Journal of Psychopathology and Behavioral Assessment, Vol. 26, No. 3, September 2004 (© 2004)

Behavior Ratings and Observations of Externalizing Symptoms in Girls: The Role of Child Popularity With Adults[1]

Amori Yee Mikami,[2] Terry C. Chi,[3] and Stephen P. Hinshaw[2,4]

Accepted November 7, 2003

This study examined staff ratings and live observations of externalizing behavior in 149 girls with and without attention-deficit/hyperactivity disorder (ADHD), who participated in all-female naturalistic research camps. Girls' popularity with adult camp staff was hypothesized to explain discrepancies between ratings and observations. Compared to behavior observations, staff ratings overestimated the externalizing behavior of girls who were disliked by staff. In contrast, ratings and observations were consistent for girls who were liked by staff. Among girls who were disliked by staff, unpopularity with peers predicted a larger discrepancy between staff ratings and observations, but peer status made little difference in rating–observation discrepancies of girls liked by staff. All results held after controlling for the participants' ADHD versus comparison status. Results suggest that staff ratings may be biased by personal feelings about children and that direct observations may be more immune to such bias.

KEY WORDS: behavior ratings; rater biases; observation coding systems; attention-deficit/hyperactivity disorder (ADHD); popularity with adults.

Adult ratings of children's actions are used widely in clinical research, under the assumption that ratings provide accurate assessments of child behavior patterns. Typically, adults report on a global scale about the child's personality and behavior tendencies. Such ratings differ from discrete behavior observations, in which trained observers watch a child for a particular amount of time and record the child's demonstrated behavior during that time only (Cairns & Green, 1979; Hinshaw & Nigg, 1999). A substantial body of research suggests that because global rating scales are more subjective than behavior observation coding systems, they are susceptible to bias.

Critically, global rating scales depend on the complex information-processing capabilities of the rater (Cairns & Green, 1979). The rater is assumed to share with the investigator, and with other raters, a theoretical idea about the attribute to be rated. However, attributes pertinent to externalizing behavior like "quarrelsome" or "has temper outbursts" are abstract. Is the child arguing unjustifiably, admirably standing up for his or her rights, or just joking around? The rater is also assumed to share the same underlying "scale" on which the attribute will be judged. If two raters have different ideas about how often temper outbursts occur in the average child, then this disparity will affect their perceptions of such anchor points as "happened frequently" versus "happened seldom." Additionally, rating scales require the rater to think about a summary of behavior that has occurred over a period of time, which may include hours, days, or even longer. The rater must remember occurrences, determine whether they are relevant, and finally synthesize this into a number—for example, a rating from "0" to "3" (Achenbach, 1991a, 1991b; Cairns & Green, 1979).

By contrast, in observational coding systems observers aspire to record the actual activities of children without making complex information-processing judgments or inferences. Subjective interpretations on the part of the observers are limited, or, ideally, nonexistent (Bakeman & Gottman, 1997). In assessing externalizing behavior, a well-designed behavior observation system includes specific behaviors—"did not comply when issued

[1]Portions of this paper were presented at the American Psychological Society 12th Annual Conference, Miami, Florida, June 2000 and at the Association for Advancement of Behavior Therapy 37th Annual Convention, Boston, Massachusetts, November 2003.

[2]University of California, Berkeley.

[3]Vanderbilt University, Nashville, Tennessee.

[4]To whom correspondence should be addressed at Department of Psychology, 2205 Tolman Hall #1650, University of California, Berkeley, California 94720-1650; e-mail: hinshaw@socrates.berkeley.edu.

0882-2689/04/0900-0151/0

a request by an adult" or "hit, kicked, or bit a peer." Observers watch the target child for a set, short amount of time such as a few seconds or 1 min, and then record whether or not the child engaged in each behavior during that period. Unlike adults who complete global rating scales, behavior observers ideally are trained in order to achieve precise consensus regarding the definition of the events under consideration (Bakeman & Gottman, 1997), a process requiring a range from several hours to as long as 1 month of training (e.g., Ritter & Langlois, 1988).

Because ratings require inference and information processing on the part of the rater, they are susceptible to bias for several different interpersonal, psychological, and contextual reasons (Chi & Hinshaw, 2002; Hart & Lahey, 1999; Richters, 1992). Personal biases toward the individual being rated (halos), as well as biases towards groups (stereotypes), can distort the memory or interpretation of events. Observations require less inference, so observers who record actual instances of behavior ideally should be more precise than raters (Cairns & Green, 1979). For example, participants used either a rating scale or an observation coding system to assess disruptive behavior in a series of videotapes of children (Kent, O'Leary, Diament, & Dietz, 1974; Michelson, Mannarino, Marchione, Kazdin, & Costello, 1985). Participants were told that they were viewing either (a) baseline tapes then posttreatment tapes, with a decrease in disruptive behavior predicted, or (b) baseline tapes then "generalization" tapes, with no change in disruptive behavior predicted. In actuality, all participants saw the same series of videotapes. Participants using rating scales were significantly biased by this expectation, reporting "improvement" when they thought they were viewing the treatment condition; in contrast, those who used the behavior observation system were not affected. In another investigation, researchers told participants that videotaped infants were either "born premature" or "born full-term," when the labels had been randomly assigned (Miller & Ottinger, 1986). Participants who assessed the infant's health and temperament via observation systems reported no difference as a function of the label, but those utilizing rating scales showed a significant bias towards infants labeled premature—as less healthy and harder to care for.

Whereas bias in the aforementioned studies can compromise the validity of research evaluating treatment effectiveness or group differences, making raters unaware of research hypotheses can mitigate the potential for distortion. However, some biases are more difficult to eliminate. For instance, biases against children because of their appearance cannot be easily controlled. Ritter and Langlois (1988) asked participants to rate the social skills of both physically attractive and unattractive children in videotapes, using either a global rating scale or a behavior observation system. In rating scales, attractive targets, compared to unattractive targets, were rated as having higher social skills when faces were visible, but these differences were reduced when the faces were occluded to hide differences in attractiveness. However, with the observation system, participants showed no such bias for the attractiveness of targets.

Biases against children because of ethnicity are also difficult to control, yet such biases have been shown to affect rating scales but not observations. When rating their pupils of African descent, U.S. teachers reported higher problem scores than did Jamaican teachers, yet observational data revealed the opposite—that such children in the United States in fact were better behaved than children in Jamaica. Ratings and observations were congruent for Jamaican teachers, but U.S. teachers consistently overestimated externalizing behavior in ratings versus observations. Because teachers in Jamaica were the same race as students but teachers in the U.S. were almost all White, it is possible that the evaluation of behavioral problems was affected by the ethnic match of teachers and students (Puig et al., 1999).

