Published in final edited form as:
Psychol Sch. 2010 April ; 47(4): 374–390. doi:10.1002/pits.20476.

Child and Informant Influences on Behavioral Ratings of Preschool Children

Beth M. Phillips and
Florida State University, Florida Center for Reading Research, 2010 Levy Avenue Suite 100, Tallahassee FL 32310, bphillips@fcrr.org

Christopher J. Lonigan
Florida State University, Florida Center for Reading Research, 2010 Levy Avenue Suite 100, Tallahassee FL 32310, lonigan@psy.fsu.edu

Abstract

This study investigated relations among teacher, parent, and observer behavioral ratings of 3- and 4-year-old children using intraclass correlations and ANOVA. Comparisons within and across children from middle- (*N* =166; Mean age 54.25 months, *SD* = 8.74) and low-income (*N* =199; Mean age 51.21 months, *SD* = 7.22) backgrounds revealed significant agreement between the raters but also considerable differences in both ranking and absolute scores between raters. Teachers and parents consistently rated children from low-income classrooms as having more behavioral problems and fewer prosocial behaviors. Results are conceptualized with respect to how differential expectations, comparison groups, and types of interaction with children can affect the evaluation of child behavior. Results point to the need for multiple sources of evaluation when assessing a child for behavioral difficulties, particularly in children from lower income backgrounds.

Keywords

behavior; ratings; preschool; teachers; agreement

Behavioral problems in the preschool age group recently have received a considerable amount of attention. Research indicates that the problematic behavior of children in preschool is predictive of later behavior problems, and in some instances even psychopathology in the school-age years (e.g., Campbell, Pierce, March, Ewing, & Szumowski, 1994; Heller, Baker, Henker, & Hinshaw, 1996; Kerr, Lunkenheimer, & Olson, 2007; Miner & Clarke-Steward, 2008). Typical problematic behaviors seen in preschool children include hyperactivity, impulsivity, inattention, and compliance difficulties (Campbell et al., 1994; Olson & Hoza, 1993); a more severe subgroup also may show developmentally inappropriate levels of aggression, poor self-regulatory skills and problematic peer relationships (Bates, Pettit, Dodge, & Ridge, 1998).

The identification of problem behaviors in the preschool period is complicated by the fact that many of the behaviors later seen as symptomatic of an externalizing disorder are normative and, in some cases, developmentally appropriate in this age group (Campbell, Pierce, Moore, Marakovitz, & Newby, 1996; Wakschlag et al., 2005). Some children will

display noncompliant behavior toward their parents as they test limits and learn to function autonomously. Aggression toward peers may occur as children learn how to share and resolve interpersonal conflicts and to master the significant transitions of entering childcare and beginning school. Whereas most children develop self-regulatory and social skills and show a decline in behavioral issues over time (Miner & Clarke-Stewart, 2008), a subgroup experiences substantial difficulties in the preschool period and beyond and frequently are diagnosed with behavioral disorders (Campbell et al., 1996; Miller-Lewis et al., 2006; Shaw, Gilliom, Ingoldsby, & Nagin, 2003).

Agreement Among Raters

The complexities surrounding identification of young children with behavioral problems necessitate the availability of reliable and valid assessment tools. Disruptive behaviors in children currently are assessed in many different ways, including rating scales (Nolan & Gadow, 1994; Schachar, Sandberg, & Rutter, 1986) observational coding (Milich & Landau, 1988; Tryon & Pinto, 1994), and formal diagnostic instruments (e.g., Achenbach & Edelbrock, 1983). Further, whereas various assessment protocols rely on the distinct perspectives of parents, teachers, or classroom observers, diagnostic systems like the Diagnostic and Statistical Manual (DSM-IV-TR, American Psychiatric Association, 2000) assume compatibility between informants by requiring evidence of problems in multiple contexts.

Mixed findings exist as to the strength of the relation between multiple types of assessments and/or different informants including teachers, parents, self-report, observer ratings, and activity coding (e.g., Achenbach, McConaughy, & Howell, 1987; Danforth & DuPaul, 1996; Nolan & Gadow, 1994; Tryon & Pinto, 1994; Reid & Maag, 1994). The seminal meta-analysis by Achenbach et al. (1987) indicated modest agreement between raters, on average. Likewise, the more recent review by De Los Reyes and Kazdin (2005) indicates that across multiple combinations and age ranges informant agreement for hyperactivity and inattention is moderate at best. For preschoolers, the challenge of obtaining reliable self-reports means the focus is typically on teacher and parent reports. Recently, Gross, Fogg, Garvey, and Julion (2004) reported a correlation of just .17 between parent and teacher ratings of two- to four-year olds, and Kerr et al. (2007) found that mother's ratings of three-year-olds' externalizing behaviors to be uncorrelated with teacher ratings. In general, informants with the same role (i.e., two teachers, mother and father; Duhig, Renk, Epstein, & Phares, 2000; Grietens et al. 2004), and with the same purpose in completing the rating (De Los Reyes & Kazdin, 2005), are those most likely to be in agreement as to the presence and severity of behavioral problems. As highlighted by Grietens et al. (2004), whereas there is general consensus that discrepancies between informants does not necessarily imply flaws with the measures themselves, some threats to the validity of various informants' ratings do exist. Moreover, which factors promote or suppress such agreement remains a pressing theoretical and practical issue.

Potential Sources of Rater Bias or Distortion

There is evidence that teacher ratings can be affected by potentially significant threats to validity. Teacher ratings can be altered by the order in which they complete ratings (Brandon, Kehle, Jenson, & Clark, 1990) and by the instructions received before observing and rating various children (Shuller & McNamara, 1976). Teacher and parent ratings also may be significantly influenced by the context in which the observations are conducted and individual raters often need to see children in the same contexts to get agreement on their ratings (e.g., Nolan & Gadow, 1994; Zentall, 1984). At the same time, more valid ratings result from observations conducted in multiple settings, suggesting that all raters ideally should observe children in several common environments (Milich & Landau, 1988). Further, teacher ratings of hyperactivity and inattention can be significantly increased when the child demonstrates aggressive or defiant behavior. Schachar et al. (1986) found that children's scores on the Conner's Teacher Rating Scale (CTRS) hyperactivity and inattention scales were higher in the presence of concurrent aggression and disobedience regardless of whether the children's hyperactivity or inattentiveness actually was more frequent or extreme when measured by direct observation. Likewise, Abikoff, Courtney, Pelham, and Koplewicz (1993) found unidirectional halo effects of aggressive and noncompliant behavior on ratings of hyperactivity and inattention.

Both teacher and parent ratings also are affected by the attributions and expectations that the informants have about the intentionality and stability of various disruptive behaviors (De Los Reyes & Kazdin, 2005; Lovejoy, 1996; Mills & Rubin, 1990). It may be that the increased salience of aggressive and noncompliant behaviors, and the subsequent influence these have on other ratings is a result, in part, of the more active negative response elicited from the teachers and parents when these behaviors occur (Chang & Sue, 2003). In this vein, Atkins, Pelham, and Licht (1989) reported that it was the presence of aggressive behaviors in their sample of hyperactive children, as opposed to behaviors that loaded on an inattention-overactivity factor that permitted hyperactive children to be distinguished from controls.

Another significant issue is the reference group to which each child is compared. As evidenced by modest relations between different sources, behavioral ratings are a subjective exercise. Each child is compared, often pursuant to explicit instructions, to other children with whom the rater has had similar contact (Kerr et al., 2007). Feil, Severson, & Walker (1995) noted that comparison of a child to peers plays a critical role in the identification of preschool behavior problems. In their study, an extreme rating in relation to the peer group was one of a series of important screening mechanisms. Parents will use their child's siblings, relatives, and frequent playmates as their reference group and teachers rely upon their experience with the other children in their current and previous classes. As Conners (1989) remarked, teachers' judgments are required because there is no "teacher-meter" to provide objective data. However, such ratings are cognitive summaries heavily dependent upon context, given that teachers are confounded with their classroom and school environments. As such, the behaviors of children within these environments are the referent to which a new child's behavior is compared.

