Journal of Child Psychology and Psychiatry 43:4 (2002), pp 507–516

'I'm gonna beat you!' SNAP!: an observational paradigm for assessing young children's disruptive behaviour in competitive play

Claire Hughes,[1] Henna Oksanen,[2] Alan Taylor,[3] Jan Jackson,[3] Lynne Murray,[4] Avshalom Caspi,[3,5] and Terrie E. Moffitt[3,5]

[1]University of Cambridge, UK; [2]University of Helsinki, Finland; [3]University of London, UK; [4]University of Reading, UK; [5]University of Wisconsin, USA

Background: This study focuses on a novel observational paradigm (SNAP) involving a rigged competitive card game (Murray, Woolgar, Cooper, & Hipwell, 2001) designed to expose children to the threat of losing. Recent work suggests that this paradigm is useful for assessing disruptive behaviour in young children (Hughes, Cutting, & Dunn, 2001). **Method:** We report on a large study (involving 800 five-year-olds) that compares observational ratings of disruptive behaviour on the SNAP game with mother and teacher reports of externalising behaviour on the CBCL and TRF (Achenbach, 1991a, 1991b). To ensure independence of data, playmates were randomly assigned to two different sub-samples. The validity of this rigged game for examining individual differences in disruptive behaviour was supported (in both sub-samples) by modest but significant correlations with both mother and teacher ratings of externalising problems, and by significantly elevated SNAP ratings among children rated by mothers and teachers as showing extreme (≥95th %) levels of externalising problems, compared with the remaining majority of children. **Results:** Significant gender differences in disruptive behaviour were found on all three measures: observational SNAP ratings and mother/teacher questionnaire ratings. Factors that may contribute to this gender difference are discussed. **Conclusions:** Our findings emphasise the importance of multi-method, multi-informant measures of disruptive behaviour, and suggest that the rigged card game used in this study is a valuable adjunct to more standard methods of rating disruptive behaviour. **Keywords:** Conduct disorder, disruptive behaviour, gender, methodology.

Research into disruptive behaviour disorders has highlighted the importance of identifying children with behavioural problems at an early age. Early onset of behavioural problems is a strong predictor of a 'life-course persistent' prognosis (Moffitt, 1993). In addition, clinical interventions are likely to be much more successful with younger children, whose problems are not as entrenched or complex as those of older children (Carey, 1997).

How should behavioural problems in young children be assessed? Most studies rely upon questionnaire ratings provided by parents or teachers, since the self-report measures used with older children or adolescents are developmentally inappropriate and direct observations are typically time-consuming, difficult to standardise, and strongly influenced by day-to-day variability in behaviour, such that they show little or no agreement with aggregate rating scales (Epstein, 1983; Hops, Davis, & Longoria, 1995; Jones, Reid, & Patterson, 1975; Stoolmiller, Eddy, & Reid, 2000). However, it is now well recognised that parent and teacher ratings of behavioural problems show only modest agreement with each other (Loeber, Green, Lahey, & Stouthamer-Loeber, 1989), raising interesting questions about context- and informant-effects. For example, findings from several studies suggest that both maternal factors such as depression (Briggs-Gowan, Carter, & Schwab-Stone, 1996; Hay et al., 1999) and transactional effects (Masten & Curtis, 2000; Patterson, Dishion, & Chamberlain, 1993) influence mother ratings of problem behaviours. Similarly, teacher ratings are likely to be influenced by the child's reputation in the school (Realmuto, August, & Hektner, 2000).

Direct observations provide a valuable means of avoiding both informant effects and influences of past transactions on how a current behaviour is interpreted. An important research challenge is therefore the design of observational techniques that can be applied in a standardised format and that are not overly time-consuming to conduct or code. In response to this challenge, in this study we report findings from a novel dyadic play scenario for assessing individual differences in disruptive behaviour. This particular play scenario was chosen because it involves a potential threat (losing a competitive game), and several prominent theoretical accounts of disruptive behaviour focus on heightened perception of/response to threat. For example, Dodge and Frame (1982) found that aggressive children showed a 'hostile attribution bias' when presented with stories involving either neutral or ambiguous actions; this bias is particularly apparent in situations that directly involve the child (Dodge & Somberg, 1987).

Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF, UK and 350 Main Street, Malden, MA 02148, USA

The paradigm used in the present study was a competitive game of SNAP, rigged to expose both players to a mildly stressful experience (a losing streak within the game). From an adult perspective, losing a game may not seem especially frustrating or threatening, but for school-aged children success or failure in competitive play is very important. This point is highlighted by recent findings from two recent experimental investigations using rigged-game designs. In the first of these, both self-report and filmed observations of 8- to 12-year-olds in competitive play with a child actor showed that these children found the rigged situation provoking (Underwood, Hurley, Johanson, & Mosley, 1999). Moreover, as Underwood et al. (1999) discuss, more overt displays of anxiety or frustration might be expected from children in competition with a playmate rather than an unfamiliar peer.

In the second investigation, Murray, Woolgar, Cooper, and Hipwell (2001) developed the SNAP game (used in the present study) to examine depressive cognitions among five-year-old children of depressed mothers. As predicted, children of depressed mothers expressed significantly more hopelessness, pessimism and low self-worth during the losing streak of the game than their typically developing peers. These findings challenge previous studies of young children that failed to show associations between depressive risk status and children's self-reported depressive cognitions (Goodman, Brogan, Lynch, & Fielding, 1993; Nolen-Hoeksema, Girgus, & Seligman, 1986; Rholes, Blackwell, Jordan, & Walters, 1980). This contrast suggests that eliciting spontaneous responses within a salient and ecologically realistic context may provide a sensitive index of problems linked to young children's perception of threat.