Similar patterns occurred when depressed versus nondepressed mothers reported information about their child's psychopathology on rating scales as opposed to structured diagnostic interviews, which emphasize specificity of behavioral descriptions. Chi and Hinshaw (2002) showed that on rating scales, depressed mothers overestimated child psychopathology (relative to teacher- and child-reports), but the data from the structured interviews were relatively impervious to the effects of such depressive bias.

Note that bias has not been ubiquitous in investigations of rating scales: for example, Eaton and Enns (1986) found that observational codes and global ratings equally differentiated male from female infants with respect to levels of activity. Additionally, although bias may be theoretically absent in observational systems, this may not be completely true in practice (e.g., Michelson et al., 1985). Indeed, expectations remain powerful even when appraising behavior in observation systems. Furthermore, ratings may capture personality traits, assumed to be maximally predictive of future behavior (Moskowitz, 1990). Rating scale procedures that require raters to aggregate behavior over longer time periods may be more meaningful than observational coding systems, which capture a "snapshot" of behavior (Funder & Colvin, 1991). That raters draw on subtle and unquantifiable memories of behavior when making ratings may in fact make ratings more ecologically valid (Moskowitz, 1986). Yet particularly for inferential constructs like externalizing behavior, it is important to

assess the potential for bias when rating methodologies are used.

This study examined an important but underinvestigated source of potential bias in rating scales: the extent to which the target child is liked by adult raters. Related research from the peer rejection literature suggests that popularity may compromise accurate assessment. Peers underestimate the competencies of rejected children with respect to dimensions ranging from intelligence to physical prowess (Koslin, Harlow, Karlins, & Pargament, 1968). Also, peers interpret the intentions of peer-rejected children as negative while interpreting the intentions of the same behavior in an accepted classmate as positive (see review in Hymel, Wagner, & Butler, 1990). Indeed, children hold prejudices against classmates whom they dislike and hold stereotypes that these rejected children engage in maladaptive, antisocial behavior. Such stereotypes appear resistant to change. That is, people seem less likely to attend to, and therefore encode or remember, information that is inconsistent with their stereotypes (Johnston & Macrae, 1994) and may even deny the truth of such information (O'Sullivan & Durso, 1984).

In this work, it was hypothesized that adults as well as children hold prejudices and stereotypes against children whom they do not like and that these prejudices may create information-processing biases that influence rating scale responses, which require interpretation of behavior. This study examined this largely unexplored possibility. Even though much existing literature has documented that ratings can be influenced by previous expectations, such biases can often be minimized in research studies, for example, by keeping raters unaware of the study hypotheses. However, because adult raters may hold opinions on how much they personally like the child being rated, this bias is pervasive and not easily controlled. Further, standard assessments for children, which inform diagnosis and clinical services, depend heavily on rating scales from parents and teachers. Thus, the real-life implications of bias associated with popularity with the rater are significant.

This study examined the discrepancy between global staff ratings and behavior observations of externalizing behavior of 149 girls with and without attention-deficit/hyperactivity disorder (ADHD), who participated in all-female naturalistic summer research camps. Using sociometric measures similar to those used to assess popularity with peers, girls' popularity with adult camp staff was assessed to see if this factor could explain discrepancies between two chief measures of camp externalizing behavior: ratings versus behavior observations. The large proportion of children with ADHD in this sample was ideal for this study, because children with ADHD typically display high rates of externalizing behavior and tend to be disliked by both peers (Blachman & Hinshaw, 2002) and adults (Mikami & Hinshaw, 2003). Evidence also suggests that knowing a child has ADHD can bias interpretation of his or her behavior. When participants expect that a child with whom they are going to interact has ADHD, a self-fulfilling prophecy may be created that elicits more externalizing behavior from the target child (Harris, Milich, & McAninch, 1998). Although staff in this study was unaware of ADHD diagnostic status by design, this study investigated the potential contribution of diagnostic status to bias on rating scales.

Hypotheses were that, with statistical control of ADHD status (a) compared to behavior observations, global staff ratings would *over*estimate the externalizing behavior of girls who were unpopular with staff; and (b) compared to behavior observations, global staff ratings would *under*estimate the externalizing behavior of girls who were popular with staff. Additionally, this study explored the relative contributions of the target child's popularity with adults versus popularity with peers. Some children get along better with adults than with peers, or vice versa; these distinctions are associated with different patterns of externalizing behavior (Mikami & Hinshaw, 2003). Because being liked by adults versus peers has different implications for adjustment, this study investigated how popularity with adults and popularity with peers might interact in predicting bias in rating scales, such that peer status would contribute to rating versus observational discrepancies only when the child was highly popular or unpopular with adults.

METHOD

Overview

This research was conducted during the course of two 5-week summer enrichment programs at the University of California, Berkeley, in 1998 and 1999, the purpose of which was to examine social behavior among girls with ADHD versus comparison girls (see Hinshaw, 2002). Participants were 149 girls from 6 to 12 years of age, 91 (61%) of whom were selected with the primary inclusion criteria of ADHD and 58 (39%) of whom were comparison girls. Prior to their involvement in the summer programs, prospective participants and their families took part in several hours of assessments, including adult informant rating scales and structured clinical interviews to confirm ADHD or comparison status. The participants were also administered self-report measures of emotional functioning, cognitive and achievement testing, and a neuropsychological

battery (see Hinshaw, Carte, Sami, Treuting, & Zupan, 2002).

The structure of the summer program and the activities provided were similar to those from typical, nontherapeutic camps except that many observations and interviews were conducted for research purposes only. During the summer program, girls with ADHD and comparison girls were intermixed for all activities. Girls were divided into three classrooms based on age (6–8, 8.5–10, and 10.5–12.5 years of age), with about 25–26 girls in each classroom. The summer program continued for 5 weeks from 8:30 a.m. to 3 p.m. The daily schedule included two classes, two outdoor play activities, art, drama, snack, and lunch. Families of any girls with ADHD who had been receiving stimulant medication prior to the program were requested to allow their daughters participation while unmedicated. The majority did so; for the 14 girls whose families requested a medication trial, data herein reflect behavior patterns during unmedicated periods. During some of the program periods, girls were taken aside for individual interviews for research purposes to assess popularity with peers, among other variables (e.g., self-perceptions, neuropsychological assessment, laboratory tasks).