Relation to SES and Ethnicity

Higher rates of disruptive behavior problems are reported among children from families of lower socioeconomic status (SES) than among children from more affluent families, particularly in the context of lower parental education and harsh parenting (e.g., Ackerman, Brown, & Izard, 2003; McDermott & Schafer, 1996). SES has also been significantly associated with an increase in behavior problems and with the likelihood of an ADHD, Conduct Disorder, or ODD diagnosis (e.g., Achenbach & Howell, 1993; Duncan, Brooks-Gunn, & Klebanov, 1994). Consistent with these patterns, Dawkins, Fullilove, and Dawkins (1995) found that inner-city Head Start children had significantly more and more frequent behavior problems than did a normative sample, although still somewhat fewer than a clinic-referred group of preschool children. Similarly, in Heller et al. (1996) SES in preschool uniquely predicted first grade teachers' ratings of externalizing behaviors over and above the contribution of initial behavior, parenting, and maternal stress. Further, the adversity associated with limited resources is a significant risk factor in the longitudinal stability of problem behaviors in early childhood (Biederman et al., 1995; Deater-Decker, Dodge, Bates, & Pettit, 1998; McGee, Partridge, Williams, & Silva, 1991).

Whereas in some cases the absolute frequency of behavioral problems is higher in African American and Latino children, these group differences are usually eliminated once SES is taken into consideration (Deater-Decker et al., 1998; Dodge, Pettit, & Bates, 1994; McDermott & Schafer, 1996). However, a number of studies also indicate that African American children are rated as displaying more behavioral problems on teacher-report scales (e.g., Epstein et al., 2005; Piggott & Cowen, 2000; Reid et al., 1998). Moreover, because many minority families live at or near the poverty level, the direct and indirect effects of SES on child behavior may impact them disproportionately (Koblinksy, Kuvalanka, & Randolph, 2006).

Possible Teacher Bias in Ratings

Mixed evidence supports the theories that incongruence between teacher and child ethnicity, or bias against minority children leads to more reports of behavioral problems. Studies have found that non-Hispanic Caucasian teachers rated African American children as higher on ADHD-like behaviors (Downey & Pribesh, 2004; Puig et al., 1999; Sonuga – Barke et al., 1993). Saft and Pianta (2001) also found that teachers rated their relationship with young students as closer and having less conflict when they shared an ethnic background. However, not all recent studies have produced findings consistent with an interpretation of bias or of incongruence in ethnic background. Whereas Piggott & Cowen (2000) found that teachers gave African American children higher ratings of behavioral problems, this was found in ratings by both African American and Caucasian teachers, with no interaction effects. In a study comparing Hispanic and non-Hispanic Caucasian teachers' ratings of both types of children, De Ramirez & Shapiro (2005) found a main effect for teacher ethnicity, such that Hispanic teachers rated children higher, but no effect for child ethnicity, and no interaction. Similarly, Hosterman, Du Paul, and Jitendra (2008) did not find teacher bias against African American or Hispanic students when their ratings were compared to

observations. Notably, all but one of these studies (Saft & Pianta, 2001) has been conducted with students in kindergarten or older.

The Present Study

Given the significant risk for long-term morbidity and the number of secondary disruptions in adaptive functioning related to early behavioral problems, especially in children from less advantaged backgrounds, early identification of children with early behavioral problems seems to be in their best interest. Intervention may be most effective at the preschool age, before behavioral patterns become entrenched. Questions arise as to how best to identify children at risk for behavioral disorders. Given the research on the susceptibility of teacher and parent ratings to biasing and reference group effects, one source of ratings alone may not be optimal for assessing preschool behavior problems, especially in low-income populations. Moreover, much less is known about agreement between raters in low-income, particularly minority preschool populations, because the preponderance of research within the United States has been conducted with more affluent, and predominantly Caucasian populations (de Ramirez & Shapiro, 2005; McDermott, & Schaefer, 1996). Of two recent exceptions, (e.g., Hosterman et al., 2008; Milfort & Greenfield, 2002), one was conducted with elementary age children and neither included parental ratings. There are significant gaps in the literature about the reliability and validity of ratings of low-income preschool children's behavior. Further, questions about teacher bias against ethnic minorities can be explored in this study, in which, unlike most prior studies including minority populations, virtually all children were evaluated by at least one teacher matched in ethnic background.

This study was designed to address some of these questions about the assessment of problem behaviors in the preschool years, particularly in children from traditionally underrepresented groups. To our knowledge, this study is just the second investigation with preschool children in which teachers, parents, and direct observation provided behavioral ratings (Kerr et al., 2007), and one of relatively few studies of rater agreement with preschool children from lower-income and ethnically diverse backrounds (e.g., Milfort & Greenfield, 2002). Of primary interest was the comparison of teacher-, parent-, and trained observer-ratings of disruptive and inattentive behaviors. Prior research suggests that these informants view children from unique perspectives and make their subjective ratings with access to reference groups of varying heterogeneity and size. These three different sources of ratings thus constituted three methods of assessment whose covariance and agreement were evaluated, and compared, for groups of children representing middle- and lower-income backgrounds.

Two specific hypotheses consistent with prior research and emerging from this methodology were evaluated: (a) The agreement between teacher and observer ratings will be stronger than that between either teacher or observer and the parent ratings, in both income groups; this would stem from a shared classroom context and opportunity to observe multiple same-age children, rather than being relatively restricted as parents to a much smaller reference group, and (b) In keeping with prior research, mean ratings will be higher for lower income, predominantly minority children than middle income, predominantly Caucasian children.

Method

Participants

Middle-Income—Classrooms were eligible for inclusion in the middle-income group if they were not Head Start centers or centers predominantly serving children receiving subsidized child care vouchers. Children in the middle-income (MI) sample were recruited for participation in a longitudinal project assessing emergent literacy skills in preschool children. In year 1 children attended 11 classrooms within two faith-based, morning only fee-for-service preschool centers. At least one observer, who accumulated at least 4.5 hours of observation time, and at least one teacher provided data on 85 children, representing approximately 52% of the children in classrooms serving 3- to 5-year-old children. Parents completed forms for 71 of these children, reflecting a return rate of 84%. In year 2, children attended one of the same preschools from year 1 plus several other eligible private preschool centers. At least one observer and one or two teachers provided data for 128 children, representing well over half of the eligible children in these centers. Parents completed questionnaires for 96 of these children, reflecting a return rate of 75%. Data from the two years were aggregated into a final sample of 166, after excluding year 1 data for a child for whom year 2 data also was available. Children ranged in age from 32 to 76 months (M = 54.25; SD = 8.74) and 47% were girls. Virtually all children were Caucasian (i.e., 95.8%); 1.2% were African American, and 3% represented other ethnicities. Almost all (i.e., over 85%) teachers in these MI classrooms were Caucasian.

