate degree of consistency of individual differences across situations, rather than direct comparability in the types or rates of behavior seen. The same considerations apply to comparisons between different locations such as home and clinic, which are discussed in the next section.

The work of Barkley and colleagues (Barkley, 1989; Barkley, Karlsson, Pollard, & Murphy, 1985; Befera & Barkley, 1985) illustrates many of the points made in the last few paragraphs. These authors observed mothers and children with hyperactivity interacting during different task settings in the laboratory. In several of their studies, the authors found that mother and child behavior during ''free play,'' where dyads were instructed to play together 'as they might do at home' with a range of attractive toys, did not discriminate between hyperactive and normal children, nor between children given different dose levels of stimulant drug treatment. However, in more structured, goal-directed tasks, particularly in those that were more difficult for the child, they found significant group differences. Mothers of hyperactive children were more controlling and negative, and the children were more off-task and noncompliant (Befera & Barkley, 1985). Similar effects were found for stimulant medication, which resulted in a reduced frequency of these negative behaviors and an increase in appropriate play behaviors, compared to placebo (Barkley, 1989; Barkley *et al.*, 1985). These studies provide a useful illustration of how some tasks produce more examples of the behaviors of interest and importantly have greater validity for discriminating between groups or examining treatment effects. However, it is not known which of the two types of lab task would be more natural to the participants, or, more importantly for evaluating treatment effects, whether either give us useful information about behavioral difficulties at home. Although there were

presumably different levels of negative parent and child behavior across the two types of task, data are not reported to show if there was consistency across tasks.

Within the clinic setting, Webster-Stratton (1985) found low correlations between structured play-tasks and unstructured activity in the frequency of child noncompliance, and modest correlations for maternal commands. Dunn, Stocker, and Plomin (1990) investigated sibling rather than parent–child interaction, but their study is mentioned here because it is one of the few designed to compare structured and unstructured activities in the home. Using observers' global ratings of different qualities of sibling interaction, they found significant correlations across different structured tasks imposed by the observer (e.g., cooperative and competitive board games; building and creative tasks). However, there were lower correlations between the ratings of the quality of sibling relationship made during the structured tasks and frequency counts of sibling interactions during more natural free play. It may be that cross-situational consistency is higher where global ratings are used, rather than frequency counts, although unfortunately in this study it was not possible to separate out the effects of these methods from the setting effects.

One of the few studies to examine the effects on mother-child interaction of different tasks or settings within the home was carried out by Pett *et al.* (1992). They compared a joint play task with a more natural family mealtime, in a sample of normal preschoolers. They found little consistency between these two home activities in mother and child controlling and positive behaviors. Generally there was better consistency across time, but within activities, than there was across activities.

A recent longitudinal study of parenting style and conduct problems by Gardner *et al.* (1999b; Gardner, Burton, Wilson, & Ward, 2000) was unusual in making direct comparisons of observational data between structured and unstructured settings in the home. Interaction between mothers and pre-schoolers with conduct problems was observed in a series of structured but realistic tasks in the home, as well as during a separate hour-long 'naturalistic' visit, where families were asked to carry on their normal activities, as they might if the observer were not there. To reduce reactivity, families were familiarized with the video recording procedure during an earlier visit, and few restrictions were placed on family activities during the 'naturalistic' visit. The structured tasks were designed to be ones that introduce mild stress, and consisted of situations where mother was busy, including filling in questionnaires, making lunch and talking to the researcher, and those where she had to make demands on the child including clear up, switching off a video and eating lunch.