Staff including undergraduate students, graduate students, and staff at the B.A. level with clinical training completed Daily Behavior Ratings (DBRs) of the girls' externalizing behavior at the end of each day; this constituted our global rating scale. Next, trained undergraduate staff made live observations of externalizing behavior (OBS) throughout the program; this constituted the behavior observation coding system. Finally, at the end of each camp, graduate students and staff at the B.A. level with clinical training completed sociometric ratings on how much they personally liked and disliked all girls. The staff who provided DBRs, OBS, and the child's popularity with adult staff were not told of the diagnostic status of participants. Additional key measures included (a) peer sociometric interviews to appraise the girls' popularity with peers, and (b) precamp diagnostic assessments, in order to control for ADHD versus comparison status in this investigation. For a detailed overview of the summer program methodology and characteristics of the sample, see Hinshaw (2002). Note that only two of the three summer programs were included in the present investigation, because ratings of the child's popularity with adult staff were collected only during the 1998 and 1999 summer programs.

Participants

Participants were recruited through medical centers, local school districts, clinics, and groups for parents of children with ADHD (e.g., CHADD) as well as advertising in daily newspapers. Wide ethnic diversity was achieved, with half of the sample White (50%) and the remainder divided between African American (26%), Latina (13%), and Asian American (11%); one girl (<1%) was Native American. This ethnic distribution was fairly representative of the population in the San Francisco Bay Area. Participants ranged in age from 6 through 12 years.

A multigated procedure was used to select eligible girls with ADHD and comparison girls (see Hinshaw, 2002, for complete details). Following telephone screening, the girls in the ADHD group met initial criteria for ADHD (five of nine items positive—i.e., at a level of 2 "pretty much" or 3 "very much" on the 0–3 metric) on the Swanson, Nolan, and Pelham Parent Inattention and Teacher Inattention scales (SNAP; Swanson, 1992) and a *t*-score of at least 60 (a cutoff validated by Chen, Faraone, Biederman, & Tsuang, 1994) on the Attention Problems subscale of the Child Behavior Checklist and Teacher Report Form (CBCL and TRF; Achenbach, 1991a, 1991b). These cutoffs were intentionally set "low" to prevent false negatives during the initial screening phase. However, for final study entry, participants were required to meet full criteria for diagnosis of ADHD, either Combined or Inattentive type, through the parent-administered Diagnostic Interview Schedule for Children—4th Edition (DISC-IV; see Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000). Eligibility for the comparison group required that girls were below cutoffs on all clinical scales and did not receive a diagnosis of ADHD on the DISC-IV (American Psychiatric Association, 1994). Of the 155 girls who met criteria for either the ADHD or comparison group and were invited to participate in the study, 149 accepted, and all 149 continued throughout the entire 5 week program. Nobody dropped out once the program began. For further details about participation rates in the full sample, please see Hinshaw (2002).

The major exclusion criteria for all participants were Verbal IQ below 70, gross neurological injury, psychosis, pervasive developmental disorder, seizure disorder, Tourette's disorder, severe obsessive–compulsive disorder, or severe physical impairment. Summer camp activities were chosen to elicit social behaviors across different domains that are part of the lives of most preadolescent children, with and without ADHD (e.g., P.E., art, reading class). Children with disorders that interfere with movement or those with IQ below 70 would have had great difficulty participating in such camp activities, given that staff and facility were not specially configured for accommodating such youth. Disorders like oppositional defiant disorder, conduct disorder, depression, some anxiety disorders, and learning disabilities were not among

Table I. Descriptive Statistics for ADHD and Comparison Groups

Variable	ADHD ($n = 91$)		Comparison ($n = 58$)		
	M	*SD*	*M*	*SD*	
Demographic					
Age (months)	115.55	20.61	113.67	20.08	$t(147) = 0.55$
Income[a]	6.29	2.62	6.28	2.59	$t(147) = 0.03$
Public assistance (any%)	0.14	0.35	0.14	0.35	$t(147) = 0.08$
Number of parents in household	1.62	0.49	1.71	0.46	$t(147) = -0.80$
Maternal education[b]	4.76	0.99	4.95	1.02	$t(147) = -1.13$
White (%)	51.65		46.55		$\chi^2(1) = 0.33$
Academic					
History of grade retention (%)	17.58		5.26		$\chi^2(1) = 4.86^*$
Behavioral					
CBCL Attention (*t* score)	74.33	8.72	52.05	4.71	$t(147) = 17.86^{**}$
CBCL Externalizing (*t* score)	64.84	10.12	45.74	8.85	$t(147) = 11.78^{**}$
CBCL Internalizing (*t* score)	60.45	10.24	48.24	12.04	$t(147) = 6.62^{**}$

[a]For total annual family income, 1 = <$10,000; 9 = >$75,000.
[b]For maternal education, 1 = less than 8th grade; 6 = advanced or professional degree.
$^*p < .05$. $^{**}p < .01$.

the exclusion criteria, given the moderate to high rates of overlap between ADHD and each of these conditions (Biederman, Newcorn, & Sprich, 1991; Jensen, Martin, & Cantwell, 1997; Jensen, Shervette, Xenakis, & Richters, 1993) and the desire to investigate a representative, yet not "super normal," comparison sample.

As shown in Table I, the ADHD and comparison samples were compared with respect to demographic, academic, and behavioral measures. The two groups did not differ significantly with respect to any demographic variable. As expected, however, girls with ADHD manifested higher levels of attention problems, externalizing behavior and internalizing behavior as reported by parents, and history of grade retention.

Measures

Daily Behavior Ratings (DBRs)

In order to yield the global rating measure of interest for this investigation, camp staff completed DBRs about each girl's externalizing behavior at the end of each day of the program, from Weeks 2–5 (see Hinshaw, 2002). DBR data were provided by 83 different camp staff, in total, over the two summers. Each of these staff members completed a DBR at the end of a day for which he or she had spent time with the girl. However, because all 83 staff did not work every day or spend time with the all the girls every day, the data for each girl represented an average of reports from 3–4 raters per day, amalgamated over the summer program. These DBR staff members represented college undergraduates, clinical psychology graduate students, and other staff at the B.A. level with clinical training. The externalizing DBR was an 11-item scale tapping aggression, defiance, and noncompliance with a 4-point Likert metric, with higher scores corresponding to greater externalizing behavior. Sample items included "was defiant," "was quarrelsome," and "had temper outbursts," Cronbach's α was .95 for this scale. The mean per-item, interrater reliability across the 11 items was .54 (across a random subsample of days of the program). Note that this figure might be lower than expected because of the different rater combinations across different periods and days. Children's scores on the externalizing DBR correlated .58 and .65 with mother report on the CBCL and teacher report on the TRF Externalizing Scales respectively, showing convergence with established measures.