Low-income—Classrooms were eligible for inclusion in the lower-income group by virtue of being Head Start centers that by definition served a population of children from lower-income backgrounds. Children were recruited from six classrooms in two Head Start centers during year 1, and from these two and four other Head Start centers, representing 15 classrooms, during year 2. Parallel to the MI centers, the overwhelming majority (i.e., over 85%) of the teachers in the LI classrooms were African American, thus matching most of their students. At least one observer and at least one teacher provided data for 92 children in year 1, which represented approximately 85% of the children enrolled in each classroom. Parent questionnaires were completed for 74 of these children, reflecting a return rate of 80%. Four children were excluded from the sample because at least one observer did not achieve 4.5 hours of total observation time. At least one observer and at least one teacher provided data for 144 children in year 2. Parent questionnaires were completed for 129 of the children, reflecting a return rate of 90%. Data from the two years were again aggregated into a final sample of 199. Children in this sample ranged in age from 35 to 67 months (M = 51.21; SD = 7.22), and included approximately 54% girls. The children were 92% African American, 4.5% Caucasian, and 3.5% other ethnic groups. Chi-square analyses indicated no significant difference in the proportion of each sex between the two groups. Analysis of the ethnic composition of the two samples indicated that they were significantly different, $\chi^2(2, 364) = 310.91$, $p < .001$. Children in the LI sample were significantly, although not substantially younger than children in the MI sample, $F(1, 364) = 13.25$, $p < .001$.

Measures

Conners Teacher Rating Scale (CTRS)—A modified version of the CTRS-28 (Conners, 1989) was used in this study. The CTRS, as opposed to the Conners Parent Rating Scale (CPRS), was used for the parents and the observers to facilitate direct comparisons of ratings and because the parent scale covers somatic and anxiety items not of direct relevance for this study. Fifteen of the 23 CTRS items also appear on the CPRS. Four items, which do not load onto any of the three factors (e.g., Conduct Problems, Hyperactivity, and Inattentive) and are not used in scoring any of the three subscales or the total score, were removed to lessen the time for measure completion for raters, particularly teachers completing multiple packets. These items include “submissive toward authority,” “no sense of fair play,” “appears to be unaccepted by group’” and “does not get along well with classmates.” Additionally, two items measuring uncooperative behavior, “uncooperative with classmates” and “uncooperative with teacher” were collapsed into a single “uncooperative with teacher” item. Lonigan et al. (1999) found the internal consistencies of the three subscales, which consisted of the remaining 23 items, to be high; alpha values ranged from .83 to .91 for LI and MI preschool children. For each item, observers, teachers, and parents rated children’s behavior during the past month on a scale that varied from 0 (*not at all*) to 3 (*very much*). The sum scores for the three subscales of Conduct Problems, Hyperactivity, and Inattentive were used in all analyses.

Emotionality, Activity, Sociability, and Impulsivity Temperament Survey (EASI)—The EASI (Buss & Plomin, 1975) is a measure appropriate for children age 2 to 6 years. It contains 20 items with five items each representing the traits of Emotionality, Activity, Sociability and Impulsivity. Although designed to measure temperament, all items index behavioral tendencies, and the measure was chosen to provide the Activity and Impulsivity scales to complement the CTRS. Neale and Stevenson (1989) found that correlations between mothers’ and fathers’ ratings of their children on individual subscales averaged .53. Each child was rated on the 20 items using a 0-4 Likert scale of whether each statement applied to the child *not at all* to *very much*. Sum scores for each of the 4 subscales were used in analyses.

Procedure

Design—All children in participating classrooms were eligible for participation and were included if parents granted consent. Primary parental caregivers completed the CTRS and EASI as part of a larger packet of questionnaires. Packets were mailed to parents of MI children with stamped return envelopes. Parents of Head Start children received their questionnaires at the preschool center returned them there; they received five-dollar gift certificates for completing the questionnaires whereas MI parents were not compensated. Two teachers (typically lead and aide), completed the CTRS and EASI for each child in packets delivered to and collected from the preschools. Teachers received $10 gifts of school supplies for their classrooms. In year 1, 21 teachers of MI children and 13 teachers of LI children completed questionnaires, and in year 2 30 teachers of MI children and 34 teachers of LI children completed questionnaires. Analyses included children for whom at least one teacher, one observer, and a parent completed ratings.

Observations—Trained female undergraduate students completed observations of the children for course credit. Six students and 30 students participated in years 1 and 2, respectively. Ethnic data was available for some, but not all participating observers, yielding approximate percentages of observers as 80% Caucasian, 15% African American, and 5% Latino. Training lasting 3 to 6 weeks involved instruction with videotapes and in pilot classrooms on how to observe multiple children's activities while minimizing interaction and disruption. Observers learned to identify typical and atypical behaviors and examples consistent with each item on the two measures. They were instructed to compare the behavior of a child to that of a typical child representative of all children they had observed when assigning ratings, not just those from the same classroom or income group. Training facilitated the collective development of normative standards for the intensity and frequency of behaviors consistent with each rating scale level.

Within a period of 7-8 weeks, two observers watched each child during the same 2- to- 3-week interval. During each interval, each pair of observers was assigned two groups of approximately eight children, with one group representing each of the income groups. This design prevented order effects from coming into play wherein the first SES group rated by an observer may have become the point of reference for any later group from the other SES sample. The CTRS and EASI were completed for each group at the end of the observation interval. The observers were randomly paired during each observation interval and no observer watched more than one group of children in a classroom. Wherever possible, no observers rated more than one set (i.e., one group from each income level) of children with a single partner. To increase the likelihood of seeing the children in a wide variety of structured and unstructured activities (i.e., story time, art activities, free play on the playground) observation periods were distributed throughout morning and afternoon hours. Children retained for analyses were observed for no fewer than 4.5 hours by at least one observer; averages were six hours by each assigned observer.

Results

Preliminary Analyses and Descriptive Statistics

Within each combination of informant and income group, less than 2% of item-level data were missing. Missing values in the teacher and observer data were replaced with the median value for an item for each individual rater. Missing values in the parent data were replaced with the median value for an item reported by the parents of each group of children (i.e., MI or LI). Fourteen teachers from MI classrooms and six teachers from LI classrooms did not complete item 13 on the EASI; likely because it requested information regarding child behavior upon morning awakening to which they were not privy. The Activity scale was prorated for these teachers to account for the missing score. CTRS and EASI subscale scores were computed for each source. The internal consistency of subscales was adequate to excellent across informant and income group (see Table 1); with the exception being unexpectedly low alphas for Sociability. Given these low values, further analyses were not conducted with the Sociability subscale. For all analyses the two observer and, separately, the two teacher subscale scores for each child were averaged. These averages were used to maximize the reliability of the ratings from these two sources. Bivariate correlations for

teachers were significant for all CTRS and EASI scales; all but one correlation was significant for observers. Table 1 includes means and standard deviations for averaged observer, averaged teacher, and parent ratings for children from MI and LI centers. For all included subscales a higher score indicates more behavioral problems.

Relative Relations Between Sources

Investigation of the first hypothesis required calculation of the relative agreement in ratings between the three informant types. These were assessed with intraclass correlations (ICC), a measure of the agreement between rater pairs. Highly comparable findings analyzing the correlated correlations are available from the first author. Following Shrout and Fleiss (1979), an appropriate intraclass formula was selected that allowed for the decomposition of variance into that attributable to differences between participants (i.e., mean square between) and differences within participants (i.e., error variance). Cross-rater ICCs were computed for each pair of sources and for all three raters simultaneously on the six subscales, and agreement between different pairs was compared both within and between income groups for each scale. Results of the ICC analyses are in Table 2. Shrout and Fleiss (1979) also provided guidance as to the significance of ICCs by supplying an *F* test for the null hypothesis that the difference between the ICC value and zero is not significant. The findings of the ICC calculation and the *F* tests are provided for each group and measure below.