Using the same measure of frequency of mother-child conflict in all settings, there were significant but modest correlations (of the order of $r = .3$) between the structured tasks and the more naturalistic setting. The correlation between frequency of conflict during the three tasks where mother was busy, and those where she had to persuade the child to do something were higher ($r = .46$). When predictive validity was examined, the pattern was for mother-child conflict in the naturalistic setting to be more highly correlated ($r = .42$–$.47$) with concurrent questionnaire (CBCL) and interview measures of conduct problems than was conflict during the structured tasks ($r = .18$–$.28$). Within the various structured tasks, only the amount of conflict during mealtime was consistently related to conduct problem scores. A possible reason for this is that mealtime was perhaps the most natural of the various imposed tasks. Interestingly, conflict during the clear up task, which is frequently used in observational studies as a measure of child problem behavior, was only weakly related to conduct problems, with coefficients of borderline significance ($r = .25$). This is consistent with the findings of an earlier study of clear up (Gardner *et al.*, 1999a). Conflict during the ''mother busy'' tasks was also unrelated to conduct scores. The study also examined mothers' use of positive discipline strategies, which were analyzed only during the structured tasks. Results were somewhat different than for conflict, as this time it was positive strategies during tasks where mother was busy, and not during the child demand tasks (including mealtime and clear up), that best predicted lower conduct problem scores.

The theoretical implications of these findings are discussed elsewhere (Gardner *et al.*, 1999a; 1999b), but the methodological implications are that even within the home setting, there may be modest consistency between natural and more structured conditions, and that conflict observed in unstructured settings may have more predictive value. The clear up task did not perform particularly well as neither frequency of conflict nor maternal positive strategies in this setting predicted conduct problem scores.

Location of the Observations

As long ago as 1975 Rapoport and Benoit expressed surprise at the paucity of studies examining

whether clinic observations provide information about child behavior that is comparable with home observations. Clearly this is a very important question, since observation in the artificial setting of the clinic or laboratory is so frequently used to plan and evaluate treatment, and to test research hypotheses. Yet there do not appear to be many further studies addressing this issue over the last 25 years. Many studies support the validity of clinic and laboratory observations for discriminating between groups (Forehand & McMahon, 1981; Stormshak *et al.*, 1997) and predicting outcome, but as argued earlier, this is a different kind of question from the main one under consideration here.

There are a small number of studies that have separated out the effects of the nature of the task and its location, and these will be reviewed first. This will be followed by a review of studies that compare settings where both the location and the task demand differ. Although such studies have the drawback that task and location are confounded, they have the potential to answer the important practical question of whether observations during artificial clinic tasks provide data that are representative of usual interactions in the home.

Webster-Stratton (1985) found moderate to high correlations between behavior in the home and clinic, when, in both locations, mother and child were observed in an unstructured situation and asked to 'do whatever they would normally do'. Thus mothers who were more directive in the home also tended to be more directive in the clinic, and the same applied to child compliance. Although there appeared to be consistency of individual differences in maternal behavior across settings, not surprisingly the two settings elicited quite different mean frequencies of the same behaviors. Mothers were more directive and more praising in the clinic, and children were more deviant at home. Kniskern, Robinson, and Mitchell (1983), using the same observation system, set out to compare home and clinic interactions during three defined play tasks. Unfortunately the design of this study does not allow us to examine consistency of individual differences across situations, as different families were assigned to home and clinic observations. Like Webster-Stratton, they found significant differences in mean rates on some variables, namely that children were more noncompliant and mothers used more restraint in the clinic than at home. Mean rates of other behaviors such as commands and praise did not differ between settings.

Belsky (1980) found poor consistency between home and laboratory interactions between mothers and their one-year-olds. Although the task requirements were relatively unstructured in both locations, the instructions given differed widely, in that mothers at home were told to do whatever they normally do, whereas in the laboratory they were told to play freely together.