Behavioral Observation (OBS)

Across the two summers, 52 trained undergraduate staff used a behavioral observation coding system for externalizing behavior (Hinshaw & Renfro, 1999). Teams of four trained observers coded 1-hr classroom and playground periods across the summer. Observers coded children for 5-s intervals, recording instances of physical aggression (e.g., hitting, kicking), verbal aggression (e.g., swearing, name-calling), and noncompliance (e.g., rule

breaking, ignoring orders, but short of aggression per se). Proportion scores were calculated representing the proportion of time the child was engaged in any of these behaviors compared to the total time the child was observed. Each child was observed at least 200 times spread over a total of 16–18 hr. Occurrence-only interrater reliability was acceptable for all categories (proportions = 0.6–0.7). Children's scores on the externalizing OBS correlated .55 and .56 with mother report on the CBCL and teacher report on the TRF Externalizing Scales respectively, also showing convergence with established measures.

Popularity With Adult Staff

Standard sociometric procedures (e.g., Coie, Dodge, & Coppotelli, 1982), typically used with peers as reporters of peer rejection, were utilized to assess this variable, with camp staff as reporters instead. Forty-eight raters nominated the three girls they personally liked most and the three girls they personally disliked most in a classroom. Raters viewed pictures of all the girls in a class to facilitate this process. Raters were those adult camp staff who had spent the most time with their respective classrooms (from 20 to 40 hr/week) and who were generally the most involved in the summer camp. Thus, raters were all graduate students and staff at the B.A. level with clinical training. Staff roles included team leaders (the adults who led the same class of girls all day throughout the summer); classroom, art, and drama teachers (the adults who taught the same class of girls 1–2 hr a day throughout the summer); and other staff who had spent a large amount of time with a particular group (discussion group facilitators, assistants to team leaders). All raters completed this measure at the end of the summer camp. However, in order to ensure that the ratings at the end of camp were reflective of the ratings during camp, a subset of four raters per class completed sociometrics during Week 1, Week 3, and Week 5 as well as at the end of camp. Correlations for Week 1–Week 5 were .50 for positive nominations and .68 for negative nominations. Additionally, correlations between Week 5 and the end-of-camp nomination scores were .63 for positive nominations and .75 for negative nominations.

Of note, none of the three groups of staff—DBR informants, OBS coders, or raters of the child's popularity with adults—were identical. DBR staff included the widest range: undergraduates, graduate students, B.A. level staff with clinical training. OBS coders were composed exclusively of trained undergraduate staff. Raters of the child's popularity with staff were exclusively graduate students or B.A. level staff with clinical training. Thus, none of the OBS coders was a popularity rater, but 25 of the 52 OBS coders and 11 of the 48 popularity with staff raters also filled out DBRs. This paper reports primary analyses using all staff providing each measure and secondary analyses restricted to overlapping staff.

Popularity With Peers

Using these same standard sociometric nomination procedures (e.g., Coie et al., 1982), in confidential interviews, all children in a given classroom nominated three classmates with whom they would most like to be friends and three classmates with whom they would least like to be friends. To facilitate this process, girls were shown large poster boards with names and pictures of their classmates. Because each class often had a different number of girls, proportion scores were calculated by dividing the number of nominations received by the number of classmates. In this way, each child had a proportion score for "most liked" nominations received as well as "most disliked" nominations received. Sociometric data were collected three times over the course of the summer camp: Week 1, Week 3, and Week 5. However, these analyses used the Week 5 data only because they were a more accurate reflection of the girls' peer rejection and because the stability of peer rejection was high (Week 1–Week 5 correlations were .51 for positive nominations and .85 for negative nominations; see Blachman & Hinshaw, 2002, for further details).

Data Reduction and Conceptualization

Because the predictors (popularity with adult staff and peers) and outcome measures (the discrepancy between the DBRs and OBS of externalizing behavior) were based on scales with different metrics, all of the key measures were standardized into *z* scores before performing analyses. Conceptually, the observation coding system (OBS) was considered the criterion from which to evaluate the veracity of the staff global ratings (DBRs). Because raw mathematical difference scores were potentially unreliable (Cohen, Cohen, West, & Aiken, 2003), standardized residual scores (SRS) were created from regression models in which OBS were used to predict DBRs, to measure the discrepancy between DBRs and OBS. Each SRS was the standardized mathematical difference between the actual values of DBRs and the predicted values of DBRs (derived from regression models with OBS as predictors) such that SRS = *z* scores of (actual DBRs – predicted

DBRs). For more detailed discussion of SRS, please see Chi and Hinshaw (2002). In the operational definition of "bias," a negative value indicated an underestimation and a positive value indicated an overestimation of child externalizing behavior from the DBRs. Finally, in order to prevent data being affected by a few extreme outliers (e.g., Cohen et al., 2003), for every measure, cases lying more than 3.5 *SD* beyond the mean were identified. The least extreme outlier was then replaced with a score that was 3.5 *SD* above the mean. In order to preserve the rank, the next greatest outlier was replaced with a score 3.75 *SD*s above the mean, and so on until all the outlying scores were changed. This transformation was only applied to 5 cases out of a total of 149 children.

Data Analytic Plan

For the central hypotheses, hierarchical multiple regressions were preformed to test the incremental contributions of the theoretically guided predictor variables. The criterion variable was the discrepancy between DBRs (global ratings) and OBS (behavior observations) of externalizing behavior, measured by SRS. All analyses were also conducted with raw difference scores as the criterion variable, and results were unchanged. In terms of predictors, diagnostic status (ADHD vs. comparison) was placed at step 1 and the proposed source of bias (e.g., negative "most disliked" nominations or positive "most liked" nominations from adult staff) at step 2. The first hypothesis (*over*estimation by staff of externalizing behavior for *disliked* children) would be confirmed if adult negative nominations predicted DBR–OBS discrepancies beyond ADHD status, with a positive beta weight. Similarly, the second hypothesis (*under*estimation by staff of externalizing behavior for *liked* children) would be confirmed if adult positive nominations predicted DBR–OBS discrepancies beyond ADHD status, with a negative beta weight. For the exploratory hypotheses, negative nominations and positive nominations from peers were placed on step 3 of the regressions and the interaction between popularity with peers and popularity with adults at step 4. In the presence of a significant interaction, results were probed in the manner suggested by Holmbeck (2002). Because participants with ADHD were in either the Inattentive or Combined types (see Hinshaw, 2002, for further details), additional analyses were conducted using subtype information (ADHD-Inattentive, ADHD-Combined, and comparison) as opposed to ADHD versus comparison status. Results were unchanged, and this study reports those results of the analyses using ADHD versus comparison status only.