The above findings suggest that through its inclusion of a salient and ecologically realistic threat the SNAP game might provide an ideal context for observing individual differences in disruptive behaviour. Hughes, Cutting, and Dunn (2001) explored this possibility in a longitudinal study of 40 'hard to manage' children. Each child was filmed playing the game with a friend at age five, and again at age seven. Compared with a typically developing control group, the 'hard to manage' group showed higher rates of negative behaviour at both time-points. In addition, individual differences in negative behaviour towards peers were stable between ages five and seven, and correlated ($r > .34$, $p < .05$ at both time-points) with earlier individual differences in violent pretend play (Dunn & Hughes, 2001). Together, these findings suggest that the SNAP game is a valid and reliable context in which to observe the social-interaction problems of young disruptive children.

However, a number of questions remain unanswered. For example, are there gender differences in children's disruptive behaviour in the SNAP game? Do ratings of disruptive behaviour in the SNAP game agree with adult reports of externalising problems? How reliable are observational ratings of disruptive behaviour in the SNAP game? The present study (which was part of a wider research programme) addressed these questions in the following ways.

Significant contrasts between girls and boys have been reported in the prevalence, nature and severity of behavioural problems (Zahn-Waxler, 1993). However, these studies typically rely on parental reports, and parents may well evaluate the same behaviours differently for boys and girls (Condry & Condry, 1976; Stevenson-Hinde & Glover, 1996). Direct observational studies typically involve small samples (often composed exclusively of boys), and so rarely have sufficient power to examine effects of gender. Here we report findings from 800 children (200 boy–boy pairs and 200 girl–girl pairs). Our first aim was to capitalise on this large sample to establish whether ratings from this new observational paradigm confirm the gender differences in disruptive behaviour typically reported in questionnaire-based studies.

Second, the study included parent and teacher ratings of externalising behaviour, using widely used, well-validated questionnaire measures: the Child Behaviour Check-list (Achenbach, 1991a) and Teacher Report Form (Achenbach, 1991b). This enabled us to compare our ratings of negative behaviour in the specific context of competitive play with another child with adult ratings of externalising behaviour that reflect how children behave in a wide variety of situations. Assessing the agreement between these different perspectives is important for establishing the *external validity* of the SNAP measure, as well as for deciding how these different sources of information should be combined to provide a robust multi-measure index of behavioural problems in young children.

Third, before rating began, we randomly assigned data from individual children within each pair to either the main sample (S1) or the replication sample (S2) in order to assess the *replicability* of our findings. This was an important goal, since the SNAP game used in this study takes only five minutes to administer, whereas most observational studies of disruptive behaviour involve much longer sampling periods. In addition, since the present study involved 800 children it was necessary to develop a quick and simple coding system; this had the added bonus of making clinical applications of the instrument more feasible, but did raise the question of whether the simplified coding would also be reliable.

Fourth, although the SNAP paradigm has been used successfully in two previous studies (Hughes et al., 2001; Murray et al., 2001), the question of reciprocal influences between social partners has yet to be addressed. Hughes et al. (2001) found that

seven-year-olds who began with a winning streak only to have victory snatched from them showed significantly more negative behaviour than those children who began by losing but later caught up with their friend; however, this effect of winning order was non-significant among five-year-olds. Unfortunately, the composition of child dyads in their study was very variable (some 'hard to manage' children were friends with each other, while others were friends with children in the control group, or with other children). As a result of this between-dyad variation, it was not possible to explore within-dyad reciprocal influences on disruptive behaviour. In contrast, all children in the present study were filmed playing with a same-sex, twin sibling. By bringing together data from S1 and S2 it was therefore possible to examine not only effects of winning order, but also reciprocal influences on disruptive behaviour.

In sum, a rigged card game was used in this study to examine individual differences in five-year-olds' disruptive responses to competitive threat. First, we asked whether boys would show more disruptive behaviour than girls in response to the threat of losing a competitive game. Our second question concerned the external validity of ratings from this rigged situation. This issue was addressed (i) by examining the correlations between ratings of disruptive behaviour on the SNAP game and parents' and teachers' ratings of externalising problems, and (ii) by assessing whether children whose questionnaire scores indicate clinically significant levels of externalising problems showed elevated levels of disruptive behaviour in the SNAP game. Third, to examine the replicability of our results, data from individual children within each pair were randomly assigned to two different samples. In the interest of space, the results from both sets of analyses will be presented together. Fourth, effects of winning order and similarities within each pair were briefly explored.

Method

Participants

Recruitment and participant characteristics. The children in this study were all taking part in a large-scale investigation of environmental influences on early development that in turn was part of a broader programme of research using an epidemiological sample of twins born in England and Wales between 1994 and 1995 (Dale et al., 1998). Data for this study were collected during home-visits to families with same-sex twins, conducted within 60 days of the children's fifth birthday (mean age = 60 months, *SD* = 1.8). The families selected for this study were the first 400 families to receive home-visits. Because all families were seen on a schedule tightly tied to the twins' fifth birthday, the first 400 families seen are not biased by having been more compliant or eager to participate than the remainder of the cohort.[1] Limited resources precluded the inclusion of all study families; one practical aim of the study was to demonstrate the validity of the SNAP paradigm in order to attract funding to extend coding to include the full sample of children.

Participants in this study included 218 monozygotic twin pairs (106 girl–girl pairs and 112 boy–boy pairs) and 182 dizygotic twin pairs (94 girl–girl pairs and 88 boy–boy pairs). To ensure independent data points and assess the replicability of our findings, individual children in each pair were randomly assigned to either 'sample 1' (S1) or 'sample 2' (S2). Since one child from each pair was a participant in either S1 or S2, the two samples were identical in age and zygosity.