There are a larger number of studies comparing clinic and home, but using very different kinds of activity in each setting. Zangwill and Kniskern (1982) found high correlations between behavior during clinic play tasks and during dinnertime at home, although overall mean rates of these behaviors were higher in the clinic than home. This was the case for mothers' rates of reinforcement and punishment and children's rates of noncompliance, but not for other deviant behavior. Webster-Stratton's (1985) findings were rather different. She found similar mean rates of child noncompliance in unstructured home observations compared to during clearly defined clinic tasks, but extremely low cross-setting correlations, suggesting that structured clinic observations cannot tell us which children are more difficult at home compared to their peers. It is difficult to reconcile the two studies, which both observed young children with conduct problems using similar coding systems, although Webster-Stratton's sample size was larger ($n = 40$ vs. 15). Forehand and McMahon (1981) also carried out naturalistic observations in the home and used defined tasks in the clinic, for evaluating parent–training interventions. In one of their studies, Peed, Roberts and Forehand (1977) found that posttreatment changes in child compliance in some clinic tasks in a small group of children ($n = 6$) were reflected in changes in home-observed compliance. As expected, home and clinic measures of compliance in the control group ($n = 6$) did not change. However, in a larger sample ($n = 18$), Forehand, Wells and Sturgis (1978) found that child noncompliance in the clinic was a poor predictor of noncompliance in the home.

In the very different setting of a summer camp for boys with ADHD, Hinshaw, Simmel and Heller (1995) compared measures of a covert antisocial behavior, property destruction, observed in a laboratory compared to those drawn from staff records. Property destruction was observed when each boy was left alone in the laboratory during the course of puzzle task. They found substantial correlations ($r = .76$) between the two settings, considerably higher that studies of other behavior in other settings reviewed here. It should be noted that the staff records did not constitute formal direct observations of the be-

havior as it occurred, as this would be difficult when much of the behavior is carried out covertly. The authors comment that this study also differed from others in that the children had become very familiar with the camp and its laboratory after spending several weeks there, which may have reduced the artificiality of the task. It is worth noting that both the familiar context of the imposed tasks, and the use of a different, presumably more judgmental, observational assessment, may have contributed to the high correlation found in this study between the two settings.

CONCLUSIONS

The studies reviewed here suggest that observers need to be cautious about assuming that observations made during structured tasks and in unusual settings such as the lab or clinic will necessarily yield similar findings to those derived from the relatively more natural setting of the home. More reassuringly, however, the problem of observer reactivity does not appear to pose a substantial threat to the validity of observational data. Questions remain about whether this optimistic conclusion can be applied to videorecording in the home, and about the exact effects of different recommendations for reducing reactivity, as these have not generally been subjected to comparative tests of their effects on participants. Even in the absence of a great deal of systematic evidence, researchers' recommendations about how to reduce reactivity can be very useful.

The evidence is much less reassuring about whether observations carried out in an artificial setting, or when the researcher has imposed a task, can tell us about interaction under more natural conditions. There are a limited number of studies, with somewhat conflicting results. Where the results have been slightly more encouraging, sample sizes have often been very small or comparisons have been based on observer global ratings, rather than frequency or duration measures. Possibly observer ratings may be more open to systematic biases such as halo effects which potentially could inflate correlations across settings. Other factors contributing to conflicting findings may include the large numbers of correlations reported in some studies, some of which may be found by chance alone, and the reliability problem discussed in the introduction, of researchers not always sampling sufficient amounts of interaction to give stable estimates of frequencies of behaviors (Hartmann & Wood, 1990; Patterson, 1982; Stoolmiller *et al.*, 2000). Low reliability may underestimate effect sizes with respect to observed behavior in both theoretical and intervention studies. The extent of the problem in the studies reviewed here is not known, as generally authors do not report test-retest reliability. In many studies of parent-child interaction observation sessions are considerably longer than those in the Stoolmiller *et al.* (2000) study, which may increase reliability compared to their school-based data. However, if the problem of low reliability can underestimate theoretically important effect sizes, we may need also to be somewhat cautious about findings which show no relation between measures in different settings, as it is possible that some of these correlations may have been reduced by low reliability across sessions. However, the main point is that these effects have to be measured and taken into account, and the Stoolmiller *et al.* (2000) study makes a considerable methodological advance by attempting to do this.