RESULTS

Group Comparison

Group differences with respect to the predictor and criterion variables are shown in Table II. Although with ADHD displayed significantly more externalizing behavior than did comparison girls (regarding both the DBRs and the OBS), the *discrepancy* between DBRs and OBS did not differ between the groups. Girls with ADHD received significantly more negative nominations and significantly fewer positive nominations from adult staff and from peers than did comparison girls. This finding was parallel to the peer rejection literature (Blachman & Hinshaw, 2002; Erhardt & Hinshaw, 1994), suggesting that children with ADHD have compromised social relationships.

By taking a median split of the popularity with peers and the popularity with adult staff measures, ADHD and comparison girls were next classified into four categories: "doubly liked," "liked by peers and disliked by staff,"

Table II. Comparison of Girls With and Without ADHD on Relevant Variables

	ADHD ($n = 91$)		Comparison ($n = 58$)		
Variable	*M*	*SD*	*M*	*SD*	
Criterion measure[a]					
Discrepancy between DBRs and OBS	0.11	0.38	−0.18	1.20	$t(147) = 1.76$
Hypothesized predictors[a]					
Positive nominations from adults	−0.24	0.82	0.38	1.12	$t(147) = 3.91^{**}$
Negative nominations from adults	0.26	1.15	−0.41	0.50	$t(147) = -4.19^{**}$
Positive nominations from peers	−0.22	0.90	0.34	1.07	$t(147) = 3.44^{**}$
Negative nominations from peers	0.35	1.13	−0.55	0.24	$t(147) = -5.95^{**}$

[a]These numbers reflect *z* scores, where positive numbers correspond to greater levels of the construct.
$^{**}p < .01$.

Table III. Popularity With Peers and Staff for Girls With ADHD and Comparison Girls

Popularity classifications	Comparison ($n = 58$)	ADHD ($n = 91$)
Doubly disliked	11	67
Staff dislike & peer like	15	9
Staff like & peer dislike	8	10
Doubly liked	24	5

Note. $\chi^2(3, N = 149) = 49.50$, $p = .000$; $\phi^2 = 0.576$, $p = .000$.

"liked by staff and disliked by peers," and "doubly disliked" (see Table III). Girls with ADHD in our sample were more likely to be classified in the "doubly-disliked" category and less likely to be classified in "doubly-liked" category than would have been expected by chance. Conversely, the comparison girls were more likely to be classified in the "doubly-liked" category and less likely to be classified in the "doubly-disliked" category, $\chi^2(149) = 49.50$, $p < .001$; $\phi = 0.58$, $p < .001$. In fact, 86% of the girls in the "doubly-disliked" category had ADHD, and 83% of the girls in the "doubly-liked" category were comparison girls.

Correlations Between Relevant Variables

Zero-order correlations shown in Table IV revealed that that negative nominations from staff were positively correlated with the discrepancy scores at .31, signifying overestimation of externalizing behavior on DBRs (the global ratings) compared to OBS (the behavior observations). However, positive nominations from staff were uncorrelated with the discrepancy scores. Regarding nominations from peers, negative nominations were positively correlated at .41 with the discrepancy scores, and positive nominations were negatively correlated at −.21 with the discrepancy scores. Furthermore, negative nominations and positive nominations from staff, as well as from peers, were all found to be associated at modest levels (.19–.45).

Table IV. Correlations Among Study Variables

Construct	1	2	3	4	5
1. Discrepancy between DBRs and OBS	—	−.01	.31**	−.21*	.41**
2. Positive noms from adults		—	−.30**	.26**	−.19*
3. Negative noms from adults			—	−.20*	.45**
4. Positive noms from peers				—	−.44**
5. Negative noms from peers					—

* $p < .05$. ** $p < .01$.

Popularity With Staff as Predicting DBR–OBS Discrepancies

Tables V and VI present key analyses regarding popularity with adult staff as a predictor of discrepancies between DBRs and OBS. First, at step 1, ADHD status did not predict discrepancy scores; at step 2, negative nominations from staff did predict discrepancy scores (R^2 change = .08, $p < .001$), with a positive beta-weight (see Table V). Thus, unpopular girls' externalizing behavior was overestimated by DBRs compared to OBS. In Table VI, regarding girls who were popular with staff, ADHD status again failed to predict discrepancy scores at step 1, and positive nominations from staff also failed to predict the discrepancy (R^2 change = .00, $p > .05$). Thus, popular girls' externalizing behavior was *not* underestimated by DBRs compared to OBS. In sum, whereas DBR scores tended to overestimate the externalizing behavior of children who were disliked by staff, DBR scores were not underestimated for children who were liked by staff.

Popularity With Peers as Predicting DBR–OBS Discrepancies

Tables V and VI also include analyses regarding popularity with peers as a both a predictor and potential moderator variable. The predictor hypotheses were tested by adding an additional step to each regression equation: the child's negative nominations or positive nominations from peers. Negative nominations from peers added incremental prediction of variance in the discrepancy (R^2 change = .09, $p < .001$), with a positive beta-weight. Thus, in addition to negative nominations from staff, negative nominations from peers predicted overestimation of the child's externalizing behavior in DBRs versus OBS. Additionally, positive nominations from peers accounted for a small amount of variance in the discrepancy (R^2 change = .04, $p < .05$), suggesting that if a child were liked by peers, staff underestimated her externalizing behavior in DBRs versus OBS.

Finally, to test potential moderator effects of peer nominations, the interaction between nominations from staff and peers was included at the final step of each regression equation. Regarding negative nominations, the interaction was significant (R^2 change = .05, $p < .01$), so it was probed in the manner suggested by Holmbeck (2002). Results (see Fig. 1) suggested that the additive effect of being disliked by peers occurred mainly when the child was disliked by adult staff. Thus, if the child was disliked by staff, being disliked by peers predicted a large increase in the overestimation of externalizing behavior in

Table V. Predicting Discrepancy Between DBRs and OBS From Negative Nominations

Variable[a]	Total R^2	R^2 change	*B*	*SE B*	β
Step 1					
ADHD status	.02	.02	.29	.17	.14
Step 2					
Negative nominations from staff	.10	.08**	.30**	.08	.30**
Step 3					
Negative nominations from peers	.19	.09**	.36**	.09	.36**
Step 4					
Interaction between 2 and 3	.24	.05**	.21**	.07	.28**

[a] Dependent variable: discrepancy between DBRs and OBS.
** $p < .01$.

