Journal of Abnormal Child Psychology, Vol. 14, No. 2, 1986, pp. 331-345

Agreement Between Teachers' Ratings and Observations of Hyperactivity, Inattentiveness, and Defiance[1]

Russell Schachar[2]
The Hospital for Sick Children, Toronto

Seija Sandberg and Michael Rutter
Institute of Psychiatry, London

The relationship of directly observed classroom behavior and teacher ratings on questionnaires with a judgmental (Conners Teacher Rating Scale) or operational format was examined for 33 boys aged 6 years 5 months to 7 years 7 months. Results showed a high degree of association between observed and rated behavior. This association did not vary with the format of the rating scales but did vary with the nature of the behavior being rated. Defiance was more reliably rated than hyperactivity or inattentiveness. Several behaviors exerted a halo effect on ratings of hyperactivity, inattentiveness, and behavior problems. In particular, defiance toward a teacher increased the likelihood that a child would be rated as hyperactive or inattentive regardless of his observed level of activity or attentiveness. These results support the validity of behavior rating scales as screening measures for hyperactivity, inattentiveness, and defiance and indicate that a child's defiance and disobedience are significant causes of misclassification.

In the study of the hyperkinetic syndrome of childhood (ICD-9; World Health Organization, 1978) or the Attention Deficit Disorder (DSM-III; Spitzer, 1980), the Conners Teacher Rating Scale (CTRS) is perhaps the

Manuscript received in final form October 16, 1985

[1]The first author was supported by the Ontario Mental Health Foundation and the Medical Research Council of Canada. The authors would like to thank Mrs. Penny Roy for her help in the reliability study. The manuscript was prepared with the assistance of the Medical Publications Department, The Hospital for Sick Children, Toronto.

[2]Address all correspondence to Russell Schachar, Department of Psychiatry, The Hospital for Sick Children, 555 University Avenue, Toronto, Ontario, Canada M5G 1X8.

0091-0627/86/0600-0331$05.00/0

most commonly used behavior rating. This scale, developed to measure change resulting from drug treatment among hyperkinetic children (Conners, 1969, 1973), consists of 39 general statements about childhood behavior, and the teacher must decide to what degree each statement is characteristic of a particular child. The CTRS and its 10-item abbreviated version have been recommended as an aid to the diagnosis of the hyperkinetic syndrome (Goyette, Conners, & Ulrich, 1978; Sprague & Sleator, 1973; Sprague, Christensen, & Werry, 1974), as an economical diagnostic screening instrument, and as a basis for assigning subjects to experimental groups that Conners (1970) contends may be "more objective...than the psychiatric diagnosis" (p. 680) (see also Douglas, Parry, Marton, & Garson, 1976).

Conclusions about the validity of the CTRS are based on its high test-retest and interrater reliability (Kazdin, Esveldt-Dawson, & Loar, 1983; Zentall & Barack, 1979), its sensitivity to drug-induced behavior change (Conners & Werry, 1979; Eisenberg & Conners, 1971; Werry & Sprague, 1974), and its ability to differentiate normal children from those diagnosed as hyperkinetic according to usual clinical practice (Conners, 1970; Kupietz, Bialer, & Winsberg, 1972; Sprague, Cohen, & Werry, 1974). However, these features do not establish the validity of CTRS as a diagnostic instrument or a measure of activity. To be valid, the ratings should correlate with independent measures of the same behavior, not be unduly influenced by the presence of other behaviors, and distinguish hyperactive from nonhyperactive disturbed children as well as from normal controls.

CTRS hyperactivity ratings show a small but significant correlation with a variety of independent measures of activity (Sandberg, 1981) and with a composite of observed disturbing behaviors (Jacob, O'Leary, & Rosenblad, 1978). However, correlation of individual CTRS factors and observations of similar behaviors are frequently low (Campbell, Schleifer, & Weiss, 1978; Copeland & Weissbrod, 1978; Rapoport & Benoit, 1975; Sandberg, 1981).

There are several possible explanations for these low correlations. First, CTRS ratings might be subject to a halo effect whereby the presence of some particular behaviors affects ratings of phenomenologically different behaviors (Guilford, 1954). This conclusion has been supported by several studies. For example, Abikoff, Gittelman-Klein, and Klein (1977, 1980) found that children rated as hyperactive were better differentiated from nonhyperactive children by a combination of observed disruptiveness and inattentiveness than by gross motor activity.