Middle Income—ICC results for the CTRS subscales are shown in Table 2. All overall and pairwise ICC values were significantly greater than zero. The overall average agreement between the three sources for the CTRS scales was .33. Within this sample, the average ICC between observer and teacher ratings was .33 and the average ICC between observer and parent ratings was .22 (after *r* to *z* transformation). There was better agreement between the ratings of teachers and parents in this group; the average ICC was .41, with particularly good agreement seen for Hyperactivity and Inattention, which assessed higher base-rate behaviors than Conduct Problems. To ascertain whether hypothesis 1 could be supported at the scale level, the ICC between each combination of informants was compared to all other combinations for each subscale. Statistical significance for these t-test analyses of correlated correlations (after *r* to *z* transformation) was set at $p < .01$ using Bonferroni correction for the number of interrelated subscale scores. After correction, none of the paired ICCs for Hyperactivity differed from one another significantly. For Inattention, the agreement between the teacher and parent ratings was significantly better than that between observer and parent ratings. For Conduct Problems, both the teacher-observer and the teacher-parent agreement on ratings were better than that between the observers and the parents, but not significantly so after correction for nonindependence. Taken as a whole, these results provide limited support for hypothesis 1 within the CTRS measures. ICC results for the EASI subscales are shown in Table 2. All overall and pairwise ICC values were significantly greater than zero and the overall average agreement for the EASI scales was .31. Findings indicated moderate agreement between teachers and parents and teachers and observers; these average ICCs were .43 and .38, respectively. For observer-parent agreement, the average was .32. There were no significant differences in the magnitude of agreement

between the three rater pairings after Bonferroni correction, although agreement between teachers and parents on Activity was somewhat greater than that between other pairs.

Low Income—Results of the ICC for CTRS ratings of children from LI centers indicated that there was somewhat poorer cross-rater agreement on these scales (see Table 2) than for ratings of children from MI centers. However, as with the results for the MI group, all overall ICC values were significantly greater than zero; overall the agreement across all three sources averaged .19. The pairwise comparisons were significant for all combinations and measures except for the parent-observer relation on Hyperactivity and Conduct Problems. The average teacher-parent agreement was .22, whereas the average teacher-observer agreement was .25. The average observer-parent agreement was substantially lower at .08. For Hyperactivity, only the difference in agreement between the teacher-observer and the parent-observer pairings was significant. For Inattention, none of the paired comparisons differed significantly, despite somewhat higher agreement for teacher-parent pairings. For Conduct Problems, both the teacher-parent agreement and the teacher-observer agreement were better than the observer-parent agreement, although only the latter was significant after correction for nonindependence. These results provided partial, somewhat better, support for hypothesis 1, in that teacher-observer agreement was stronger than observer-parent agreement on some subscales, although teacher-observer agreement was never significantly stronger than teacher-parent agreement.

ICC results for the three retained EASI subscales are shown in Table 2. Examination indicated that again the average agreement between sources was relatively low, but again all overall relations were significantly greater than zero. The overall average agreement across all three sources was .18. All pairwise comparisons were significant except for those between observers and parents for Emotionality and Impulsivity. Average teacher-parent agreement was .20, whereas average teacher-observer agreement was .27. The average agreement between observers and parents was lower at .09. On Activity and Impulsivity there were no significant differences in agreement between sources after correction. However, for Emotionality, teacher-observer agreement was significantly higher than observer-parent agreement. Supporting hypothesis 1, the EASI results for LI children suggest that the observer-teacher combination had the best agreement, although not always significantly so.

Comparison Between Income Groups

The final question regarding relative agreement between informants was whether these relations differed across type of preschool. That is, for example, was the observer-teacher ICC for Hyperactivity significantly different in the MI sample than in the LI sample? Analyses of independent correlations indicated that for the observer-teacher pairing, there were no significant differences across preschool type for any subscale. For the observer-parent pairing, there were no significant differences for CTRS subscales or for Emotionality but there were significant differences favoring the children from MI centers for Activity, z (362) = −2.31, $p < .05$, and Impulsivity, z (362) = −3.05, $p < .01$. Whereas teacher-parent agreement was higher for children from MI centers than for children from LI centers for both the Hyperactivity and Inattention subscales, the differences were not significant after

correction (e.g., z [362] = −2.46, $p < .05$ and z [362] = −2.38, $p < .05$, respectively). In contrast, on EASI subscales there were significant differences in agreement magnitude across income group for the teacher-parent agreement on Activity, z (362) = −4.01, $p < .001$, and Impulsivity, z (362) = −3.07, $p < .01$.

Differences in Group Means across Rater Types

The second main hypothesis evaluated was that ratings of children from the LI centers would be higher, indicative of more behavioral problems, than those given to children from MI centers. Three analytic questions related to potential mean score differences were addressed to evaluate this hypothesis and explore whether there were differences in support for this hypothesis by rater type. First, were average scores within rater type different across the two groups? Second, were average scores for the three sources significantly different? Third, did these differences vary across the two groups for any informant combinations? To answer these questions a series of repeated measure ANOVAs was conducted with a 3 (type of rater) by 2 (type of center) analysis for each subscale. Further, these analyses investigated any significant interactions between group and the relation of average scores across raters.

Income Group Differences

The ANOVA results regarding main effects and interaction effects for group, shown in Table 3, indicated that there were overall group differences for three of the six analyzed subscales, namely CTRS Inattention, EASI Emotionality, and EASI Impulsivity. However, analyses of the individual rater types for CTRS Inattention indicated that whereas teacher ratings were on average higher for the children from LI centers, $F(1, 364) = 10.59$, $p < .001$, neither the observer, $F(1, 364) = 2.20$, $p > .05$, nor the parent, $F(1, 364) = 2.11$, $p > .05$, mean ratings differed significantly between the groups. For EASI Emotionality, follow-up ANOVAs indicated that teachers in the LI classrooms rated children higher than did teachers in the MI classrooms, $F(1, 364) = 11.48$, $p < .001$. Likewise, parent results for Emotionality indicated a significant difference between scores for the two groups, $F(1, 364) = 5.08$, $p < .05$. However, there was not a significant difference between the two groups for observer ratings on this subscale, $F(1, 364) = 0.00$, $p > .05$, with children in MI and LI classrooms receiving virtually identical mean scores. In keeping with the general pattern, follow-up ANOVAs for the three rater types on EASI Impulsivity indicated that teacher ratings were significantly higher for children from the LI classrooms than for children from the MI classrooms, $F(1, 364) = 14.16$, $p < .001$ but observer average scores, F (1, 364) = 1.55, $p > .05$, and parent average scores, F (1, 364) = 3.16, $p < .05$, did not differ significantly across the two groups.

Differences Among Raters

ANOVA results for all six CTRS and EASI subscales indicated significant mean differences for observer ratings relative to teachers, parents, or both; in all cases the mean observer ratings indicated fewer behavior problems than the mean teacher or parent ratings (see Table 3). Follow-up contrast analyses of CTRS Hyperactivity revealed that mean scores were significantly different between observers and both other rater types, and that teachers and parents also were significantly different in their average ratings. There was a significant

interaction for the relation between the teacher and observer average scores, such that these ratings were more similar for children from MI centers than for children from LI centers. Follow-up analyses of the separate groups supported this pattern, indicating that there were significant differences between observers and teachers in the LI group, $F(1, 198) = 83.91$, $p < .001$, as there were, although of lesser magnitude, for the MI group, $F(1, 165) = 11.50$, $p < .001$. There was no significant interaction for the other relations between rating sources.

Average ratings of the Inattention subscale were significantly different between observers and teachers and between observers and parents, but not between teachers and parents. For observer-teacher and observer-parent scores there was a significant interaction with group; convergence was higher in MI centers than in LI centers. Analyses of the separate groups indicated that mean score differences between observers and teachers were significant for both groups, $F(1, 165) = 29.79$, $p < .001$, and $F(1, 198) = 117.32$, $p < .001$, for MI and LI ratings, respectively, although there was a greater discrepancy within the LI group. Similarly, analyses of observer and parent scores indicated significant differences for children from the MI group, $F(1, 166) = 57.34$, $p < .001$ and for children from the LI group, $F(1, 198) = 78.88$, $p < .001$. Results for the CTRS Conduct Problems subscale indicated that average ratings differed between all pairs of raters and supported a significant interaction between rater type and group. In particular, the relation between observers and teachers differed significantly across type of center. Although both income groups yielded significant mean differences between observers and teachers, there was more similarity between the scores for children from MI centers, $F(1, 165) = 57.12$, $p < .001$, than between scores for children from LI centers, $F(1, 198) = 151.73$, $p < .001$.