DBRs versus OBS. By contrast, if the child was not disliked by staff, then her peer status made little difference in predicting the discrepancy. The interaction between positive nominations from staff and positive nominations from peers was not significant (R^2 change $= .00$, $p > .05$).

As highlighted above, there was not complete overlap between the staff who (a) filled out the DBRs, (b) conducted the OBS, and (c) rated the child's popularity with adult staff. Thus, the discrepancy scores between DBRs and OBS, as used above, reflected differences among raters as well as differences among rating systems. In an attempt to control for the confound of rater differences, core analyses were reconducted in two ways, first restricting the data set to that yielded by the staff who conducted both the OBS and DBRs (representing 2,575 of 8,796 total DBR data points, or 29%) and then restricting it to staff who conducted both the popularity with adult staff measure and DBRs (representing 1,914 of the 8,796 total DBR data points, or 22%). Note that DBR data from these two, completely nonoverlapping, subsets correlated .90 with one another. When the primary regression analyses were reconducted with both restricted data sets, the core findings remained unchanged: diagnostic status failed to make a significant contribution at step 1, and negative nominations (but not positive nominations) from staff made a significant (and positive) contribution at step 2, signifying again that staff overestimated the externalizing behavior of unpopular children in ratings versus observations. The contribution of peer status changed slightly: although negative nominations from peers continued to make a significant contribution on step 3, positive nominations from peers were no longer significant in either restricted subset of data. The interaction between negative nominations from staff and negative nominations from peers remained significant, and the interaction between positive nominations from staff and peers remained nonsignificant. Thus, in both restricted data sets, core results regarding unpopularity with staff and unpopularity with peers were entirely preserved; but the previously-found positive bias related to peer popularity was not replicated.

DISCUSSION

Findings from this study suggest that staff's global ratings of child externalizing behavior may be susceptible to the target child's popularity with adult staff. Specifically, with respect to children who are unpopular

Table VI. Predicting Discrepancy Between DBRs and OBS From Positive Nominations

Variable[a]	Total R^2	R^2 change	*B*	*SE B*	β
Step 1					
ADHD status	.02	.02	.29	.17	.14
Step 2					
Positive nominations from staff	.02	.00	.04	.09	.04
Step 3					
Positive nominations from peers	.06	.04*	−.20*	.09	−.20*
Step 4					
Interaction between 2 and 3	.06	.00	.04	.09	.03

[a] Dependent variable: discrepancy between DBRs and OBS.
* $p < .05$.

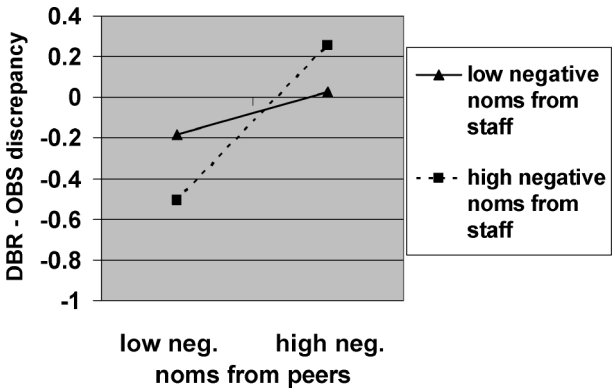

Fig. 1. Interactive effects of peer and staff negative nominations.

with staff, ratings appeared to overestimate the child's externalizing behavior in contrast to more objective behavior observations. Yet no favorable bias on rating scales was found for children who were particularly popular with staff. Peer status supplemented these predictions such that adult raters further overestimated the externalizing behavior of children who were unpopular with peers. On the other hand, popularity with peers attenuated rating bias only in the full data set and not in the data sets restricted to staff who overlapped in ratings and observations or ratings and popularity nominations. Negative peer status also served as a moderator, such that the association between unpopularity with staff and bias was stronger for children disliked by peers than for children not disliked by peers. No moderating effect occurred for being popular with peers.

What are the mechanisms underlying this apparent bias found in rating scales? First, dislike of the child by an adult may create a negative halo around that child such that the informant becomes more willing to interpret ambiguous behavior as psychopathology. Additionally, informants may have an easier time remembering negative behavior of children whom they (or other staff) dislike, because this is stereotype-congruent information. Indeed, people attend to and remember information that is congruent with their preexisting stereotypes and are more likely to discount information that does not support their stereotype (Johnston & Macrae, 1994). Similar to cognitive processes traditionally described as pertinent to aggressive and rejected children (Dodge, 1980), adults may have hostile attribution biases towards children whom they do not like—assuming that the child's ambiguous behavior is hostile in intent. This pattern of negative cognitive biases is similar to findings from the peer rejection literature revealing that peers have a large number of prejudices toward interpreting, and remembering, the behavior of classmates whom they dislike (Hymel et al., 1990). In sum, further research is needed to investigate the underlying mechanisms responsible for this effect of popularity on ratings.

Findings also suggest that although raters overestimated the externalizing behavior of children who were unpopular with staff, they were not favorably biased when evaluating children who were popular with staff. In other words, being popular with staff did not create a positive halo effect around the child's behavior, but being unpopular led to a negative bias. Again, this finding was similar to the peer rejection literature, which suggests that negative nominations from peers are more stable over time, as well as more significant predictors of future adjustment, than are positive nominations. Being peer rejected clearly carries negative consequences relative to being of average status, but being popular does not consistently add benefit beyond being average (e.g., see McDougall, Hymel, Vaillancourt, & Mercer, 2001; Putallaz & Dunn, 1990). Likewise, negative ascriptions about children from adults may have a more powerful biasing effect than do positive ascriptions. Supporting this idea, a teacher's negative expectation about a child's academic ability creates a self-fulfilling prophecy leading to decreased child performance, but a teacher's positive expectation is less likely to lead to increased child performance (Alvidrez & Weinstein, 1999; McKown & Weinstein, 2002). However, this pattern could have occurred for psychometric reasons, in that positive nominations from staff were less stable over the summer than were negative nominations (r for Week 1–Week 5 = .50 for positive nominations, .68 for negative nominations). This difference in stability was similar to the pattern from the peer sociometrics. Because the DBR data were collected continuously over the course of the summer, fluctuation in the children who were most popular at any given time could have diluted any positive halo effects.