Clearly further studies are needed, particularly of the effects of imposing tasks on families, as such tasks are frequently used but rarely evaluated. The few studies reviewed here suggest that even within the home setting, there may be poor consistency across different tasks. Gardner *et al.*'s (2000) findings suggest that tasks that have high 'face validity' in that they produce plenty of examples of the behavior of interest, may not necessarily reflect real life behavior, nor have good predictive value. Furthermore, the same behaviors observed in different contexts may show different patterns of prediction to child outcomes.

It may be useful for future studies to compare the predictive validity of observations carried out under different types and degrees of task structure, and to show how the choice of setting varies with the behavior being sampled. A preliminary conclusion from the studies reviewed is that researchers and clinicians need to be cautious about generalizing from one type of task or setting to another, and that it may be preferable where possible to carry out naturalistic home observations. Where home observations are not possible, artificial clinic or laboratory ones may provide useful information under some circumstances. If the aim of observation is to evaluate an intervention for child noncompliance in the home, then it is vital that noncompliance observed in a clinic task tells us something about noncompliance as it occurs in more natural situations. Observations in clinic and laboratory settings appear not to be very good for this purpose. If on the other hand a clinic

task aims rapidly to elicit a behavior of interest (e.g., mothers' critical comments; child noncompliance) in order to intervene clinically, then observing in this setting may be useful. Even if they are not highly representative of real life, artificial observations clearly have predictive value for some purposes, and in any case are likely to be better than no observations, as they may complement or improve on information obtained from parental reports.

In the absence of better evidence about test–retest reliability and about which observation settings are most valid for what purpose, one sensible solution is to carry out observations across multiple settings and tasks and then combine the results, thereby attempting to sample sufficient amounts of interaction across a range of settings. Many researchers have adopted this approach, frequently combining findings across different tasks, or between home and lab (e.g., Forehand & McMahon, 1981; Kochanska & Aksan, 1995; Shaw *et al.*, 1998; Zahn-Waxler *et al.*, 1990). The drawback of this approach is that it is expensive in assessment time and does not directly address the issue of which settings or combinations of settings would be most useful for any given purpose. Further work could help elucidate which tasks or combinations of tasks are most representative of 'typical' interactions and which most predictive of outcome, using the construct building methods of Patterson *et al.* (1992). Particularly for researchers working under high resource constraints, this kind of research would help guide the very difficult choices that are made about how and where to carry out observations of family interaction.

ACKNOWLEDGMENTS

The author is grateful to the Wellcome Trust and the UK Medical Research Council for their support, to Jenny Burton, Charlotte Wilson and Sarah Ward for their contributions to preparing this paper, and to Mike Stoolmiller and three anonymous reviewers for their thoughtful comments on the manuscript.

REFERENCES

Arnold, D., O'Leary, S., Wolff, L., & Acker, M. (1993). The Parenting Scale: A measure of dysfunctional parenting in discipline situations. *Psychological Assessment*, *5*, 137–144.

Bakeman, R., & Gottman, J. (1997). *Observing interaction*: *An introduction to sequential analysis*. (2nd edn.) Cambridge, UK: Cambridge University Press.

Barker, R. (1963). *The stream of behavior*. NY: Appleton–Century–Crofts.

Barkley, R. A. (1989). Hyperactive girls and boys: stimulant drug effects on mother-child interactions. *Journal of Child Psychology and Psychiatry*, *30*, 379–391.

Barkley, R. A., Karlsson, J., Pollard, S., & Murphy, J. (1985). Developmental changes in the mother-child interactions of hyperactive boys: effects of two dose levels of Ritalin. *Journal of Child Psychology and Psychiatry*, *26*, 705–717.

Befera, M. S., & Barkley, R. A. (1985). Hyperactive and normal girls and boys: mother-child interaction, parent psychiatric status and child psychopathology. *Journal of Child Psychology and Psychiatry*, *26*, 439–453.