Second, these low correlations might be a function of the global nature of the CTRS. Frequency counts of behavior based on direct observations and rating scales composed of items as discrete and readily identifiable

as those of behavior observation categories are more accurate and less subject to a halo effect (Kent, O'Leary, Diament, & Dietz 1974; Neisworth, Kurtz, Jones, & Madle, 1974; Shuller & McNamara, 1976; Siegel, Dragovich, & Marholin, 1976; Wahler & Leske, 1973; Ysseldyke & Foster, 1978).

Third, despite the fact that the hyperactivity dimension of behavior is thought to be conceptually and statistically distinct from conduct problems or inattentiveness, the CTRS hyperactivity factor contains items such as disturbs others and teases others. Therefore, CTRS hyperactivity ratings may be influenced by behaviors that are conceptually more like conduct disturbance than they are like hyperactivity.

With these problems in mind, we assessed the agreement between CTRS ratings of hyperactivity and observed classroom hyperactivity. The measure of observed hyperactivity was calculated in two ways. First, scores for observations of all behaviors included in the CTRS hyperactivity factor were summed, and second, observation items related to disturbing and teasing others were omitted in order to determine the effect of these items on agreement. In addition, we attempted to determine the extent to which a questionnaire consisting of clearly defined behaviors with operational rather than judgmental response categories can result in greater agreement between rated and observed behavior, and to examine the extent to which these two types of activity ratings are subject to a halo effect.

For comparison, we assessed the agreement of inattentiveness and behavior problem ratings on both types of questionnaires with observed classroom behavior. We also examined the extent to which ratings of these behaviors were subject to a halo effect.

METHOD

Subjects

Teachers completed a CTRS for each of the 185 boys in the final year of infant school (North American grade 1 equivalent) in six London schools. The children were grouped according to scores on the hyperactivity factor (Conners, 1973) using norms obtained from a previous study of 226 children from the same part of London (Sandberg, Wieselberg, & Shaffer, 1980). A high activity level was defined as a CTRS hyperactivity factor score >1 standard deviation above the mean for the larger sample and low activity level was a score >1 standard deviation below the mean; any score between these two extremes was considered to represent a midactivity level.

The study sample of 33 children was selected at random from these three strata, with the restriction that there must not be an equal number in each stratum at any particular school. This was intended to make it less likely that the observer could guess the rated activity levels of the children being observed. There were 5 or 6 children in each of the six schools.

Observation Procedure

The 33 children selected were studied in their classrooms every school day for two 1-week periods by an observer who was blind to the CTRS scores. In most cases, behavior was recorded during consecutive weeks and there was never more than a 2-week interval, usually due to school holidays. The teacher identified the target child to the observer, who was introduced to the class as a student teacher (student teachers frequently audit these classrooms). No information about the target child was given to the observer.

A time-sampling procedure was used to document the behavior of the target child. We observed the behaviors that corresponded to the items of the CTRS hyperactivity, inattentiveness, and behavior problem factors. Piloting was carried out in order to define the observation categories as clearly as possible.

The observers wore earphones attached to cassette recorders hidden beneath their clothing. This recording provided the timing for the observation and recording intervals. The observers spent 30 minutes in the classroom before beginning the first observation period. This allowed the children to get used to the observers' presence and allowed the observers to establish the best view of the child without making it obvious who was being observed.

During each 10-second observation period the investigator noted the assigned task (structured or unstructured task), whether the child was on or off task, his physical activity (gross and fine motor movements, fiddling with objects), his location in relation to other children, and the nature of his interactions with peers and teacher (see Table I). Following each 10-second observation, 15 seconds was spent in recording the occurrence of any category of behavior listed in Table I. The investigator also noted how many times the target child looked at or approached him. Then the observation and recording cycle was repeated for a total of 10 minutes.

After each target child in the school was studied for 10 minutes, the observer repeated the cycle of 10-minute observation periods so a sample of each child's behavior was obtained approximately every hour continuously throughout the school day.