Family background. Information about parental education and occupation was collected during an interview with the mother. Note that in the UK, CSE and GCSE exams are taken at age 16 (the former are recognised to be much less challenging than the latter); A levels are taken at age 18 and are usually required for university-entry. The distribution of family ethnicity, maternal educational and head-of-household occupation for the participants is shown separately for girls and boys in Table 1. Independent sample *t*-tests showed no gender differences in ethnicity, parental education or parental occupation.

Intellectual ability. This was assessed during individual testing of each child using the vocabulary and block-design subtests from the Wechsler Pre-school and Primary Scales of Intelligence (WPPSI) (Wechsler, 1990).

Behavioural questionnaire ratings. Mothers completed the CBCL (Achenbach, 1991a) during the home-visit. With parents' consent, teachers were sent the TRF (Achenbach, 1991b) by post, together with a pre-paid response envelope and a pen. In this study we focus on mother and teacher ratings of aggression, delinquency and overall externalising problems.

Direct observations

Materials. The rigged SNAP game was played with two decks of 40 playing cards, each showing a picture of a farm animal. Each card was numbered, so that before the game, the decks could be arranged in the correct order to ensure the rigged pattern of animal matches and mismatches for each child. In addition, an A4-sized metal picture board with a picture of two snakes (each numbered 1–10 from head to tail) was used, so that

[1] Post-hoc tests comparing the 400 families in this study with the remaining 718 families showed no sample difference in ethnic background (χ^2 (9, 1116) = 14.92, *ns*, 94% White in this study vs. 89% in the remaining sample) or maternal age ($t(1116) = .30$, *ns*). However, there was a non-significant trend for mothers of children in this study to have fewer educational qualifications than mothers in the remaining sample ($t(1116) = 1.72$, $p = .09$). The mothers of children in this study also showed significantly lower reading scores on the WRAT (Wide Range Achievement Test) (Wilkinson, 1993) than mothers in the remaining sample ($t(1116) = 2.72$, $p < .01$).

Table 1 Child and family characteristics of study participants

		Boys		Girls	
		Sample 1	Sample 2	Sample 1	Sample 2
Number	*N* (total)	200	200	200	200
Age	Mean in months	60	60	60	60
	(*SD*)	(5)	(5)	(5)	(5)
Full Scale IQ	Mean	95.71	96.37	94.32	94.81
	(*SD*)	(14.90)	(15.90)	(13.36)	(14.78)
Aggression	Mother – mean (/40)	12.84	13.33	11.17	11.45
	(*SD*)	(8.28)	(8.40)	(6.94)	(7.42)
	Teacher – mean (/40)	5.68	6.12	3.39	3.77
	(*SD*)	(7.50)	(8.40)	(5.26)	(5.33)
Delinquency	Mother – mean (/26)	2.84	2.80	2.85	2.24
	(*SD*)	(2.54)	(2.54)	(2.14)	(2.08)
	Teacher – mean (/26)	1.02	1.00	0.60	0.57
	(*SD*)	(1.73)	(1.85)	(1.24)	(1.16)
Externalising	Mother – mean (/66)	15.68	16.13	13.52	13.69
	(*SD*)	(10.31)	(10.32)	(8.47)	(8.82)
	Teacher – mean (/66)	6.64	7.16	3.99	4.35
	(*SD*)	(8.96)	(9.95)	(6.21)	(6.09)
Disruption	SNAP rating – mean (/10)	3.94	3.75	3.16	3.11
		(2.11)	(2.04)	(1.51)	(1.46)
Ethnicity	Caucasian	94%		93.5%	
Mothers' education	Degree or higher	9.5%		9.5%	
	No academic qualifications	33%		36%	
Occupational status	Professional /Managerial	52.5%		34%	
	Unskilled /unemployed	16%		17.5%	
Family structure	Lone parents	22%		19%	

each child could move a magnetic counter along his or her snake when he or she received a matching pair of picture cards (i.e., a 'SNAP').

Procedures. The SNAP game was administered at the end of a three-hour home-visit. A small Sony Camcorder was mounted on a tripod in a corner of the room at the start of the visit, so that by the time the game was played, the presence of the camera was no longer intrusive. The game was used as originally described by Murray and colleagues (2001), with two minor modifications: (i) the researcher dealt cards to each child simultaneously, rather than consecutively, and (ii) magnetic counters rather than sticky stars were used to mark each child's progress in the game. These modifications were introduced to ensure the participation of children with short attention spans, and reduced the duration of the game from approximately ten down to five minutes. Note that the key feature of the SNAP game was its rigged design, which was exactly the same as in the study by Murray et al. (2001). In all but a small minority of cases, the children played the game on the floor in a quiet room with a researcher. Full instructions for the game are given in Appendix 1. Briefly, on each deal of the game the researcher simultaneously presented each child with a pair of picture cards. If a child received a matching pair he or she was allowed to move a magnetic counter one place along a racetrack – the children were told that the first to the end was the winner. The cards were rigged so that each child received a winning streak and a losing streak, so that the children were level-pegging by the end of the game, which ended in a tie, with each child being given a prize of a colouring book. Given the potential ethical concerns of exposing children to the threat of losing, it is worth noting that post-visit feedback from families showed that the children greatly enjoyed the game and hoped for another opportunity to play soon.