Results for EASI Emotionality revealed a significant interaction between rater type and group, as well as significant main effects for both rater type and group (see Table 3). Follow-up contrasts indicated significant differences for all three pairings of rater type. Observers and teachers differed significantly on their ratings of children from MI centers $F(1, 165) = 58.47$, $p < .001$, and even more so in their ratings of children from LI centers, $F(1, 198) = 203.77$, $p < .001$. Significant differences also were found for observer and parent ratings of children in the MI centers, $F(1, 165) = 216.67$, $p < .001$, and LI centers $F(1, 198) = 460.31$, $p < .001$. Similarly, for the teacher-parent pairing, there were significant differences for ratings of children from MI centers, $F(1, 165) = 56.36$, $p < .001$, and LI centers, $F(1, 198) = 97.95$, $p < .001$.

The analyses for EASI Activity found that the overall effect of rater was significant, and that the rater types were significantly different across all pairings. Ratings did not differ across the two types of centers, and the interaction between rater and type of center was not significant therefore no follow-up analyses were conducted. Finally, the results for EASI Impulsivity indicated an overall significant effect of rater type, group, and the interaction between rater and group. Average ratings were significantly different across all three pairings of raters. The interaction between rater and group was significant for the observer-teacher and the observer-parent pairings. Follow-up results showed that the difference between teachers and observers was significant but small for the children from the MI centers, $F(1, 165) = 6.02$, $p < .05$, and significant and larger for the children in LI centers, $F(1, 198) = 93.66$, $p < .001$. The average scores provided by observers and parents were

significantly and substantially different both for the children from MI classrooms, $F(1, 165) = 119.12$, $p < .001$, and for the children from the LI classrooms, $F(1, 198) = 185.66$, $p < .001$.

Discussion

This study yielded a number of significant findings about the relations between multiple raters of preschool children's behavior. To our knowledge it is the first study to explore the relations among behavior ratings from teachers, parents, and observers in a sample that included lower-income and minority preschool children. As well, this is one of a few studies of preschool children in which all raters used common measures. Results showed that in general the intraclass correlations were significantly greater than chance in both income groups. Despite distinct classroom environments, agreement between teacher and observer ratings was largely consistent across both income groups. However, rater agreement was moderate at best and generally better for the MI group. Further, mean score comparisons indicated that teacher ratings were significantly higher in the LI classrooms, whereas observers consistently provided the lowest average ratings of behavior problems. In general, these results provide evidence both of the commonality and of the significant differences found between different sources when rating the same children's behaviors, and they illustrate the need for multiple methods of evaluation when assessing a child's behavior, particularly when rating children from lower income backgrounds.

Partial support was found for hypothesis 1, that teacher-observer agreement would surpass either relation with parents, although this trend was more frequent in the LI group. Teacher-observer relations were stable across income groups, and generally comparable to the teacher-parent relation; the latter was typically better in the MI group. Observers and parents were the least consistent; only showing meaningful agreement for the MI group. Results of the ANOVAs evaluating the relations between raters' average scores generally supported these findings. Hypotheses 2, regarding higher ratings (i.e., more behavior problems) for LI children, also was partially supported, in that whereas there were overall group differences for most subscales, only teachers' and occasionally parents' ratings were significantly different. There were significant differences between rater types for most scales, and frequent interactions between rater type convergence and center type.

Across both groups of children, observer ratings were almost always significantly lower than the parent and teacher ratings. Observers may have 'normalized' behaviors because of their training in typical child behaviors and exposure to many more children in more settings than other raters. Consequently, extreme misbehaviors would be needed to produce higher than average ratings Teachers, and especially parents, who were exposed to far smaller reference groups, were likely more prone to rate children based on the presence or absence of the itemized behaviors, rather than in terms of whether the child evidenced these behaviors to an atypical degree. Importantly, these results cannot determine objective accuracy; that conclusion requires assessment of predictions from these ratings to later diagnostic findings. Observers were, by design, not engaged with the children in disciplinary or other behavior control functions. Therefore, no observer experienced a child's noncompliance or lack of attention as a direct violation of her directions and thus may not have had an affective

response engendered by such refusals. Moreover, whereas observers were trained to attend to all child behaviors, including appropriate, positive examples, parents and teachers may have allocated attention to a child more frequently when the child was being disruptive or noncompliant, therefore creating a skewed representation of their behavior (i.e., biased recall, see, e.g., Tversky & Marsh, 2000).

Prior research has suggested (e.g., Lovejoy, 1996; Mills and Rubin, 1990) that teachers and parents are most bothered by, and thus more likely to base evaluations on, behaviors that directly impact them. Likewise, earlier research (e.g., Abikoff et al., 1993; De Los Reyes & Kazdin, 2005; Schachar et al., 1986) supports the theory that halo effects can influence ratings. In this study, it is possible that experiences of disobedience and inattention were inordinately attended to by teachers and parents, remembered more than moments of appropriate behavior, and ultimately weighed heavily in their ratings. In a similar way, most parent ratings were higher than those from other informants, a finding that held across both groups to roughly the same degree. Parents, regardless of SES, who see their children primarily in one-on-one situations, may be exposed infrequently to their children's behaviors in large group situations and with the child's peers. This limited context may provide parents with very few other children against which to compare their children's behavior, and may lead them to rate their children's behaviors as being relatively extreme.

Supporting the second hypothesis, children from LI centers were rated by teachers and sometimes parents as having more behavioral problems. These results are consistent with literature suggesting that higher rates of behavior problems are reported more frequently for children from lower income backgrounds (e.g., Achenbach & Howell, 1993; Duncan et al., 1994; Kaiser, Cai, Hancock, & Foster, 2002). Low SES may contribute directly and indirectly through socialization and associated factors, to the increased prevalence in behavioral problems (Dodge et al., 1994; McLeod & Shanahan, 1993). Detailed investigation revealed, however, that observer ratings were not affected by children's group membership and that parent ratings were not always different. If observer ratings actually were more accurate, these findings suggest that higher teacher ratings of LI children here, and previously, may have been the product of differential expectations and comparisons, rather than a genuinely higher frequency of behavior problems. Longitudinal studies are needed to discern which ratings have the best predictive validity.

The current results contrast with Lonigan et al., (1999) where teachers in LI classrooms rated children as having significantly fewer behavior problems and rule violations. One key difference between the current study and Lonigan et al. is that in that study the MI classrooms were much more structured than the LI classrooms. This structure was thought to place additional attention and compliance demands on children from the MI classrooms, resulting in more opportunities for rule violation and behavioral difficulties. Informal evaluations of the environments of classrooms in the current study revealed a quite different pattern. In general, the MI classroom environments were child-centered, with play-based curricula and opportunities for child choice of activities. The teachers also shared frequent, positive interactions with the children, as a group and individually. In contrast, the LI classrooms were characterized by lengthy whole group, teacher-directed activities and by fewer positive communications and more disciplinary interactions between teachers and

children. It may be that the environments in the LI classrooms increased the likelihood that children would violate teachers' expectations for attention and compliance with rules, and therefore, that these teachers would then rate these children as having more behavioral problems than their counterparts in the MI classrooms. Teacher-child interactions in the LI settings may have been subject to similar transactional influences as those that affect parent-child relationships. Coercive teacher behaviors and noncompliant child behaviors may have each increased the likelihood of the other.