Table I. Interrater Reliability of Observation Categories[a]

Category	Beginning of study		End of study	
	Nominal agreement[b]	Kappa[c]	Nominal agreement	Kappa
Minor motor movements	.85	.67	.85	.67
Gross motor movements	.96	.96	.94	.79
Fiddling	.95	.73	.95	.75
Self-vocalization	.93	.72	.92	.67
Daydreaming	.85	.67	.93	.79
Disruptive-noisy	.99	.67	1.00	1.00
Negative interaction with peers	.99	.80	.99	.99
Negative interaction with teacher	.99	.90	.98	.80
Positive interaction with peers	.95	.84	.95	.85
Positive interaction with teacher	.98	.94	.99	.94
Alone	.98	.92	.98	.88
Nature of task	.97	.93	.97	.93
On/off task	.95	.88	.91	.81

[a]The observation protocol is available upon request from the first author.
[b]Number of agreements as a percent of total number of times the behavior was recorded by either observer.
[c]As above, corrected for chance agreements.

A score for each category of observed behavior was calculated as a percentage of the observation periods in which the behavior was noted. We calculated a score for observed hypearctivity, inattentiveness, and behavior problem dimensions by summing items that corresponded to those included in each CTRS factor (see below). The direct observation (DO) hyperactivity score was a sum of fiddling, self-vocalization, gross motor movement (total), minor motor movement, negative interaction with peers, and disruptive-noisy. A second measure of DO hyperactivity was calculated by omitting negative interaction with peers and disruptive-noisy. The DO inattentiveness measure was a sum of not on structured task, not on unstructured task, and daydreaming. The DO behavior problem score was a sum of negative interaction with peers, negative interaction with teacher, and disruptive-noisy.

Collection of observation data was not begun until interobserver agreement for all categories was at least 80%. To reduce the possibility of criterion drift (Reid, 1970), interrater reliability of the direct observations was checked at the beginning and toward the end of the study for a total of 500 observations. Three observers were involved in determining reliability and two in the study itself. A split earplug was used during reliability assessment to ensure that both observers had the same time signals.

Direct Observation Questionnaire

An 11-item Direct Observation Questionnaire (DOQ) was developed in consultation with teachers to assess behaviors similar to those noted in the CTRS and direct observations.[3] In the DOQ each behavior being rated was defined as clearly as possible and the responses were requested either as the number of instances of the behavior noted (for low-frequency behaviors) or the proportion of time during which a behavior occurred (for high-frequency behaviors). Ratings were indicated on a scale from 0 to 100.

The teachers were asked to familiarize themselves with the DOQ at the beginning of each week that the children were being observed and to complete the questionnaire at the end of the week. They were asked to avoid making judgments based on their expectations of age-appropriate behavior but to complete the form solely on the basis of their observations of the child's behavior during that week.

We calculated a DOQ hyperactivity score corresponding to the CTRS hyperactivity factor score for each child by adding the scores for the following six items: fiddling, self-vocalization, gross motor movement, minor motor movement, negative interaction with peers, and disruptive-noisy. As was done for the DO scores, a second DOQ hyperactivity score was calculated by omitting negative interaction with peers and disruptive-noisy. The DOQ inattentiveness score was based on ratings of on-task behavior and daydreaming, and the DOQ behavior problem score was the sum of disruptive-noisy, negative interaction with peers, and negative interaction with teacher.

CTRS Ratings

At the end of each week of observation, teachers completed a CTRS for each child along with the DOQ. On the basis of these ratings, CTRS hyperactivity, inattention, and behavior problem factor scores were calculated (Conners, 1973). The individual CTRS items constituting the hyperactivity factor are fidgets, makes odd noises, overactive, excitable and/or impulsive, disturbs others, and teases others. Those constituting the inattention factor are coordination poor, distractible, inattentive, daydreams, easily led, and lacks leadership. The CTRS items included in the behavior problem factor are sullen or sulks, quarrelsome, destructive, steals, lies, acts "smart," explosive, no sense of fair play, (not) submissive, defiant, impudent, stubborn, and uncooperative with teacher.

[3]A copy of the questionnaire is available on request from the first author.

RESULTS

Interrater Agreement

Interrater agreement was calculated using both observer agreement coefficients and Cohen's Kappa, a reliability coefficient that takes into account the likelihood of chance agreement (Frick & Semmel, 1978).