Coding. Coding of disruptive behaviour for each individual child was made using both a global scale and an event frequency scale (each applied from videotape across the whole session). Global ratings reflected both minor and major disruptive acts. Minor acts of disruption included surreptitiously moving counters to wrong place, interrupting the researcher, and singing while the researcher was trying to explain the game. Major disruptive acts included: grabbing the board, knocking the board over, throwing counters, trying to snatch the researcher's cards, refusing to relinquish cards, swearing or other forms of verbal aggression, hitting the playmate or storming out of the room. The criteria for each of the ratings on the 5-point global scale were as follows:

1 = child cooperative throughout the game
2 = child not fully cooperative (e.g., responded sluggishly to a request)
3 = child failed to cooperate more than once, or at least one minor disruption
4 = child shows one overt disruptive act, or several minor disruptive acts
5 = child's disruptive behaviour results in premature game termination

The event scale indexed the frequency of rule-violations (e.g., calling 'SNAP' without having a matching pair of cards, trying to move counter more than one square along). Note that these incidences of rule violation (coded on a 3-point scale: 1 = not present, 3 = occurred once or twice, 5 = occurred

more than twice) did not contribute to global ratings of disruptive behaviour. (The 1/3/5 rating was used rather than 0/1/2 to give equal weight to the event scale and the global scale.) A trained researcher (who was unaware of the children's CBCL scores) coded the child on the left-hand side of the screen for all 400 pairs, before returning to code the child on the right-hand side of the screen. All coding was done in real time, so coding time per child equalled administration time (mean = 5.5 minutes, range = 3–15 minutes). A second researcher (who was also unaware of the children's CBCL scores) independently coded 46 randomly selected children to establish reliability; all kappa values exceeded .83, indicating a good level of inter-rater agreement.

Data reduction. Global and event ratings of disruptive behaviour in the SNAP game were significantly correlated with each other ($r(399) = .51, .52$ for samples 1 and 2 respectively, $p < .001$ for both samples). These two scales were therefore summed to create an aggregate disruptive behaviour score. To minimise distorting effects of outliers without loss of data, all outliers were set to 2 *SD* above the group mean using gender-specific means and standard deviations.

Questionnaire measures

Ratings of externalising behaviour were also collected during interviews with mothers and by questionnaires from teachers using the Aggression, Delinquency and Externalising (Aggression + Delinquency) subscales of the Child Behaviour Check List (CBCL) (Achenbach, 1991a) and Teacher Report Form (TRF) (Achenbach, 1991b).

Results

Descriptive statistics

Table 1 shows the mean estimated full-scale IQ (FSIQ) separately for boys and girls in each sample. Paired *t*-tests showed no sample differences in FSIQ, and independent samples *t*-tests showed no gender difference in FSIQ.

Mean mother (CBCL) and teacher (TRF) questionnaire ratings of aggression, delinquency and externalising are presented separately by gender and sample in Table 1. Paired *t*-tests showed no differences in ratings between the two samples, but a significant effect of informant. For all three scales (aggression, delinquency and externalising), mother ratings were higher than teacher ratings, in both S1 and S2 ($t(398) > 13.85$, $p < .001$ for all 6 comparisons). Table 1 also shows the mean ratings of disruptive behaviour on the SNAP game, by gender and sample. Paired *t*-tests showed no sample difference in mean ratings of disruption ($t(398) = 1.08$, *ns*). Independent samples *t*-tests showed no effect of winning order on mean ratings of disruption (for both S1 and S2, $t(398) = 1.10$, *ns*). Gender differences in questionnaire and observational ratings are presented in the next section.

Gender differences in disruptive behaviour

As predicted, both mothers and teachers rated boys higher than girls for aggression, delinquency and overall externalising behaviour (for mothers, in S1 $t(398) = 2.16, 2.08, 2.29$, respectively; in S2 $t(398) = 2.37, 2.40, 2.54$ respectively, $p < .05$ for all three scales in each sample; for teachers, in S1 $t(398) = 3.52, 2.80, 3.51$, respectively, in S2, $t(398) = 3.32, 2.73, 3.35$ respectively, $p < .01$ for all three scales in each sample). SNAP ratings of disruptive behaviour were also higher for boys than for girls ($t(398) = 4.29, 3.58$ for S1 and S2 respectively, $p < .01$ for each sample). That is, regardless of the informant, rating method and context, boys appeared more disruptive than girls. The effect size for this gender difference ranged from .20 (mothers' ratings of externalising in S1) to .37 (SNAP ratings of disruption in S1). From Cohen (1988), $r = d/\sqrt{(d^2 + 4)}$, so corresponding r^2 values ranged from .10 to .18. That is, across all rating methods, gender accounted for 10–18% of the variance in disruptive behaviour.

Validity and replicability of SNAP ratings

The external validity of disruptive behaviour ratings on the SNAP game was assessed in relation to mother (CBCL) and teacher (TRF) questionnaire ratings of aggression, delinquency and overall externalising behaviour. First we examined the correlations between these different rating scales. Was there significant agreement between observational ratings on the SNAP game and adult-reports on these well-validated questionnaires? Next, we examined whether the sub-group of children with extreme CBCL and TRF ratings of externalising behaviour were significantly more disruptive on the SNAP game than the remaining majority of children. That is, did ratings from the SNAP game support the questionnaire-based clinical cut-off? Finally, by using co-twins as a replication sample we were also able to investigate the replicability of SNAP ratings.