Thus, teachers in LI classrooms indeed may have been exposed to higher levels of misbehavior than were teachers in MI classrooms. Instead of becoming desensitized to this behavior, they may have developed a reduced tolerance for children's misbehaviors. Whereas the observers may, as discussed above, have viewed the behavior of most children in the LI classrooms as typical and developmentally appropriate, these teachers may have seen the behavior as typical yet inappropriate, and worthy of higher ratings. It also is possible, as de Ramirez and Shapiro (2005) found, that minority teachers such as those in the LI classrooms may be more stringent when it comes to their behavioral expectations, especially for children of their own ethnicity.

In this study, SES and ethnicity were confounded for both children and teachers as the MI group was mostly Caucasian and the LI group predominantly African-American. This general match between teacher and child ethnicity contrasts with prior studies in which, regardless of the children's ethnicity or SES, virtually all teachers were Caucasian. As such, higher teacher ratings for the almost exclusively African American children in the LI classrooms cannot be ascribed to bias of Caucasian teachers against African American children. Child and teacher ethnicity cannot be fully distentangled from SES in these data. As such, we cannot determine whether one or more of these factors influenced the higher teacher (and sometimes parent) ratings of children in the LI classrooms. Prior evidence suggests that child ethnicity does not always contribute significant variance to the differences seen between SES groups once mediating factors such as parenting styles, familial backgrounds, and stressors are taken into account (e.g., Deater-Decker et al., 1998; Dodge et al., 1994; McDermott & Schafer, 1996). As discussed, setting factors such as teacher interactional style, perhaps impacted by resources and trainings less available to teachers in LI classrooms, may also be a significant influence on the behavior of the teachers and children in those classrooms. Future research should include larger percentages of Caucasians and other ethnic groups from LI centers, as well as more minorities, both children and teachers, at MI centers. Such work will allow needed differentiation of contributions of child and teacher ethnicity, the match between teacher and child ethnicity, and SES to children's behavioral problems and to the pattern of ratings the children receive.

These and previous results (e.g., De Los Reyes, & Kazdin, 2005; Milfort & Greenfield, 2002) support the use of multiple sources of data when assessing the behaviors of young children, as the teacher, observer, and parent ratings were clearly not interchangeable. This is most applicable for the children in the LI Head Start centers where relations between different raters' scores were sometimes quite poor. It may be that the uniform elevation of teacher reports in these LI classrooms decreases the specificity of ratings for children with the most significant behavior problems. Likewise, Kaiser et al. (2002) found that Head Start

teachers' years of experience was significantly related to their behavior ratings of children in their classrooms. As such, these children, who may benefit from early intervention services, may be more easily identified by the convergence of multiple sources of data, representing both home and school behaviors. Prior studies of Head Start centers (e.g., Mowder, Unterspan, Knuter, Goode, & Pedro, 1993; Piotrowski, Collins, Knitzer, & Robinson, 1994) indicated that despite often receiving higher teacher and parent ratings, children in the LI classrooms are less likely to be referred for intervention services while still in preschool. Rather, the probable pattern is that many of these children will first receive intervention when they display behavior problems or academic difficulties in elementary school. The use of multiple ratings, and other evaluation methods, for preschool children may increase the potential for earlier identification and services for those children most at-risk for later problems.

Although providing a solid basis for future research in this area, this study had limitations. First, as discussed we could not fully separate influences of SES and ethnicity within this study. Second, in large part because of absences, there was more variability than desired in the amount of time that each child was observed. Despite this, however, the children were observed on average for longer than is typically seen in the literature. Moreover, the children were observed in a variety of free-play and group activities both inside the classrooms and on the school playgrounds. Third, there was a larger than anticipated length of time between completion of observer ratings and teacher and parent ratings, which may have affected the strength of the relation between these different sources. Multiple attempts and considerable efforts were needed to obtain the teacher and parent ratings, especially for the children in the LI centers. Future studies in which the interval for which ratings are completed is better controlled may help to minimize the influence of this factor. Conversely, the use of common rating scales for all raters may have reduced error arising from having to equate across different measures and allowed for direct comparison. Finally, the lower than expected internal consistency for the Sociability scale did not permit the inclusion of that subscale, which was intended to offer a prosocial counterpoint to the measures of negative behaviors. Future research thus needs to explore comparability of ratings on prosocial measures within preschool samples. Given that we used older behavioral measures used often in research studies, but not currently used in classroom contexts, follow-up studies should replicate these results with more common measures. Future work that also further minimizes age differences and minor procedural differences between samples would provide needed replication.

References

Abikoff H, Courtney M, Pelham WE, Koplewicz HS. Teachers' ratings of disruptive behaviors: The influence of halo effects. Journal of Abnormal Child Psychology. 1993; 21:519–533. [PubMed: 8294651]

Achenbach, TM.; Edelbrock, C. Manual for the Child Behavior Checklist and Revised Child Behavior Profile. Author; Burlington VT: 1983.

Achenbach TM, Howell CT. Are American children's problems getting worse? A 13-year comparison. Journal of the American Academy of Child and Adolescent Psychiatry. 1993; 32:1145–1154. [PubMed: 8282658]

Achenbach TM, McConaughy SH, Howell CT. Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. Psychological Bulletin. 1987; 101:213–232. [PubMed: 3562706]

Ackerman BP, Brown E, Izard CE. Continuity and change in levels of externalizing behavior in school children from economically disadvantaged families. Child Development. 2003; 74:694–709. [PubMed: 12795385]

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th ed. Author; Washington, DC: 2000. Text Revision

Atkins MS, Pelham WE, Licht MH. The differential validity of teacher ratings in inattention/hyperactivity and aggression. Journal of Abnormal Child Psychology. 1989; 17:423–435. [PubMed: 2794255]

Biederman J, Milberger S, Faraone SV, Kiely K, Guite J, Mick E, Ablon S, Warburton R, Reed E. Family-environment risk factors for Attention Deficit Hyperactivity Disorder. Archives of General Psychiatry. 1995; 52:464–470. [PubMed: 7771916]

Brandon KA, Kehle TJ, Jenson WR, Clark E. Regression, practice and expectation effects on the Revised Conners Teacher Rating Scale. Journal of Psychoeducational Assessment. 1990; 8:456–466.

Buss, AH. The EAS theory of temperament. In: Strelau, J.; Angleitner, A., editors. Explorations in temperament: International perspectives on theory and measurement. Perspectives on individual differences. Plenum Press; New York, NY: 1991. p. 43-60.

Buss, AH.; Plomin, R. Temperament: Early Developing Personality Traits. Erlbaum; Hillsdale, NJ: 1984.

Buss, AH.; Plomin, R. A temperament theory of personality development. Wiley; New York: 1975.

Campbell SB, Pierce EW, March CL, Ewing LJ, Szumowski EK. Hard-to manage preschool boys: Symptomatic behavior across contexts and time. Child Development. 1994; 65:836–851. [PubMed: 8045171]

Campbell SB, Pierce EW, Moore G, Marakovitz S, Newby K. Boys externalizing problems at elementary school age: Pathways from early behavior problems, maternal control, and family stress. Development and Psychopathology. 1996; 8:701–719.

Chang DF, Stanley S. The effects of race and problem type on teacher's assessments of student behavior. Journal of Consulting and Clinical Psychology. 2003; 71:235–242. [PubMed: 12699018]

Conners, CK. Conners' Ratings Scales Manual. Multi-Health Systems; North Tonawanda, NY: 1989.

Danforth JS, DuPaul GJ. Interrater reliability of teacher rating scales for children with Attention-Deficit Hyperactivity Disorder. Journal of Psychopathology and Behavioral Assessment. 1996; 18:227–237.