Belsky, J. (1980). Mother-infant interaction at home and in the laboratory: a comparative study. *Journal of Genetic Psychology*, *137*, 37–47.

Bernal, M. E., Gibson, D. M., William, D. E., & Pesses, D. I. (1971). A device for automatic audio tape recording. *Journal of Applied Behavior Analysis*, *4*, 151–156.

Blurton-Jones (ed.) (1972). *Ethological studies of child behaviour*. Cambridge, UK: Cambridge University Press.

Campbell, S. B. (1995). Behavior problems in preschool children: a review of recent research. *Journal of Child Psychology and Psychiatry*, *36*, 113–149.

Capaldi, D., Chamberlain, P., & Patterson, G. (1997). Ineffective discipline and conduct problems in males: association, late adolescent outcomes, and prevention. *Aggression and violent behavior*, *2*, 343–353.

Cone, J. D. (1982). Validity of direct observation assessment procedures. In D. P. Hartmann (Ed.), *Using Observers to Study Behavior* (pp. 67–79). San Francisco: Jossey Bass.

Crockenberg, S., & Litman, C. (1990). Autonomy as competence in 2-year-olds: Maternal correlates of child defiance, compliance, and self-assertion. *Developmental Psychology*, *26*, 961–971.

Deater-Deckard, K., Dodge, K. A., Bates, J. E., & Pettit, G. S. (1996). Physical discipline among African American and European American mothers: Links to children's externalizing behaviors. *Developmental Psychology*, *32*, 1065–1072.

Dishion, T. J., Burraston, B., & Li, F. (in press). A multimethod and multitrait analysis of family management practices: Convergent and predictive validity. In B. Bukowski & Z. Amsel (Eds.) *Handbook for drug abuse prevention theory*, *science*, *and practice*. New York: Plenum.

Dishion, T. J., French, D. C., & Patterson, G. R. (1995). The development and ecology of antisocial behavior. In D. Cicchetti & D. Cohen (Eds.), *Developmental psychopathology* (Vol. 2, pp. 421–470). New York: Wiley.

Dowdney, L., Mrazek, D., Quinton, D., & Rutter, M. (1984). Observation of parent–child interaction with two-to-three year olds. *Journal of Child Psychology and Psychiatry*, *25*, 379–409.

Dumas, J. E. (1987). INTERACT-A computer-based coding and data management system to assess family interactions. In R. J. Prinz (Ed.), *Advances in Behavioral Assessment of Children and Families*, *3*, 177–203.

Dumas, J. E., LaFreniere, P. J., & Serketich, W. J. (1995). "Balance of power": A transactional analysis of control in mother-child dyads involving socially competent aggressive, and anxious children. *Journal of Abnormal Psychology*, *104*, 104–113.

Dunn, J., Stocker, C., & Plomin, R. (1990). Assessing the relationship between young siblings: a research note. *Journal of Child Psychology and Psychiatry*, *31*, 983–991.

Dunn, J., & Kendrick, C. (1982). *Siblings*: *Love*, *envy and understanding*. London: Grant McIntyre.

Dunn, J., & Kendrick, C. (1980). The arrival of a sibling: changes in patterns of interaction between mother and first-born child. *Journal of Child Psychology and Psychiatry*, *21*, 119–132.

Eddy, J. M., Dishion, T., & Stoolmiller, M. (1998). The analysis of intervention change in children and families: Methodologi-

cal and conceptual issues embedded in intervention studies. *Journal of Abnormal Child Psychology*, *26*, 53–71.

Eddy, J. M., Leve, L., & Fagot, B. (2000). Coercive family processes: A replication and extension of Patterson's Coercion Model. *Aggressive Behavior*, *in press*.

Fergusson, D., Lynskey, M., & Horwood, L. (1993). The effect of maternal depression on maternal ratings of child behavior. *Journal of Abnormal Child Psychology*, *21*, 245–271.