Importantly, no evidence was found that staff ratings were biased by the girls' ADHD status. This finding was surprising, because other research suggests that an expectation that a child has ADHD can bias perceptions negatively (e.g., Harris et al., 1998). Perhaps this lack of effect was related to having kept raters unaware of diagnostic status throughout the programs. Additionally, camp staff was carefully screened for clinical skills and sensitivity towards children with ADHD, and they may have constituted a somewhat atypical population of adults.

Results have implications for methodology in developmental psychopathology research. Despite staff training in which the importance of objectivity and lack of emotional involvement in completing the global ratings were emphasized, it appeared that general unpopularity of the child served to bias her raters' perceptions of externalizing

behavior. Although observation coding systems are often expensive and impractical to implement, these results point to the importance of including, when feasible, observations or other molecular measures of behavior over short periods of time, as these may be more objective than global behavior ratings. Results also showed that in addition to the myriad problems associated with social rejection of children (Parker & Asher, 1987), being disliked by adults may have negative consequences as well. The "dislike-bias" shown in this study raises a potentially troubling possibility that a certain percentage of "behavioral disturbance" of rejected children may stem from a combination of actual negative behaviors, bias on the part of adult informants, and reciprocal cycle of self-fulfilling prophecies. Indeed, a substantial body of research suggests that negative expectations can actually elicit stereotype-congruent behavior. When naïve children were placed with peers who expected them to show externalizing behavior versus peers who had no such expectation, the target children conformed and did in fact show more externalizing behavior (Harris et al., 1998; Olson, 1992). Furthermore, it has been shown that adults who expected to interact with a child who had ADHD (as opposed to a child without ADHD) ingested more alcohol as a way to "deaden" their expected stress (e.g., Pelham et al., 1997). Thus, results from this study showed the effects of expectancies on the validity of clinical data.

One limitation of this study was that staff who completed the DBRs overlapped with, but were not identical to, staff who reported the child's popularity with adults or the staff who conducted the OBS. Thus, it was possible that the measure of the child's popularity did not adequately reflect the feelings of the staff who completed the DBRs. Nonetheless, reanalyses indicated that the main results remained when the samples were restricted to overlapping staff, particularly for negative appraisals from staff and peers. Additionally, the lack of full overlap would actually reduce the power of the study, to the extent that the measure of popularity with adults might not perfectly reflect the feelings of the DBR raters, diluting potential effects of popularity with adults in biasing DBR ratings. Thus, findings may in fact have *underestimated* how much staff's personal feelings about the child biased ratings. An additional limitation was that the DBRs and OBS may not have measured identical constructs. The global nature of the 4-point DBR scale required raters to judge externalizing behavior on the basis of both frequency and intensity. By contrast, raters in the OBS system measured only frequency of externalizing behavior, coding 1 if *present* and 0 if *absent*, irrespective of intensity. Thus, results may have partially reflected differences in the construct being measured as well as differences in raters' perceptions.

To conclude, this study represented an important step in examining the influence of adult-rated popularity on behavior ratings of research staff working with child clinical samples. Because assessment of a child's behavior can result in labeling, psychiatric diagnosis, and access or denial of services, biases in assessment yield powerful consequences. A child who is disliked by adults may be more likely to receive diagnoses of Oppositional Defiant Disorder and Conduct Disorder partially from exaggerated adult ratings of externalizing behavior. These diagnoses may result in stigmatization and association with deviant peer networks, creating a self-fulfilling prophecy that leads to delinquency and disengagement from school (e.g., Dishion, McCord, & Poulin, 1999). Hence, investigators as well as clinicians who work with children must consider potential sources of bias in describing children's behaviors and psychopathology, such as the raters' personal feelings about the target child. Results affirm the high importance of thoroughly assessing child psychopathology with multiple informants, so as to lessen bias from any one adult's liking of the child, as well as supplementing rating scales with more objective observational measures when possible.

ACKNOWLEDGMENTS

Work on this research was supported by National Institute of Mental Health Grant R01 MH45064. We express our great appreciation to the girls and parents who participated in our summer research programs and to the many staff who assisted with data collection and data coding, whose work made this study possible. We would also like to thank Elizabeth Owens, David Sobel, and Brian Zupan for their helpful consultation on this study.

REFERENCES

Achenbach, T. M. (1991a). *Manual for Child Behavior Checklist and Revised Child Behavior Profile*. Burlington, VT: University Associates in Psychiatry.

Achenbach, T. M. (1991b). *Manual for Teacher's Report Form and 1991 Profile*. Burlington, VT: University Associates in Psychiatry.

Alvidrez, J. L., & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology, 91,* 731–746.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

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

Biederman, J., Newcorn, J., & Sprich, S. (1991). Comorbidity of attention-deficit hyperactivity disorder with conduct, depressive, anxiety, and other disorders. *American Journal of Psychiatry, 148,* 564–577.

Blachman, D. R., & Hinshaw, S. P. (2002). Patterns of friendship among girls with and without attention-deficit/hyperactivity disorder. *Journal of Abnormal Child Psychology, 30,* 625–640.

Cairns, R. B., & Green, J. A. (1979). How to assess personality and social patterns: Observations or ratings? In R. B. Cairns (Ed.), *The analysis of social interactions* (pp. 209–226). Hillsdale, NJ: Erlbaum.

Chen, W. J., Faraone, S. V., Biederman, J., & Tsuang, M. T. (1994). Diagnostic accuracy of the Child Behavior Checklist Scales for attention-deficit hyperactivity disorder: A receiver-operating characteristic analysis. *Journal of Consulting and Clinical Psychology, 62,* 1017–1025.

Chi, T. C., & Hinshaw, S. P. (2002). Mother–child relationships of children with ADHD: The role of maternal depressive symptoms and depression-related distortions. *Journal of Abnormal Child Psychology, 30,* 387–400.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. A. (2003). *Applied multiple regression/correlation analyses for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Coie, J. D., Dodge, K. A., & Coppotelli, H. (1982). Dimensions and types of social status: A cross-age perspective. *Developmental Psychology, 18,* 557–570.

Dishion, T. J., McCord, J., & Poulin, F. (1999). When interventions harm: Peer groups and problem behavior. *American Psychologist, 54,* 755–764.

Dodge, K. A. (1980). Social cognition and children's aggressive behavior. *Child Development, 51,* 162–170.

Eaton, W. O., & Enns, L. R. (1986). Sex differences in human motor activity level. *Psychological Bulletin, 100,* 19–28.

Erhardt, D., & Hinshaw, S. P. (1994). Initial sociometric impressions of attention-deficit hyperactivity disorder and comparison boys: Predictions from social behaviors and from nonbehavioral variables. *Journal of Consulting and Clinical Psychology, 62,* 833–842.