Dawkins MP, Fullilove C, Dawkins M. Early assessment of problem behavior among young children in high-risk environments. Family Therapy. 1995; 22:133–141.

Deater-Decker K, Dodge KA, Bates JE, Pettit GS. Multiple risk factors in the development of externalizing behavior problems: Group and individuals. Development and Psychopathology. 1998; 10:469–493. [PubMed: 9741678]

De Los Reyes A, Kazdin A. Informant discrepancies in the assessment of child psychopathology: A critical review, theoretical framework, and recommendations for further study. Psychological Bulletin. 2005; 131:483–509. [PubMed: 16060799]

de Ramirez RD, Shapiro ES. Effects of student ethnicity on judgments of ADHD symptoms among Hispanic and White Teachers. School Psychology Quarterly. 2005; 20:268–287.

Dodge KA, Pettit GS, Bates JE. Socialization mediators of the relation between socioeconomic status and child conduct problems. Child Development. 1994; 64:649–665. [PubMed: 8013245]

Downey DB, Pribesh S. When race matters: Teachers' evaluations of students' classroom behavior. Sociology of Education. 2004; 77:267–282.

Duhig AM, Renk K, Epstein MK, Phares V. Interparental agreement on internalizing, externalizing, and total behavior problems: A meta-analysis. Clinical Psychology: Science and Practice. 2000; 7:435–453.

Duncan GJ, Brooks-Gunn J, Klebanov PK. Economic deprivation and early childhood development. Child Development. 1994; 65:296–318. [PubMed: 7516849]

Epstein JN, Willoughby M, Valencia EY, Tonev ST, Abikoff HB, Arnold LE, et al. The role of children's ethnicity in the relationship between teacher ratings of Attention-Deficit/Hyperactivity Disorder and observed classroom behavior. Journal of Consulting and Clinical Psychology. 2005; 73:424–434. [PubMed: 15982140]

Feil EG, Severson HH, Walker HM. Identification of critical factors in the assessment of preschool behavior problems. Education and Treatment of Children. 1995; 18:261–271.

Grietens H, Onghena P, Prinzie P, Gadeyne E, Van Assche V, Ghesquiere P, Hellinckx W. Comparison of mothers', fathers', and teachers' reports on problem behaviors in 5- to 6-year-old children. Journal of Psychopathology and Behavioral Assessment. 2004; 26:137–146.

Gross D, Fogg L, Garvey C, Julion W. Behavior problems in young children: An analysis of cross-informant agreements and disagreements. Research in Nursing and Health. 2004; 27:413–425. [PubMed: 15514961]

Heller TL, Baker BL, Henker B, Hinshaw SP. Externalizing behavior and cognitive functioning from preschool to first grade: Stability and predictors. Journal of Clinical Child Psychology. 1996; 25:376–387.

Hosterman SJ, DuPaul GJ, Jitendra AK. Teacher ratings of ADHD symptoms in ethnic minority students: Bias or behavioral difference? School Psychology Quarterly. 2008; 23:418–435.

Kaiser A, Cai X, Hancock T, Foster M. Teacher-reported behavior problems and language delays in boys and girls enrolled in Head Start. Behavioral Disorders. 2002; 28:23–39.

Kerr DCR, Lunkenheimer ES, Olson SL. Assessment of child behavior problems by multiple informants: A longitudinal study from preschool to school entry. Journal of Child Psychology and Psychiatry. 2007; 48:967–975. [PubMed: 17914997]

Koblinksy SA, Kuvalanka KA, Randolph SM. Social Skills and behavior problems of urban, African American preschoolers: Role of parenting practices, family conflict, and maternal depression. American Journal of Orthopsychiatry. 2006; 76:554–563. [PubMed: 17209723]

Lonigan CJ, Bloomfield BG, Anthony JL, Bacon KD, Phillips BM, Samwel CS. Relations between emergent literacy skills, behavior problems, and social competence in preschool children: A comparison of children from low- and middle-income backgrounds. Topics in Early Childhood Special Education. 1999; 19:40–53.

Lovejoy MC. Social inferences regarding inattentive-overactive and aggressive child behaviors and their effects on teacher reports of discipline. Journal of Clinical Child Psychology. 1996; 25:33–42.

McDermott PA, Schafer BA. A demographic survey of rare and common problem behaviors among American students. Journal of Clinical Child Psychology. 1996; 25:352–362.

McGee R, Partridge F, Williams S, Silva PA. A twelve-year follow-up of preschool hyperactive children. Journal of the American Academy of Child and Adolescent Psychiatry. 1991; 30:224–232. [PubMed: 2016226]

McLeod JD, Shanahan MJ. Poverty, parenting, and children's mental health. American Sociological Review. 1993; 58:351–366.

Milfort R, Greenfield DB. Teacher and observer ratings of head start children's social skills. Early Childhood Research Quarterly. 2002; 17:581–595.

Milich R, Landau S. Teacher ratings of inattention/overactivity and aggression: Cross-validation with classroom observations. Journal of Clinical Child Psychology. 1988; 17:92–97.

Miller-Lewis LR, Baghurst PA, Sawyer MG, Prior MA, Clark JJ, Arney FM, Carbone JA. Early childhood externalizing behavior problems: Child, parenting, and family-related predictors over time. Journal of Abnormal Child Psychology. 2006; 34:891–906. [PubMed: 17103309]

Mills RSL, Rubin KH. Parental beliefs about problematic social behaviors in early childhood. Child Development. 1990; 61:138–151.

Miner JL, Clarke-Stewart KA. Trajectories of externalizing behavior from age 2 to age 9: Relations with gender, temperament, ethnicity, parenting, and rater. Developmental Psychology. 2008; 44:771–786. [PubMed: 18473643]

Mowder BA, Unterspan D, Knuter L, Goode C, Pedro MN. Psychological consultation and Head Start: Data, issues, and implications. Journal of Early Intervention. 1993; 17:1–7.

Neale MC, Stevenson J. Rater bias in the EASI temperament scales: A twin study. Journal of Personality and Social Psychology. 1989; 56:446–455. [PubMed: 2926639]

Olson SL, Hoza. Preschool developmental antecedents of conduct problems in children beginning school. Journal of Clinical Child Psychology. 1993; 22:60–67.

Piggott RL, Cowen EL. Teacher race, child race, racial congruence, and teacher ratings of children's school adjustment. Journal of School Psychology. 2000; 38:177–196.

Piotrowski CS, Collins RC, Knitzer J, Robinson R. Strengthening mental health services in Head Start. American Psychologist. 1994; 49:133–139. [PubMed: 7512314]

Puig M, Lambert MC, Rowan GT, Winfrey T, Lyubansky M, Hannah SD, et al. Behavioral and emotional problems among Jamaican and African American children, ages 6 to 11: Teacher reports versus direct observations. Journal of Emotional and Behavioral Disorders. 1999; 7:240–250.

Reid R, DuPaul GJ, Power TJ, Anastopoulos AD, Rogers-Adkinson D, Noll M-B, et al. Assessing culturally different students for attention deficit hyperactivity disorder using behavior rating scales. Journal of Abnormal Child Psychology. 1998; 26:187–198. [PubMed: 9650625]

Reid R, Maag JW. How many fidgets in a pretty much: A critique of behavior rating scales for identifying students with ADHD. Journal of School Psychology. 1994; 32:339–354.

Saft EW, Pianta RC. Teachers' perceptions of their relationships with students: Effects of child age, gender, and ethnicity of teachers and children. School Psychology Quarterly. 2001; 16:125–141.