Forehand, & McMahon (1981). *Helping the noncompliant child*. NY: Guilford Press.

Forehand, R., Wells, K. C., & Sturgis, E. T. (1978). Predictors of child noncompliant behavior in the home. *Journal of Consulting and Clinical Psychology*, *46*, 179.

Gardner, F. (1987). Positive interaction between mothers and children with conduct problems: Is there training for harmony as well as fighting? *Journal of Abnormal Child Psychology*, *15*, 283–293.

Gardner, F. (1989). Inconsistent parenting: Is there evidence for a link with children's conduct problems? *Journal of Abnormal Child Psychology*, *17*, 223–233.

Gardner, F. (1994). The quality of joint activity between mothers and their children with behavior problems. *Journal of Child Psychology and Psychiatry*, *35*, 935–948.

Gardner, F., Sonuga-Barke, E., & Sayal, K. (1999a). Parents anticipating misbehaviour: An observational study of strategies parents use to prevent conflict with behavior problem children. *Journal of Child Psychology and Psychiatry*, *40*, 1185–1196.

Gardner, F., Burton, J., Wilson, C. (1999b). Positive parenting style: How does it influence the early development of conduct problems? Paper presented at meeting of International Society for Research in Child and Adolescent Psychopathology, Barcelona, June 1999.

Gardner, F., Burton, J., Wilson, C., & Ward, S. (2000). Parent–child interaction and pre-school conduct problems: how consistent is behavior observed in different settings in the home? Manuscript in preparation.

Harris, F. C., & Lahey, B. B. (1982). Subject reactivity in direct observational assessment: a review and critical analysis. *Clinical Psychology Review*, *2*, 523–538.

Hartmann, D. P., & Wood, D. D. (1990). Observational methods. In: A. S. Bellack, M. Hersen, & A. E. Kazdin, (Eds.), *International handbook of behavior modification and therapy* (pp. 107–138). New York: Plenum.

Hinshaw, S. P., Simmel, C., & Heller, T. L. (1995). Multimethod assessment of covert antisocial behavior in children: laboratory observations, adult ratings, and child self-report. *Psychological Assessment*, *7*, 209–219.

Holden, G. W. (1983). Avoiding conflict: Mothers as tacticians in the supermarket. *Child Development*, *54*, 233–240.

Hops, H., Davis, B., & Langoria, N. (1995). Methodological issues in direct observation: Illustrations with the Living in Familial Environments (LIFE) coding system.. *Journal of Clinical Child Psychology*, *24*, 193–203.

Hops, H., Biglan, A., Sherman, L., Arthur, J., Friedman, L., & Osteen, V. (1987). Home observations of family interactions of depressed women. *Journal of Consulting and Clinical Psychology*, *55*, 341–346.

Hughes, M., Carmichael, H., Pinkerton, G., & Tizard, B. (1979). Recording children's conversations at home and at nursery school: a technique and some methodological considerations. *Journal of Child Psychology and Psychiatry*, *20*, 225–232.

Hughes, H. M., & Haynes, S. N. (1978) Structured laboratory observation in the behavioral assessment of parent-child interactions: a methodological critique. *Behavior Therapy*, *9*, 428–447.

Hutt, S. J., & Hutt, C. (1970). *Direct observation and measurement of behavior*. Springfield IL: Charles C. Thomas.

Jacob, T., Tennenbaum, D., Seilhamer, R. A., Bargiel, K. *et al.* (1994). Reactivity effects during naturalistic observation of distressed and nondistressed families. *Journal of Family Psychology*, *8*, 354–363.

Johnson, S. M., & Bolstad, O. D. (1975). Reactivity to home observation: a comparison of audio recorded behavior with observers present or absent. *Journal of Applied Behavior Analysis*, *8*, 181–185.