Agreement ranged from a high of .99 for disruptive-noisy to a low of .67 for minor motor movement and did not vary greatly from beginning to end of the study (see Table I).

Characteristics of the Sample

The mean age of the children was 7 years 2 months (range 6 years 5 months to 7 years 7 months). Of the 33 children, 8 had CTRS hyperactivity factor scores >1 standard deviation above the mean, 12 had scores >1 standard deviation below the mean, and the remaining 13 had scores between these. The 8 with high CTRS hyperactivity factor scores also had abbreviated CTRS scores >15, the proposed cutoff for diagnosing the hyperkinetic syndrome (Goyette et al., 1978; Sprague & Sleator, 1973).

The mean number of observations per child over the 2 weeks was 396 ($SD = 80$), the equivalent of 165 minutes of observation.

A wide range of DOQ and DO category scores was found (see Tables II and III). The narrowest range of scores was seen in self-vocalization and negative interaction with the teacher for the DOQ, and in self-vocalization, negative interaction with peers, negative interaction with teacher, and disruptive-noisy on DO.

No significant difference in observations and ratings from week 1 to week 2, assessed using a paired t test ($p > .05$), was noted in CTRS, DOQ, or DO scores for the group as a whole or for any activity stratum. Therefore, measures were averaged across the 2 weeks.

Relationship of Rated and Observed Behaviors

Agreement between ratings and observations of behavior was assessed in two complementary ways. First, the correlation of all observation categories and rating scale items was calculated to compare the correlations of "similar" items (e.g., rated and observed activity) and "dissimilar" items (e.g., rated activity and observed disruptiveness).

Table II. Mean Scores for Direct Observation Questionnaire Items

DOQ item	*Mean*	*SD*	Range
1. On task	51.2	21.3	6.5–97.5
2. Daydreaming	18.2	19.1	0–90
3. Fiddling	21.8	18.4	0–75
4. Self-vocalization	11.8	14.9	0–55
5. Gross motor movements	24.9	22.1	2.5–73.5
6. Minor motor movements	18.1	20.3	0–79.5
7. Positive interaction with peers	37.7	17.0	15–85
8. Positive interaction with teacher	18.6	18.1	0–91.9
9. Negative interaction with peers	16.3	17.3	0–78.5
10. Negative interaction with teacher	4.0	8.7	0–45.8
11. Disruptive-noisy	10.5	14.8	0–65.0

Correlations between ratings on individual items and DO scores of corresponding behaviors were very similar for the two types of questionnaires. However, agreement with the observational measures was higher for the CTRS scores in gross motor movements (.41 vs. .35), minor motor movements (.31 vs. .25), and negative interaction with peers (.57 vs. .45), and higher for the DOQ in negative interactions with teacher (.62 vs. .66).

On the other hand, correlations between ratings on individual items and DO scores of dissimilar items were often higher than correlations between ratings and similar items. For example, CTRS score of activity was

Table III. Mean Scores for Observation Categories

DO item	*Mean*[a]	*SD*	Range
1. Minor motor movements	35.9	16.5	8.6–75.9
2. Gross motor movements (total)	22.7	7.2	8.5–35.8
3. Gross motor movements (structured situations)	20.8	7.1	7.8–33.1
4. Fiddling	7.1	5.1	0.4–24.3
5. Self-vocalization	5.8	3.3	0.8–14.3
6. Daydreaming	9.8	5.4	2.3–22.0
7. Disruptive-noisy	1.8	2.7	0–11.7
8. Negative interaction with peers	2.5	2.8	0–10.8
9. Negative interaction with teacher	2.3	2.5	0–11.1
10. Positive interaction with peers	24.4	9.6	5.7–47.2
11. Positive interaction with teacher	9.4	5.2	2.2–26.0
12. Alone	9.0	5.8	0.5–22.2
13. Not on structured task	22.7	12.8	2.4–52.1
14. Not on unstructured task	18.4	14.2	0–58.3

[a]The mean is given as the percentage of the observation periods in which the behavior item was noted.