Agreement with questionnaire ratings. Correlations between SNAP ratings of disruptive behaviour and mother/teacher ratings of aggression, delinquency and overall externalising behaviour are presented in Table 2. Overall, the results showed significant agreement for 10/12 correlations. However, the magnitude of these correlations was only modest ($r(399)$ ranged from .16 to .21 for teacher ratings, and from .09 to .16 for mother ratings). For each sub-sample, mother-rated delinquency was the only externalising scale that did not correlate with SNAP ratings. Mean values on this scale were very low, so the lack of correlation may simply reflect the limited variance on this scale. When effects of FSIQ were controlled, partial correlations between SNAP ratings of disruptive behaviour and teacher ratings of aggression, delinquency and externalising

Table 2 Correlates of SNAP ratings of disruptive behaviour, by sample, gender and informant

Measure	Sample	All (n = 400)	Boys (n = 200)	Girls (n = 200)
WISC IQ	1	−.13*	−.15*	−.13
(Block & Vocabulary)	2	−.12*	−.20**	−.02
Parental occupational status	1	.07	.04	.10
(highest value)	2	.09	.11	.05
CBCL Aggression	1	.11*	.13	.01
	2	.16**	.28**	−.06
CBCL Delinquency	1	.09	.20**	.04
	2	.11	.21**	−.11
CBCL Externalising total	1	.11*	.13	.02
	2	.16**	.28**	−.07
TRF Aggression	1	.17**	.19**	.05
	2	.20**	.23**	.10
TRF Delinquency	1	.16**	.16**	.07
	2	.16**	.18*	.07
TRF Externalising total	1	.17**	.19**	.05
	2	.21**	.23**	.09

* p < .05, ** p < .01.
CBCL = mother ratings; TRF = teacher ratings.

remained significant (at the p < .05 level or higher) in both S1 and S2; partial correlations between SNAP ratings of disruptive behaviour and mother ratings of aggression and externalising remained significant in S2 but fell just below significance (p = .09) in S1.

Table 2 also shows the correlations between SNAP ratings of disruptive behaviour and questionnaire ratings of aggression, delinquency and externalising behaviour separately for boys and girls. These correlations appeared stronger for boys than for girls (for whom all correlations were non-significant). However, one-tailed tests using Fisher's z-transforms showed that in S1 there were no significant gender differences in the strength of these six correlations for girls and boys; while in S2, there was a significant gender difference in the strength of the correlations between SNAP ratings and all three mother (but not teacher) questionnaire ratings (z > 1.63, 1-tailed). Since questionnaire ratings are based on typical everyday behaviour, these results indicate that the SNAP game may have somewhat greater ecological validity for boys than for girls. We will return to this issue in the discussion section.

Do 'high externalisers' show high disruption on the SNAP game? Our second approach to assessing the validity of the SNAP game was to examine whether children rated on the questionnaires as showing extreme levels of externalising problems (≥95th%) obtained significantly higher SNAP ratings of disruption than did the remaining majority of children. The children in the extreme 'high externaliser' group were all rated ≥95th% for externalising problems (cut-off score = 17 for boys, 15 for girls) by *both* mothers and teachers. Table 3 shows the mean z-scores for disruption for extreme and normal groups, for each sub-sample. These means were compared using t-tests that did not assume equal variance (using natural logarithms of the z-scores to reduce skewness). A significant difference in mean SNAP ratings of disruption was found for both S1 and S2 (t(398) = 2.09, 3.59, p < .05, p < .001 respectively).

Agreement in findings from S1 and S2. Note that the data for S1 and S2 (shown in Tables 1, 2 and 3) were remarkably similar. Admittedly, S2 cannot be considered as a full replication sample for S1, since the children in each sample were related to each other. Nevertheless, the data from the two groups showed extremely similar means and distributions. In support of the *replicability* of SNAP ratings of disruptive behaviour it is worth noting that 30/32 analyses showed very similar results in both samples. (The exceptions were the correlations between SNAP ratings and CBCL aggression and externalising that were significant for boys in S2 but not S1.)

Are our findings limited by reciprocal influences between playmates?

Finally, the data from S1 and S2 were combined to examine the within-pair correlation in SNAP ratings of disruptive behaviour. This correlation (r(398) = .24, p < .001), although significant, was lower than

Table 3 Mean (gender-specific) SNAP Z scores for children with extreme vs. normal ratings of externalising problems

Sample	Measure	Extreme (≥95th %)	Normal (<95th %)	Group diff. T
1	X	.47	−.03	2.09*
	SD	(1.34)	(.97)	
	N	24	376	
2	X	.81	−.05	3.50**
	SD	(.84)	(.97)	
	N	22	373	

** p < .01; * p < .05.

within-pair correlations in ratings of externalising problems made by mothers or teachers ($r(398) = .53$, .63 respectively, $p < .001$ for both).

Recall also that there was no effect of winning order on SNAP ratings of disruption. That is, within each pair the SNAP paradigm is equally sensitive to disruptive behaviour in both children (who necessarily experienced different winning orders).[2] Taken together, these findings suggest that reciprocal influences on the SNAP game are modest, and do not limit the validity of individual-based ratings from this observational paradigm.

Discussion

The focus of this study was a new technique for observing disruptive behaviour, involving a rigged competitive game of SNAP. Previous work (Hughes et al., 2001) with the SNAP game revealed group differences in disruptive behaviour between young 'hard to manage' children (all $\geqslant 90^{th}$ % for symptoms of attention-deficit hyperactivity disorder – ADHD) and their typically developing peers (all $< 50^{th}$ % for ADHD symptoms). These differences were stable from age four to age seven, supporting the reliability of this observational paradigm. The present study extends this preliminary finding in several ways.