Johnson, S. M., & Bolstad, O. D. (1973). Methodological issues in naturalistic observations: Some problems and solutions for field research. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior change*: *Methodology*, *concepts and practice*. (pp. 7–67). Champaign, IL: Research Press.

Kazdin, A. E. (1982). Observer effects: reactivity of direct observation. In D. P. Hartmann (Ed.), *Using observers to study behavior*. (pp. 5–19). San Francisco: Jossey Bass,

Kier, C. (1996). How natural is ''naturalistic'' home observation? Observer reactivity in infant-sibling interaction. *Proceedings of the British Psychological Society*, *4*, 79.

Kniskern, J. R., Robinson, E. A., & Mitchell, S. K. (1983). Mother-child interaction in home and laboratory settings. *Child Study Journal*, *13*, 23–39.

Kochanska, G., & Aksan, N. (1995). Mother-child mutually positive affect, the quality of child compliance to requests and prohibitions, and maternal control as correlates of early internalization. *Child Development*, *66*, 236–254.

Kochanska, G., Kuczynski, L., & Radke-Yarrow, M. (1989). Correspondence between mothers' self-reported and observed child rearing practices. *Child Development*, *60*, 56–63.

Lewis, C., Kier, C., Hyder, C., Prenderville, N., Pullen, J., & Stephens, A. (1996). Observer influences on fathers and mothers: An experimental manipulation of the structure and function of parent–infant conversation. *Early Development and Parenting*, *5*, 57–68.

Loeber, R., & Hay, D. (1994). Developmental approaches to aggression and conduct problems. In: M. L. Rutter & D. Hay (Eds.), *Development through life*: *a handbook for clinicians* (pp. 488–516). Oxford, UK: Blackwell

Margolin, G., Oliver, P. H., Gordis, E. B., O'Hearn, H. G., Medina, A. M., Ghosh, C. M., & Morland, L. (1998). The nuts and bolts of behavioral observation of marital and family interaction. *Clinical Child and Family Psychology Review*, *1*, 195–213.

Mash, E. J., & Barkley, R. (1986). Assessment of family interaction with the Response Class Matrix.. In R. J. Prinz (Ed.), *Advances in Behavioral Assessment of Children and Families*, *2*, 29–67.

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin*, *36*, 376–390.

Olson, R. L., & Foster, S. L. (1991). Assessing childhood noncompliance with parental requests: current status and future directions. In R. J. Prinz (Ed.), *Advances in Behavioral Assessment of Children and Families*, *5*, 139–170.

Parten, M. (1932). Social participation among preschool children. *Journal of Abnormal and Social Psychology*, *27*, 243–269.

Patterson, G. R. (1982). *Coercive family process*. Eugene, OR: Castalia. (Chapter 3: Observations of family process).

Patterson. G. R., & Bank, L. (1986). Bootstrapping your way in the nomological thicket. *Behavioral Assessment*, *8*, 49–73.

Patterson. G. R., & Forgatch, M. S. (1995). Predicting future clinical adjustment from treatment outcome and process variables. *Psychological Assessment*, *7*, 275–285.

Patterson. G. R., Reid, J. B., & Dishion, T. J. (1992). *Antisocial Boys*. Eugene, OR: Castalia.

Peed, S., Roberts, M., & Forehand, R. (1977). Evaluation of the effectiveness of a standardized parent training program in altering the interaction of mothers and their non-compliant children. *Behavior Modification*, *1*, 323–350.

Pepler, D. J., & Craig, W. M. (1995). A peek behind the fence: naturalistic observations of aggressive children with remote

audio-visual recording. *Developmental Psychology*, *31*, 548–553.

Pett, M. A., Wampold, B. E., Vaughan-Cole, B., & East, T. D. (1992). Consistency of behaviors within a naturalistic setting: an examination of the impact of context and repeated observations on mother-child interactions. *Behavioral Assessment*, *14*, 367–385.