Table IV. Agreement Between Classifications Based on Ratings and Observations ($N = 33$)[a]

Factor rated and scale of rating	Correct positive (CP)	False negative (FN)	Sensitivity[b] (%)	False positive (FP)	Correct negative (CN)	Specificity[c] (%)	Accuracy[d] (%)
			Observed hyperactivity score				
Hyperactivity							
CTRS	11	6	65	6	10	63	64
DOQ	12	5	71	4	12	75	73
			Observed inattentiveness score				
Inattentiveness							
CTRS	10	6	63	6	11	65	64
DOQ	10	6	63	7	10	59	61
			Observed behavior problem score				
Behavior problem							
CTRS	13	4	76	2	14	88	82
DOQ	11	6	65	6	10	63	64

[a]Negative interaction with peers and disruptive-noisy are included in DOQ and DO hyperactivity scores (refer to p. 335, paragraph 1).
[b]Sensitivity = CP/(CP + FN).
[c]Specificity = CN/(FP + CN).
[d]Accuracy = (CP + CN)/(CP + FN + FP + CN).

more highly correlated with observations of negative interaction with teacher than with observed activity (.67 vs. .41). The same was true for DOQ ratings of activity, which correlated more highly with observations of negative interaction with teacher than with observed activity (.41 vs. .32).[4]

Agreement of Classification Based on Ratings and Observations

To check the agreement of ratings and observations, hyperactivity, inattentiveness, and behavior problem scores on DOQ, CTRS, and DO were categorized as "high" or "low," depending on whether they were above or at/below the group medians, and agreement between the classifications based on ratings and observations was calculated.

DOQ ratings of hyperactivity were more sensitive, specific, and accurate in identifying observed hyperactivity than were CTRS hyperactivity ratings (see Table IV). When the items negative interaction with peers and disruptive-noisy were excluded from DOQ and DO hyperactivity scores, accuracy of classification based on DOQ hyperactivity scores decreased from

[4]The table of correlations is available on request from the first author.

73% to 61%. Conversely, the accuracy of CTRS hyperactivity rating in predicting DO hyperactivity scores was similar regardless of whether these two items were included (64% and 67%). CTRS ratings of behavior problems were considerably more accurate than DOQ scores in classifying children with respect to DO behavior problem. There was little difference in the accuracy of DOQ and CTRS inattentiveness ratings. Accuracy of classifications based on ratings ranged from 82% to 61%.

Halo Effect

To assess the halo effect of one behavior on ratings of another behavior, "correctly" and "incorrectly" rated children were compared with respect to other aspects of their observed behavior. For these analyses, correctly rated children were those for whom the classification based on ratings and observations of hyperactivity, inattentiveness, and behavior problems were similar (i.e., both high or both low). The children correctly rated as hyperactive (correct positives, CP) were compared with those who were incorrectly rated as not hyperactive (false negatives, FN), and those correctly rated as not hyperactive (correct negatives, CN) were compared with those incorrectly rated as hyperactive (false positives, FP). Similar analyses were conducted for inattentiveness and behavior problems.

The numbers of children in each group that scored above the group median for each category of observed behavior were compared. The statistical significance of the differences between the groups was tested with the Fisher exact probability test.

We found several significant differences in observed behavior between correctly and incorrectly rated children (see Table V). Eighty-four comparisons were made and 14 (17%) were significant, which exceeds the number expected by chance alone ($p < .05$).

Ratings on the CTRS hyperactivity factor were subject to a greater number of these halo effects than were DOQ hyperactivity ratings, whereas DOQ behavior problem ratings were more influenced by halo effects than were CTRS ratings. Overall, the two types of ratings were equally subject to these effects. In particular, correct and incorrect ratings of hyperactivity were most often distinguished by observed negative interactions with teacher and peers, and not on task behavior. For example, 4 of 6 children incorrectly rated as hyperactive on CTRS scored above the mean for observed negative peer interactions, compared with 1 of 10 children correctly rated as not hyperactive ($p < .03$).

Exclusion of items relating to negative interaction with peers and disruptive-noisy from DOQ and DO hyperactivity scores had little impact

Table V. Significant Differences Between Observed Behaviors for Children Correctly and Incorrectly Rated as Being Hyperactive or Inattentive or Having Behavior Problems (Halo Effect)[a]

	Hyperactivity[b]		Inattentiveness		Behavior problems	
Observed behavior	CTRS	DOQ	CTRS	DOQ	CTRS	DOQ
5. Self-vocalization	FN>CP	FN>CP				
8. Negative interaction with peers	FP>CN			FP>CN		
9. Negative interaction with teacher	CP>FN	CP>FN	FP>CN	FP>CN		
10. Positive interaction with peers						FP>CN
11. Positive interaction with teacher	CN>FP					
13. Not on structured task	FP>CN					CP>FN & FP>CN
14. Not on unstructured task	CP>FN					FP>CN

[a]Only differences significant at $p < .05$ are shown.