First, the participants in this study were filmed playing with a twin sibling at home, rather than with a friend at school. Thus the observations differed in both situational context and in social partners, providing a useful test of the *generalisability* of findings from the SNAP game. Second, this study involved a much larger and more representative sample; in addition, co-twins were allocated to two separate sub-samples, to provide an internal replication study. The data reported in this study therefore also support the *replicability* of findings from the SNAP game. Third, this study included concurrent parent and teacher ratings on standardised questionnaire measures of externalising problems. These mother and teacher ratings showed modest but significant correlations with SNAP ratings of disruptive behaviour (even when effects of IQ were controlled), supporting the *external validity* of the SNAP game. Fourth, this study showed significant contrasts in SNAP ratings of disruptive behaviour between children with extreme ($\geq 95^{th}$ %) scores for externalising problems and the remaining majority of children. Compared with Hughes et al.'s (2001) study of 'hard to manage' children, the group comparisons in this study were both qualitatively different and quantitatively more conservative. The positive findings from the study therefore support the *sensitivity* of disruptive behaviour ratings from the SNAP game.

At this point it is worth considering why, although significant, correlations between direct observational ratings and adult questionnaire reports were relatively modest. Since previous studies indicate that both assessment methods are reliable, the most obvious explanation for this modest correlation is that the SNAP paradigm assesses disruptive behaviour within a specific context (competitive play with another child) at a specific time, whereas questionnaires such as the CBCL provide global ratings of everyday behaviour across a range of contexts and a time-frame of months rather than minutes. Although, as discussed in the introduction, the SNAP game provides a window onto children's behaviour in a highly salient context, numerous previous studies have highlighted both the context-specificity (Gardner, 2000; Hops et al., 1995; Stoolmiller et al., 2000) and day-to-day variability (Jones et al., 1975; Stoolmiller et al., 2000) of disruptive behaviour. As a result, only a modest agreement with global questionnaire ratings can be expected, since each method assesses different facets of an underlying behavioural disposition (Epstein, 1983). In view of this long history of observational measures of disruptive behaviour failing validity checks, we believe that the significant correlations between SNAP ratings and adult questionnaire scores (especially teacher scores) are very encouraging.

A second extension to previous research with the SNAP paradigm comes from the fact that this study's large sample size enabled effects of gender to be explored. Significant gender differences were obtained from all three informants: parent and teacher questionnaire ratings of externalising behaviours, and researcher's direct observational ratings of disruptive behaviour in the SNAP game. Power analysis showed that gender accounted for 10–18% of the variance in disruptive behaviour – this is somewhat larger than the effect sizes of gender that are typically reported in studies that adopt a traditional individual differences framework. This finding confirms Maccoby's (1998) view that gender differences become more striking when viewed through the lens of *child–child interactions*.

Note that each child in this study was filmed playing with a same-age, same-sex sibling. Evidence from several studies of a marked gender divergence in play-styles is therefore relevant. Compared with girls, boys' play is more competitive, and more often on the edge of aggression (Charlesworth & Dzur, 1987; Flannery & Watson, 1993; Smith & Boulton, 1990), while boys' speech is more power assertive (Leaper, 1991; Miller, Danaher, & Forbes, 1986). Thus the fact that only same-sex pairs were involved in this study may have heightened the observed gender difference in disruptive behaviour. Further work with mixed-sex pairs playing the SNAP game is needed to explore this possibility properly.

[2] This conclusion may not hold true for all age-groups, since in a previous study Hughes et al. (2001) found no effect of winning order in five-year-olds, but a significant effect when the children played the game again at age seven.

Gender differences have also been reported in children's relationships with adults. As toddlers, boys have been shown to be more likely than girls to ignore mothers' initial low-key remonstrances, so that mothers become more likely to resort to more forceful methods of control (Minton, Kagan, & Levine, 1971). Similar findings have been obtained in studies of adult–child interactions in nurseries (Fagot, Hagan, Leinbach, & Kronsberg, 1985) and in the first year of school (Grant, 1985). An adult researcher administered the SNAP game, and many of the examples of 'minor disruption' involved acts that were directed towards this adult (e.g., interrupting researcher or trying to snatch researcher's cards). To our knowledge, the extent to which contrasting attitudes and/or responsiveness to adults can explain gender differences in disruptive behaviour has not been investigated, but is an interesting avenue for future research in this field.

In addition, the competitive threat within the SNAP game (and the possibility of rule-breaking) may have been especially arousing for boys. Thorne and Luria (1986, cited in Maccoby (1998)) found that, unlike girls, boys show great excitement in rule-breaking, while Eisenberg, Fabes, Nyman, Bernzweig, and Pinuelas (1994) found that boys became more aroused than girls when watching a film that contained an element of threat. Boys' positive enjoyment of this kind of arousal has been posited as one factor contributing to early gender segregation in peer interactions (Maccoby, 1998), and so it may also be that the SNAP game was not only more exciting for boys, but also more representative of boys' everyday social interactions. This hypothesis is supported by the stronger agreement of SNAP ratings with adult ratings for boys than for girls (although this gender difference was only statistically significant for mother ratings in S2, and may simply reflect the greater variance in boys' ratings). In support of this view, careful naturalistic observations suggest that for *both* boys and girls, individual differences in response to arousing or stressful situations are valuable, since 'even by preschool age individual differences in children styles of regulating themselves… are related to their everyday anger-related behaviours' (Eisenberg et al., 1994, p. 126).

Finally, it should also be noted that whilst the present coding system focused on male-relevant behaviours, the SNAP game is very versatile, and could equally well be applied to assess female-relevant behaviours. This point is clearly demonstrated by Murray et al.'s (2001) original use of the SNAP game to study childhood vulnerability to depression.