Prescott, A., Bank, L., Reid, J., Knutson, J., Burraston, B., & Eddy, J. M. (2000). The veridicality of punitive childhood experiences reported by adolescents and young adults. *Child Abuse & Neglect*, *24*, in press.

Rapoport, J. L., & Benoit, M. (1975). The relation of direct home observations to the clinic evaluation of hyperactive school age boys. *Journal of Child Psychology and Psychiatry*, *16*, 141–147.

Reid, J. (1993). Prevention of conduct disorder before and after school entry: Relating interventions to developmental findings. *Development and Psychopathology*, *5*, 243–262.

Richters, J. (1992). Depressed mothers as informants about their children: A critical review of the evidence for distortion. *Psychological Bulletin*, *112*, 485–499.

Roberts, M. W., & Powers, S. W. (1988). The Compliance Test. *Behavioral Assessment*, *10*, 375–398.

Robinson, E. A., & Eyberg, S. M. (1981). The dyadic parent–child interaction coding system: standardization and validation. *Journal of Consulting and Clinical Psychology*, *49*, 245–250.

Russell, A., Russell, G., & Midwinter, D. (1992). Observer influences on mothers and fathers: self reported influence during a home observation. *Merrill-Palmer Quarterly*, *36*, 263–283.

Sackett, G. (Ed) (1978). *Observing behavior* (Vol. 2): Data collection and analysis methods. Baltimore: University Park Press.

Shaw, D. S., Keenan, K., & Vondra, J. I. (1994b). Developmental precursors of externalizing behavior: ages 1 to 3. *Developmental Psychology*, *30*, 355–364.

Shaw, D. S., Winslow, E. B., Owens, E. B., Vondra, J. I., Cohn, J. F., & Bell, R. Q. (1998). The development of early externalizing problems among children from low-income families: A transformational perspective. *Journal of Abnormal Child Psychology*, *26*, 95–107.

Stoolmiller, M., Eddy, J. M., & Reid, J. B. (2000). Detecting and describing preventive intervention effects in a universal school-based randomized trial targeting delinquent and violent behavior. *Journal of Consulting and Clinical Psychology*, *68*, 296–306.

Stormshak, E. A., Speltz, M. L., DeKlyen, M., & Greenberg, M. T. (1997). Observed family interaction during clinical interviews: a comparison of families containing preschool boys with and without disruptive behavior. *Journal of Abnormal Child Psychology*, *25*, 345–357.

Wahler, R., House, A., & Stambaugh, E (1976). *Ecological assessment of child problem behavior*. NY: Wiley

Wasik, B. (1989). The systematic observation of children: Rediscovery and advances. *Behavioral Assessment*, *11*, 201–217.

Webster-Stratton, C. (1985). Comparisons of behavior transactions between conduct-disordered children and their mothers in the clinic and at home. *Journal of Abnormal Child Psychology*, *13*, 169–184.

Webster-Stratton, C., Kolpacoff, M., & Hollinsworth, T. (1989). The long-term effectiveness and clinical significance of three cost-effective training programs for families with conduct problem children. *Journal of Consulting and Clinical Psychology*, *57*, 550–553.

Webster-Stratton, C., & Spitzer, A. (1991). Development, reliability, and validity of the daily telephone discipline interview. *Behavioral Assessment*, *13*, 221–239.

Webster-Stratton, C. (1994). Advancing videotape parent training: A comparison study. *Journal of Consulting and Clinical Psychology*, *62*, 583–593.

Webster-Stratton, C. (1998). Preventing conduct problems in head start children: strengthening parenting competencies. *Journal of Consulting and Clinical Psychology*, in press.

Zahn-Waxler, C., Iannotti, R. J., Cummings, E. M., & Denham, S. (1990). Antecedents of problem behaviors in children of depressed mothers. *Development and Psychopathology*, *2*, 271–291.

Zangwill, W. M., & Kniskern, J. R. (1982). Comparison of problem families in the clinic and at home. *Behavior Therapy*, *13*, 145–152.