[b]CN = correct negative, CP = correct positive, CTRS = Conners Teacher Rating Scale, DOQ = Direct Observation Questionnaire, FN = false negative, FP = false positive.

on halo effects. The only difference was that CN and FP no longer differed with respect to positive interactions with teacher.

DISCUSSION

We found considerable agreement between teachers' ratings on the two questionnaires and independent observations of behavior. The correlations were much higher than those found in some studies (Campbell et al., 1978) but similar to those of others (Kazdin et al., 1983). To some extent, these differences might result from the restricted range of scores in other investigations. In our study, random sampling of children with varying levels of activity ensured a wide range in behaviors and ratings. Moreover, there was no evidence that the degree of correlation between ratings and observations was a function of the range of scores on particular behaviors.

The extent to which the ratings and observations agreed depended on the behavior under consideration. For both the CTRS and the DOQ, correspondence of ratings and observations of defiant and disruptive behaviors was greater than for behaviors such as inattentiveness, hyperactivity, and self-vocalization. For behavior such as daydreaming, no significant correlation between ratings and observations was found. While this might suggest that the definitions used in the ratings and observations are more alike for defiance than for other behaviors, a more plausible explanation is that salient behaviors are simply more easily and accurately rated.

Despite our predictions, the DOQ (comprising behaviors and response categories defined in detail) did not produce more reliable ratings across all behaviors than did the CTRS and was no less likely to be subject to halo effects. Indeed, there was little to choose between the two instruments with respect to correlations with individual behaviors or their agreement with DO of children as above or below median scores for hyperactivity, inattentiveness, or behavior problems.

Perhaps differences were reduced because the teachers completed both types of rating scales and because of the emphasis given to careful observation by the presence of the observer. Teachers may also be too busy in a typical classroom to make detailed judgments about child behavior (see Kazdin et al., 1983).

Other studies have suggested that the low correlations between rated and observed behavior might result from a halo effect emanating from phenomenologically different behaviors. Blunden, Spring, and Greenberg (1974) reported such a halo effect in ratings of restlessness, poor concentration, and sociability on the Classroom Behavior Inventory. These ratings were all more strongly associated with observed impulsiveness than observ-

ed restlessness, attentiveness, or sociability. Studies by Abikoff et al. (1977, 1980) also suggest that CTRS ratings may be susceptible to a similar halo effect. They found that children rated as hyperactive on the CTRS were better distinguished from those not rated as hyperactive by a combination of observed disruptiveness and inattentiveness than by gross motor activity.

The present study confirms that the halo effect of various behaviors on ratings of hyperactivity, inattentiveness, and behavior problems accounts for some of the misclassification by these scales (e.g., rated but not observed as hyperactive; observed but not rated as hyperactive).

Of particular interest is the evidence of a halo effect of difficult relationships with peers and teacher on ratings of hyperactivity, and the absence of a halo effect of hyperactivity on ratings of behavior problems. Children who had poor relationships (i.e., frequent negative interactions and infrequent positive interactions with peers and teacher) were more likely to be rated as hyperactive irrespective of observed hyperactivity. This applied regardless of whether items relating to disruptiveness are included in the score of observed hyperactivity. CTRS hyperactivity ratings were most liable to this halo effect of negative relationships.

Both these findings agree with those of Prinz, Connor, and Wilson (1981). By comparison, Stevens-Long (1973) found a clear halo effect of overactivity on aggressiveness. In that study using videotapes, overactive children were judged to be more aggressive than normally active children even though the rate of aggressive acts was equal on the tapes.

The halo effect of observed self-vocalization on rated hyperactivity suggests that children who talk to themselves a great deal are judged not to be hyperactive despite high scores for observed hyperactivity. Perhaps frequent self-vocalization prompts an alternative conception of a child's behavior. This result is surprising because "hums and makes other odd noises" is included in the CTRS hyperactivity factor and self-vocalization was part of the DOQ and DO hyperactivity factors.