Taken together, the data presented in this study confirm that the SNAP game is a potentially valuable supplement to more standard questionnaire methods of assessing disruptive behaviour, as it has been shown to be for assessing depressive cognitions in young children (Murray et al., 2001). In particular, our data suggest that the adapted SNAP game is not only simple to administer and code, but also yields direct ratings of disruptive behaviour that are reliable, show significant agreement with adult questionnaire reports, and are sensitive to both gender differences and the contrast between children showing clinically significant vs. normal levels of disruptive behaviour.

Although the findings from this study are encouraging for other researchers investigating disruptive behaviour in young children, more work is needed to establish fully the reliability of the SNAP paradigm. Given the rigged nature of the game, assessing short-term test-retest reliability is likely to be problematic. However, in our future research we hope to assess the long-term predictive validity of SNAP ratings. The current data all derive from the age-five phase of this research programme; the age-seven phase is now well under way; this will enable us to analyse the SNAP paradigm's effectiveness in predicting ratings of disruptive behaviour across a two-year interval. In particular, we hope to ascertain the extent to which age-five disruptive-behaviour ratings from the SNAP game and from the CBCL/TRF predict unique or overlapping variance in disruptive behaviour at age seven. More long-term research plans with this sample include an evaluation of how well the SNAP game predicts other outcome measures such as psychiatric referral and juvenile delinquency (cf., Patterson & Forgatch, 1995). Other possible future directions require a new sample; these include comparisons of mixed-sex vs. same-sex pairs; validation of the SNAP paradigm against more traditional approaches involving longer time frames for observing disruptive behaviour; and assessing the SNAP game as a tool for assessing improvements following intervention.

Acknowledgements

The Medical Research Council funded this study. We would like to extend warm thanks to all parents, children and teachers who participated in this study.

Appendix 1
Instructions for the (revised) SNAP game

The game is conducted with two children (A and B) sitting side by side and the researcher opposite (either on the floor, or at a table), and is introduced as follows.

'Have you ever played a game called SNAP! before? Well, we're going to play a game a bit like SNAP, using these special snakes. This one is for you (A) and this one is for you (B). I'm going to give you each two cards with pictures of farm animals.' To A: *'If your cards have the same animal on them, I want you to say 'SNAP!!'* To B: *'And if your cards are the same, you can say 'SNAP!!' too!'* To both: *'When you get a*

snap, you can move your magnet ONE place along the snake.' (Give a magnet to each child.) *'NO CHEATING!! Move your magnet one place along each time you get a SNAP. The winner is the first to get to Number 10, and will get a special prize. OK, do you understand what to do? Let's have a practice first without the magnets. First I'll give two to (B). That's right, they're a SNAP. Now, in the proper game, you'd move your magnet one place, wouldn't you? Now it's (A)'s turn. Oh, so they're not the same, so you wouldn't say SNAP, would you?'*

The test-phase begins when both children understand the rules of the game. On each deal of the test phase, the cards are dealt simultaneously to Child A and Child B.[3] The researcher should encourage the children to see the game as a race (e.g., by occasionally asking *'Who's going to win?'*), but should also look out for cheating (e.g., surreptitious movements of the children's counter). The cards were prearranged so that Child A won trials 1, 2, 3, 4, 6, 7, 8, 16 and 18, Child B won the practice trial and trials 5, 9, 10, 11, 12, 13, 14, 15 and 17, and both children got a SNAP on the 19th (final) deal, and so emerged as joint winners.

Correspondence to

Claire Hughes, Centre for Family Research, University of Cambridge, Free School Lane, Cambridge CB2 3RF, UK; Tel: +44 (0) 1223 334512; Fax: +44 (0) 1223 330574; Email: ch288@cam.ac.uk

References

Achenbach, T. (1991a). *Manual for the Child Behaviour Checklist/4-18 and 1991 profile*. Burlington VT: University of Vermont Department of Psychiatry.

Achenbach, T. (1991b). *Manual for the teacher's report form and 1991 profile.* Burlington, VT: University of Vermont Department of Psychiatry.

Briggs-Gowan, M., Carter, A., & Schwab-Stone, M. (1996). Discrepancies among mother, child, and teacher reports: Examining the contributions of maternal depression and anxiety. *Journal of Abnormal Child Psychology*, *24*, 749–765.

Carey, K. (1997). Preschool interventions. In A. Goldstein & J. Conoley (Eds.), *School violence intervention: A practical handbook* (pp. 93–106). New York, NY: The Guilford Press.

Charlesworth, W., & Dzur, C. (1987). Gender comparisons of preschoolers' behaviour and resource utilization in group problem-solving. *Child Development*, *58*, 191–200.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale, NJ: Erlbaum.

Condry, J., & Condry, S. (1976). Sex differences: A study of the eye of the beholder. *Child Development*, *47*, 812–819.

Dale, P., Simonoff, E., Bishop, D., Eley, T., Oliver, B., Price, T., Purcell, S., Stevenson, J., & Plomin, R. (1998). Genetic influence on language delay in two-year-old children. *Nature Neuroscience*, *1*, 324–328.

Dodge, K., & Frame, C. (1982). Social cognitive biases and deficits in aggressive boys. *Child Development*, *53*, 620–635.

Dodge, K., & Somberg, D. (1987). Hostile attributional biases among aggressive boys are exacerbated under conditions of threat to the self. *Child Development*, *58*, 213–224.

Dunn, J., & Hughes, C. (2001). 'I got some swords and you're dead!': Fantasy and friendship in young 'hard to manage' children. *Child Development*, *72*, 491–505.