Schachar R, Sandberg S, Rutter M. Agreement between teachers' ratings and observations of hyperactivity, inattentiveness, and defiance. Journal of Abnormal Child Psychology. 1986; 14:331–345. [PubMed: 3722627]

Schaughency EA, Lahey BB. Mothers' and fathers' perceptions of child deviance: Roles of child behavior, parental depression, and marital satisfaction. Journal of Consulting & Clinical Psychology. 1985; 53:718–723. [PubMed: 4056189]

Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin. 1979; 86:420–428. [PubMed: 18839484]

Shaw DS, Gilliom M, Ingoldsby EM, Nagin DS. Trajectories leading to school-age conduct problems. Developmental Psychology. 2003; 39:189–200. [PubMed: 12661881]

Shuller DY, McNamara JR. Expectancy factors in behavioral observation. Behavior Therapy. 1976; 7:519–527.

Sonuga-Barke EJ, Minocha K, Taylor EA, Sandberg S. Inter-ethnic bias in teachers' ratings of childhood hyperactivity. British Journal of Developmental Psychology. 1993; 11:187–200.

Tryon WW, Pinto LP. Comparing activity measurements and ratings. Behavior Modification. 1994; 18:251–261.

Tversky B, Marsh EJ. Biased retellings of events yield biased memories. Cognitive Psychology. 2000; 40:1–38. [PubMed: 10692232]

Wakschlag LS, Leventhal BL, Briggs-Gowan MJ, Danis B, Keenan K, Hill C, Egger HL, Cicchetti D, Carter AS. Defining the "disruptive" in preschool behavior: What diagnostic observation can tell us. Clinical Child and Family Psychology Review. 2005; 8:183–201. [PubMed: 16151617]

Zentall SS. Context effects in the behavioral ratings of hyperactivity. Journal of Abnormal Child Psychology. 1984; 12:345–352. [PubMed: 6725788]

Table 1

Means (and Standard Deviations) for Average Observer and Teacher Ratings, Parent Ratings, and [Cronbach's alphas] for Middle-Income and Lower-Income Groups of Children

Scale	Observer Avg.	Teacher Avg.	Parent	Observer Avg.	Teacher Avg.	Parent
		Middle-Income			Lower-Income	
CTRS Hyp.	3.58 (3.45) [.88]	4.80 (4.71) [.94]	6.52 (4.13) [.84]	2.91 (3.04) [.87]	5.89 (4.80) [.89]	6.82 (4.68) [.85]
CTRS Con.	1.91 (2.62) [.87]	4.23 (4.14) [91]	7.16 (4.04) [.81]	1.66 (2.27) [.79]	5.60 (4.79) [.90]	7.60 (5.32) [.87]
CTRS Att.	2.99 (2.39) [.74]	4.68 (4.07) [.88]	5.12 (3.36) [.76]	2.64 (2.20) [.71]	6.13 (4.36) [.86]	5.75 (4.82) [.87]
EASI Emo.	2.95 (2.26) [.71]	4.78 (3.16) [.73]	7.16 (3.48) [.72]	2.96 (2.27) [.65]	5.93 (3.31) [.64]	8.05 (3.96) [.62]
EASI Act.	6.27 (3.80) [.81]	7.51 (4.06) [.82]	10.85 (4.01) [.74]	5.50 (3.44) [.76]	7.35 (3.57) [.63]	10.56 (3.99) [.61]
EASI Soc.	12.47 (2.39) [.47]	13.27 (2.16) [.25]	12.98 (2.36) [.36]	12.31 (2.35) [.47]	12.84 (2.31) [.11]	12.20 (2.94) [.37]
EASI Imp.	5.06 (3.55) [.83]	5.87 (4.18) [.86]	8.33 (3.49) [.77]	4.62 (3.19) [.82]	7.32 (3.15) [.58]	8.99 (3.55) [.55]

Note. CTRS Total score ranges 0-69; Attn. and Con. subscales range 0-24; Hyp. subscale ranges 0 -21. EASI subscale range 0-20. Alpha statistics are for pooled individual observers and teachers. CTRS = Conners Teacher Rating Scale; EASI = Emotionality Activity Sociability Impulsivity Temperament Survey; Act. = Activity; Att. = Inattention; Con. = Conduct Problems; Emo. = Emotionality; Hyp. = Hyperactivity; Imp. = Impulsivity; Soc. = Sociability. Avg. = Average.

Table 2

Interrater Intraclass Correlations for Middle-Income and Lower-Income Conners Teacher Rating Scale and Emotionality, Activity, Sociability, Impulsivity Temperament Survey Scores

	Middle-Income				Lower-Income			
Scale	**Overall**	**Observer -Teacher**	**Observer -Parent**	**Teacher-Parent**	**Overall**	**Observer-Teacher**	**Observer-Parent**	**Teacher-Parent**
CTRS Hyp.	.37	$.36_{a\,b}$	$.27_{b}$	$.45_{a}$	.21	$.34_{a}$	$.07_{b}$	$.22_{a\,b}$
CTRS Inatt.	.35	$.29_{b}$	$.24_{b}$	$.46_{a}$	.18	$.13_{a}$	$.13_{a}$	$.24_{a}$
CTRS Con.	.28	$.35_{a}$	$.14_{b}$	$.32_{a}$	.18	$.28_{a}$	$.03_{b}$	$.21_{a}$
EASI Emo.	.27	$.37_{a}$	$.21_{a}$	$.24_{a}$	.19	$.32_{a}$	$.04_{b}$	$.23_{a}$
EASI Act.	.42	$.38_{a\,b}$	$.35_{b}$	$.52_{a}$	.18	$.27_{a}$	$.12_{a}$	$.15_{a}$
EASI Imp.	.43	$.40_{a}$	$.40_{a}$	$.50_{a}$	.18	$.23_{a}$	$.10_{a}$	$.22_{a}$

Note. CTRS = Conners Teacher Rating Scale; Hyp. = Hyperactivity; Con. = Conduct; Inatt. = Inattention; EASI = Emotionality, Activity, Sociability, Impulsivity Temperament Survey; Emo. = Emotionality; Act. = Activity; Imp. = Impulsivity. Observer and Teacher correlations based on averaged scores. Within each row and income group, pairwise correlations not sharing a common subscript differ from one another at $p < .05$. or $p < .01$. All correlations > .10 are significant at $p < .05$.

Table 3

Summary of Analysis of Variance Effects from Primary and Follow-up Analyses of Average Ratings of Children across Rater Type and Center Type

	Overall Analysis			Follow-up Two-Group Contrasts					
				Observer-Teacher		Observer-Parent		Teacher-Parent	
	Rater	Rater x Group	Group	Rater	Rater x Group	Rater	Rater x Group	Rater	Rater x Group
CTRS Scores									
Hyperactivity	85.76***	5.55**	0.56	75.02***	13.07***	167.40***	3.42	21.97***	1.87
Inattention	79.08***	7.19***	4.46*	131.31***	16.00***	128.64***	4.73*	0.01	2.42
Conduct	221.53***	4.66**	3.18	194.71***	13.08***	426.05***	1.60	68.29***	2.44
EASI Scores									
Emotionality	255.64***	4.26*	8.83**	203.77***	11.65***	460.31***	4.13*	97.95***	0.35
Activity	214.75***	0.913	0.16	46.75***	---	377.02***	---	190.25***	---
Impulsivity	155.80***	9.56***	4.17*	66.82***	19.33***	294.51***	6.09*	94.38***	3.42

Note. All F tests conducted on averaged teacher and observer ratings. CTRS = Conners Teacher Rating Scale; EASI = Emotionality, Activity, Sociability, Impulsivity Temperament Survey. df = 2 for Omnibus Rater and Rater × Group Analyses; df = 1 for all other contrasts.

* $p < .05$

** $p < .01$

*** $p < .001$.