Although observed hyperactivity did not affect behavior problem ratings, other behaviors did have such an effect. Either observed inattentiveness or positive interactions with peers was associated with a greater likelihood of a child being rated as showing behavior problems regardless of observed behavior. These effects may reflect the perceived "nuisance" created by children who are overly sociable or who do not attend to assigned tasks even though these behaviors are not behavior problems in a strict sense.

Aggressive and defiant social interactions with teacher and peers had a similar halo effect on rated inattentiveness. Children were more likely to be incorrectly rated as inattentive in the presence of this type of behavior.

Our results indicate that questionnaire measures of either type can play a useful role in screening for hyperactivity, inattentiveness, or behavior

problems. About two-thirds of the children are correctly classified by these ratings using observations as a criterion of validity. However, there are differences in the ability of factor ratings to specifically measure the behaviors implied in their names. In particular, hyperactivity ratings may be influenced by observed defiance and inattentiveness, inattentiveness ratings by observed defiance, and behavior disorder by observed hyperactivity and inattentiveness. The present results are of concern in view of current emphasis on distinguishing these three behaviors for clinical and theoretical reasons and the continued reliance on rating scales with a subjective or global format (McGee, Williams, & Silva, 1984).

REFERENCES

Abikoff, H., Gittelman-Klein, R., & Klein, D. F. (1977). Validation of a classroom observation code for hyperactive children. *Journal of Consulting and Clinical Psychology, 45,* 772-783.

Abikoff, H., Gittelman-Klein, R., & Klein, D. F. (1980). Classroom observation code for hyperactive children: A replication of validity. *Journal of Consulting and Clinical Psychology, 48,* 555-565.

Blunden, D., Spring, C., & Greenberg, L. M. (1974). Validation of the classroom behavior inventory. *Journal of Consulting and Clinical Psychology, 42,* 84-88.

Campbell, S. B., Schleifer, M., & Weiss, G. (1978). Continuities in maternal reports and child behaviors over time in hyperactive and comparison groups. *Journal of Abnormal Child Psychology, 6,* 33-45.

Conners, C. K. (1969). A teacher rating scale for use in drug studies with children. *American Journal of Psychiatry, 126,* 884-888.

Conners, C. K. (1970). Symptom patterns in hyperkinetic, neurotic, and normal children. *Child Development, 41,* 667-682.

Conners, C. K. (1973). Rating scales in drug studies with children. *Psychopharmacology Bulletin,* Special Issue (Pharmacotherapy of Children), 24-29.

Conners, C. K., & Werry, J. S. (1979). Pharmacotherapy. In H. C. Quay & J. S. Werry (Eds.), *Psychopathological disorders of childhood* (2nd ed., pp. 336-386). New York: Wiley.

Copeland, A. P., & Weissbrod, C. S. (1978). Behavioral correlates of the hyperactivity factor of the Conners Teacher Questionnaire. *Journal of Abnormal Child Psychology, 6,* 339-343.

Douglas, V. I., Parry, P., Marton, P., & Garson, C. (1976). Assessment of a cognitive training program for hyperactive children. *Journal of Abnormal Child Psychology, 4,* 389-410.

Eisenberg, L., & Conners, C. K. (1971). Psychopharmacology in childhood. In N. B. Talbot, J. Kager, & L. Eisenberg (Eds.), *Behavioral science in pediatric medicine* (pp. 397-424). Philadelphia: Saunders.

Frick, T., & Semmel, M. I. (1978). Observer agreement and reliabilities of classroom observational measures. *Review of Educational Research, 48,* 157-184.

Goyette, C. H., Conners, C. K., & Ulrich, R. F. (1978). Normative data on revised Conners Parent and Teacher Rating Scales. *Journal of Abnormal Child Psychology, 6,* 221-236.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Jacob, R. G., O'Leary, K. D., & Rosenblad, C. (1978). Formal and informal classroom settings: Effects on hyperactivity. *Journal of Abnormal Child Psychology, 6,* 47-59.