Eisenberg, N., Fabes, R., Nyman, M., Bernzweig, J., & Pinuelas, A. (1994). The relations of emotionality and regulation to children's anger-related reactions. *Child Development*, *65*, 109–128.

Epstein, S. (1983). Aggregation and beyond: Some basic issues in the prediction of behaviour. *Journal of Personality*, *51*, 360–392.

Fagot, B., Hagan, R., Leinbach, M., & Kronsberg, S. (1985). Differential reactions to assertive and communicative acts of toddler boys and girls. *Child Development*, *56*, 1499–1505.

Flannery, K., & Watson, M. (1993). Are individual differences in fantasy play related to peer acceptance levels? *Journal of Genetic Psychology*, *154*, 407–416.

Gardner, F. (2000). Methodological issues in the direct observation of parent–child interaction: Do observational findings reflect the natural behaviour of participants? *Clinical Child and Family Psychology Review*, *3*, 185–198.

Goodman, S., Brogan, D., Lynch, M., & Fielding, B. (1993). Social and emotional competence in children of depressed mothers. *Child Development*, *64*, 516–531.

Grant, L. (1985). Race-gender status, classroom interactions, and children's socialization in elementary school. In L. Wilkinson & C. Marrett (Eds.), *Gender influences in classroom interaction*. Orlando: Academic Press.

Hay, D.F., Pawlby, S., Sharp, D., Schmücker, G., Mills, A., Allen, H., & Kumar, R. (1999). Parents' judgements about young children's problems: Why mothers and fathers might disagree yet still predict later outcomes. *Journal of Child Psychology and Psychiatry*, *40*, 1249–1258.

Hops, H., Davis, B., & Longoria, 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.

Hughes, C., Cutting, A.L., & Dunn, J. (2001). Acting nasty in the face of failure? Longitudinal observations of 'hard to manage' children playing a rigged competitive game with a friend. *Journal of Abnormal Child Psychology*, *29*, 403–416.

Jones, R., Reid, J.B., & Patterson, G. (1975). Naturalistic observations in clinical assessment. In P. Reynolds (Ed.), *Advances in psychological assessment* (vol. 3, pp. 42–95). San Francisco: Jossey-Bass.

[3] In Murray et al. (2001), the cards were dealt to the two children consecutively (except for the final deal), and progress along the snake was marked by sticky stars.

Leaper, C. (1991). Influence and involvement in children's discourse: Age, gender and partner effects. *Child Development*, *62*, 797–811.

Loeber, R., Green, S., Lahey, B., & Stouthamer-Loeber, M. (1989). Optimal informants on childhood disruptive behaviours. *Development and Psychopathology*, *1*, 317–337.

Maccoby, E.E. (1998). *The two sexes: Growing up apart, coming together*. Cambridge, MA: Harvard University Press.

Masten, A., & Curtis, W. (2000). Integrating competence and psychopathology: Pathways toward a comprehensive science of adaption in development. *Development and Psychopathology*, *12*, 529–550.

Miller, P., Danaher, D., & Forbes, D. (1986). Sex-related strategies for coping with interpersonal conflict in children aged five and seven. *Developmental Psychology*, *22*, 543–548.

Minton, C., Kagan, J., & Levine, J. (1971). Maternal control and obedience in the two-year-old. *Child Development*, *42*, 1873–1894.

Moffitt, T.E. (1993). The neuropsychology of conduct disorder. *Development and Psychopathology*, *5*, 135–152.

Murray, L., Woolgar, M., Cooper, P., & Hipwell, A. (2001). Cognitive vulnerability to depression in five year old children of depressed mothers. *Journal of Child Psychology and Psychiatry*, *42*, 891–900.

Nolen-Hoeksema, S., Girgus, J., & Seligman, M. (1986). Learned helplessness in children: A longitudinal study of depression, achievement, and explanatory style. *Journal of Personality and Social Psychology*, *51*, 435–442.

Patterson, G., Dishion, T., & Chamberlain, P. (1993). Outcomes and methodological issues relating to treatment of antisocial children. In T. Giles (Ed.), *Handbook of effective psychotherapy* (pp. 43–88). New York: Plenum Press.

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

Realmuto, G., August, G., & Hektner, J. (2000). Predictive power of peer behavioural assessment for subsequent maladjustment in community samples of disruptive and nondisruptive children. *Journal of Child Psychology and Psychiatry*, *41*, 181–190.

Rholes, W., Blackwell, J., Jordan, C., & Walters, C. (1980). A developmental study of learned helplessness. *Developmental Psychology*, *16*, 616–624.

Smith, P., & Boulton, M. (1990). Rough and tumble play, aggression, and dominance: Perception and behaviour in children's encounters. *Human Development*, *33*, 271–282.

Stevenson-Hinde, J., & Glover, A. (1996). TI: Shy girls and boys: A new look. *Journal of Child Psychology and Psychiatry*, *37*, 181–187.

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

Underwood, M., Hurley, J., Johanson, C., & Mosley, J. (1999). An experimental, observational investigation of children's response to peer provocation: Developmental and gender differences in middle childhood. *Child Development*, *70*, 1428–1446.

Wechsler, D. (1990). *Wechsler Preschool and Primary Scale of Intelligence-Revised.* London: The Psychological Corporation: Harcourt Brace and Company.

Wilkinson, G. (1993). *Wide range achievement test* (3rd edn). Wilmington, DE: Wide Range Inc.

Zahn-Waxler, C. (1993). Warriors and worriers: Gender and psychopathology. Special issue: Toward a developmental perspective on conduct disorder. *Development and Psychopathology*, *5*, 79–89.

Manuscript accepted 15 November 2001