Funder, D. C., & Colvin, C. R. (1991). Explorations in behavioral consistency: Properties of persons, situations, and behaviors. *Journal of Personality and Social Psychology, 60,* 773–794.

Harris, M. J., Milich, R., & McAninch, C. B. (1998). When stigma becomes self-fulfilling prophecy: Expectency effects and the causes, consequences, and treatment of peer rejection. In J. Brophy (Ed.), *Advances in research on teaching* (243–272). Greenwich, CT: JAI Press.

Hart, E. L., & Lahey, B. B. (1999). General child behavior rating scales. In D. Shaffer, C. P. Lucas, & J. E. Richters (Eds.), *Diagnostic assessment in child and adolescent psychopathology* (pp. 65–90). New York: Guilford Press.

Hinshaw, S. P. (2002). Preadolescent girls with attention-deficit/hyperactivity disorder: I. Background characteristics, comorbidity, cognitive and social functioning, and parenting practices. *Journal of Consulting and Clinical Psychology, 70,* 1086–1098.

Hinshaw, S. P., Carte, E. T., Sami, N., Treuting, J. J., & Zupan, B. A. (2002). Preadolescent girls with attention-deficit/hyperactivity disorder: II. Neuropsychological performance in relation to subtypes and individual classification. *Journal of Consulting and Clinical Psychology, 70,* 1099–1111.

Hinshaw, S. P., & Nigg, J. T. (1999). Behavior rating scales in the assessment of disruptive behavior in childhood. In D. Shaffer & P. T. Lucas (Eds.), *Diagnostic assessment in child and adolescent psychopatholgy* (pp. 91–126). New York: Guilford Press.

Hinshaw, S. P., & Renfro, R. (1999). *Social behavior observation system*. Unpublished manuscript, University of California, Berkeley, CA.

Holmbeck, G. N. (2002). Post-hoc probing of significant moderational and mediational effects in studies of pediatric populations. *Journal of Pediatric Psychology, 27,* 87–96.

Hymel, S., Wagner, E., & Butler, L. J. (1990). Reputational bias: View from the peer group. In S. R. Asher & J. D. Coie (Eds.), *Peer rejection in childhood* (pp. 156–186). New York: Cambridge University Press.

Jensen, P. S., Martin, D., & Cantwell, D. P. (1997). Comorbidity in ADHD: Implications for research practice, and, DSM-IV. *Journal of the American Academy of Child and Adolescent Psychiatry, 36,* 1065–1079.

Jensen, P. S., Shervette, R. E., Xenakis, S. N., & Richters, J. E. (1993). Anxiety and depressive disorders in attention deficit disorder with hyperactivity: New findings. *American Journal of Psychiatry, 150,* 1203–1209.

Johnston, L. C., & Macrae, C. N. (1994). Changing social stereotypes: The case of the information seeker. *European Journal of Social Psychology, 24,* 581–592.

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

Koslin, B. L., Harlow, R. N., Karlins, M., & Pargament, R. (1968). Predicting group status from members' cognitions. *Sociometry, 31,* 64–75.

McDougall, P., Hymel, S., Vaillancourt, T., & Mercer, L. (2001). The consequences of childhood peer rejection. In M. R. Leary (Ed.), *Interpersonal rejection*. (pp. 213–250). New York: Oxford University Press.

McKown, C., & Weinstein, R. S. (2002). Modeling the role of child ethnicity and gender in children's differential response to teacher expectations. *Journal of Applied Social Psychology, 32,* 159–184.

Michelson, L., Mannarino, A., Marchione, K., Kazdin, A. E., & Costello, A. (1985). Expectancy bias in behavioral observations of therapeutic outcome: An experimental analysis of treatment and halo effects. *Behavior Research and Therapy, 23,* 407–414.

Mikami, A. Y., & Hinshaw, S. P. (2003). Buffers of peer rejection among girls with and without ADHD: The role of popularity with adults and solitary play. *Journal of Abnormal Child Psychology, 31,* 381–397.

Miller, M. D., & Ottinger, D. R. (1986). Influence of labeling on ratings of infant behavior: A prematurity prejudice. *Journal of Pediatric Psychology, 11,* 1986.

Moskowitz, D. S. (1986). Comparison of self-reports, reports by knowledgeable informants, and behavioral observation data. *Journal of Personality, 54,* 294–317.

Moskowitz, D. S. (1990). Convergence of self-reports and independent observers: Dominance and friendliness. *Journal of Personality and Social Psychology, 58,* 1096–1106.

Olson, S. L. (1992). Development of conduct problems and peer rejection in preschool children: A social systems analysis. *Journal of Abnormal Child Psychology, 20,* 327–350.

O'Sullivan, C. S., & Durso, F. T. (1984). Effect of schema-incongruent information on memory for stereotypical attributes. *Journal of Personality and Social Psychology, 47,* 55–70.

Parker, J. G., & Asher, S. R. (1987). Peer relations and later personal adjustment: Are low-accepted children at risk? *Psychological Bulletin, 102,* 357–389.

Pelham, W. E., Lang, A. R., Atkeson, B., Murphy, D. A., Gnagy, E. M., & Greiner, A. R., et al. (1997). Effects of deviant child behavior on parental distress and alcohol consumption in laboratory interactions. *Journal of Abnormal Child Psychology, 25,* 413–424.

Puig, M., Lambert, M. C., Rowan, G. T., Winfrey, T., Lyubansky, M., & Hannah, S. D., et al. (1999). Behavioral and emotional problems among Jamaican and African American children, ages 6–11: Teacher reports versus direct observations. *Journal of Emotional and Behavioral Disorders, 7,* 240–250.

Putallaz, M., & Dunn, S. E. (1990). The importance of peer relations. In M. Lewis & S. M. Miller (Eds.), *Handbook of developmental psychopathology* (pp. 227–236). New York: Plenum.

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

Ritter, J. M., & Langlois, J. H. (1988). The role of physical attractiveness in the observation of adult–child interactions: Eye of the beholder or behavioral reality? *Developmental Psychology, 24,* 254–263.

Shaffer, D., Fisher, P., Lucas, C. P., Dulcan, M. K., & Schwab-Stone, M. E. (2000). NIMH Diagnostic Interview Schedule for Children, Version, IV: Description, differences from previous versions, and reliability of some common diagnoses. *Journal of the American Academy of Child and Adolescent Psychiatry, 39,* 28–38.

Swanson, J. M. (1992). *Assessment and treatment of ADD students*. Irvine, CA: K.C. Press.