Kazdin, A. E., Esveldt-Dawson, K., & Loar, L. L. (1983). Correspondence of teacher ratings and direct observations of classroom behavior of psychiatric inpatient children. *Journal of Abnormal Child Psychology, 11,* 549-564.

Kent, R. N., O'Leary, K. D., Diament, C., & Dietz, A. (1974). Expectation biases in observational evaluation of therapeutic change. *Journal of Consulting and Clinical Psychology, 42,* 774-780.

Kupietz, D., Bialer, I., & Winsberg, B. G. (1972). A behavior rating scale for assessing improvement in behaviorally deviant children: A preliminary investigation. *American Journal of Psychiatry, 128,* 1432-1436.

McGee, R., Williams, S., & Silva, P. A. (1984). Behavioral and developmental characteristics of aggressive, hyperactive and aggressive-hyperactive boys. *Journal of the American Academy of Child Psychiatry, 23,* 270-279.

Neisworth, J. T., Kurtz, P. D., Jones, R. T., & Madle, R. P. (1974). Biasing of hyperkinetic behavior ratings by diagnostic reports. *Journal of Abnormal Child Psychology, 2,* 323-329.

Prinz, R. J., Connor, P. A., & Wilson, C. C. (1981). Hyperactive and aggressive behaviors in childhood: Intertwined dimensions. *Journal of Abnormal Child Psychology, 9,* 191-202.

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

Reid, J. B. (1970). Reliability assessment of observational data: a possible methodological problem. *Child Development, 41,* 1143-1150.

Sandberg, S. T. (1981). The overinclusiveness of the diagnosis of hyperkinetic syndrome. In M. Gittelman (Ed.), *Strategic interventions for hyperactive children* (pp. 8-38). Armonk, New York: M.E. Sharpe.

Sandberg, S. T., Wieselberg, M., & Shaffer, D. (1980). Hyperkinetic and conduct problem children in a primary school population: Some epidemiological considerations. *Journal of Child Psychology and Psychiatry, 21,* 293-311.

Shuller, D. Y., & McNamara, J. R. (1976). Expectancy factors in behavioral observation. *Behavior Therapy, 7,* 519-527.

Siegel, L. J., Dragovich, S. L., & Marholin, D., III. (1976). The effects of biasing information on behavioral observations and rating scales. *Journal of Abnormal Child Psychology, 4,* 221-233.

Spitzer, R. L. (Chairperson). (1980). DSM-III: Diagnostic and statistical manual of mental disorders (3rd ed.). Washington: American Psychiatric Association.

Sprague, R. L., Christensen, D. E., & Werry, J. S. (1974). Experimental psychology and stimulant drugs. In C. K. Conners (Ed.), *Clinical use of stimulant drugs in children* (pp. 141-164). The Hague: Excerpta Medica.

Sprague, R. L., Cohen, M. N., & Werry, J. S. (1974). *Normative data on the Conners' Teaching Rating Scale and Abbreviated Scale* (Technical Report, November, 1974b, USPHS Grant #MH 18909. Final report, Children's Research Center) (pp. 140-145b). Champaign, Illinois: University of Illinois.

Sprague, R. L., & Sleator, E. K. (1973). Effects of psychopharmacologic agents on learning disorders. *Pediatric Clinics of North America, 20,* 719-735.

Stevens-Long, J. (1973). The effect of behavioral context on some aspects of adult disciplinary practice and affect. *Child Development, 44,* 476-484.

Wahler, R. G., & Leske, G. (1973). Accurate and inaccurate observer summary reports. *Journal of Nervous and Mental Disease, 156,* 386-394.

Werry, J., & Sprague, R. L. (1974). Methylphenidate in children—effect of dosage. *Australian and New Zealand Journal of Psychiatry, 8,* 9-19.

World Health Organization. (1978). *Mental disorders: Glossary and guide to their classification in accordance with the ninth revision of the International Classification of Diseases.* Geneva: Author.

Ysseldyke, J. E., & Foster, G. G. (1978). Bias in teachers' observations of emotionally disturbed and learning disabled children. *Exceptional Children, 44,* 613-615.

Zentall, S. S., & Barack, R. S. (1979). Rating scales for hyperactivity: Concurrent validity, reliability, and decisions to label for the Conners and Davids abbreviated scales. *Journal of Abnormal Child Psychology, 7,* 179-190.