Article

Hammill Institute on Disabilities

Topics in Early Childhood Special Education
2015, Vol. 34(4) 211–222

DOI: 10.1177/0271121414546008
tecse.sagepub.com
SAGE

Teacher and TA Ratings of Preschoolers' Externalizing Behavior: Agreement and Associations With Observed Classroom Behavior

Catherine Sanger Wolcott, MEd[1] and Amanda P. Williford, PhD[1]

Abstract

The present study investigated teachers' and teacher aides' (TAs) agreement in their ratings of preschoolers' externalizing behavior and their associations with observed classroom behavior for a sample of children at risk of developing a disruptive behavior disorder. One hundred twenty-two teachers rated 360 students' externalizing behavior in the first month of school, and within the next 6 weeks, children were observed in their classroom. Results indicated that teacher and TA reports were moderately correlated, teacher-rated severity of oppositional defiant disorder behaviors was related to agreement, and teacher/TA agreement did not predict observed externalizing behavior. In general, using teacher and TA ratings together versus a single informant's rating did not provide a better estimate of information gathered from observational measures. Results demonstrate the importance of gathering observations and rating scales when evaluating preschoolers' externalizing problems. Future work should explore factors that contribute to teachers', TAs', and observational methods' differential evaluations of externalizing behavior.

Keywords

assessment, teachers, personnel, prevention, behavioral, classroom, behavior

Children who display externalizing behaviors in the preschool classroom are at risk of a variety of negative outcomes, making it critical to identify these children early in development (Dunlap et al., 2006). Externalizing behaviors are characterized by overactivity, impulsivity, aggression, and non-compliance/defiance (Hinshaw, 1992). High levels of these problem behaviors are often the precursors of later, more serious developmental disorders such as attention-deficit/hyperactivity disorder (ADHD), oppositional defiant disorder (ODD), and conduct disorder (CD; Calkins, Gill, & Williford, 1999). Preschoolers displaying problem behaviors are three times as likely to be expelled from early childhood programs compared with their school age peers (Gilliam, 2005), a reality that is especially concerning, as the majority of children spend substantial time in preschool prior to kindergarten entry (Adams, Tout, & Zaslow, 2007). Importantly, high-quality preschool programs can serve as a protective factor for children at risk of developing later behavioral problems (Frede & Barnett, 1992).

Due to the negative outcomes associated with the display of externalizing problems during early childhood, it is critical to accurately identify these children to change developmental trajectories through early intervention and prevention efforts (Campbell, 2002; Querido & Eyberg, 2004). Assessment plays a key role in this process, and information from the preschool classroom is especially important, as school is often a setting in which maladjustment related to externalizing problems is likely to occur due to heightened social and academic demands (Wolraich et al., 2004).

Assessment in the Preschool Classroom

The assessment of preschool externalizing problems should gather information from the child's school or day care setting and should use multiple informants and multiple methods, such as rating scales and direct observations (Carter, Briggs-Gowan, & Davis, 2004; Egger & Angold, 2006). Direct observations in naturalistic settings are useful because they are considered an unbiased way of capturing a child's environmental functioning (Carter et al., 2004).

[1]University of Virginia, Charlottesville, USA

Corresponding Author:
Catherine Sanger Wolcott, Center for Advanced Study of Teaching and Learning, University of Virginia, 350 Old Ivy Way, Suite 100, Charlottesville, VA 22903, USA.
Email: ces2jg@virginia.edu

These observations take place in-context, evaluate actions at the time of their occurrence, and are assumed to be representative of a child's usual behavior in these settings (Qi & Kaiser, 2004; Thomas, Shapiro, DuPaul, Lutz, & Kern, 2011). However, direct observations are timely and costly, frequently making them impractical for everyday use (Pelham, Fabiano, & Massetti, 2005; Thomas et al., 2011). Another major concern is that these observations use a time sampling approach, which can lead to the failure to detect behaviors that have a low base rate but are clinically significant, such as conflict and aggression (Pelham et al., 2005).

Due to the limitations of direct observations, indirect measurements, such as rating scales, are often used to assess externalizing problems (Thomas et al., 2011). Teacher rating scales are considered an important source of information because teachers interact with children in the school context, where problem behaviors often manifest themselves (Wolraich et al., 2004). In addition, teachers are in a position to assess children's irregular patterns of externalizing behavior, a tendency that is attributed to their ability to compare children's behaviors to typical patterns of development and to their experience interacting with a variety of students (Atkins & Pelham, 1991). Finally, teachers are frequently involved in the expulsion of preschool children (Gilliam, 2005), highlighting that teachers' perceptions of children's behavior is associated with important child outcomes. Despite these strengths, teachers may artificially inflate their report of children's externalizing behavior; for example, studies have shown that teacher ratings of hyperactivity spuriously increase when children display behaviors characteristic of ODD (Abikoff, Courtney, Pelham, & Koplewicz, 1993; Jackson & King, 2004). This suggests that teachers may be influenced by halo effects, which can lead to bias in their reports.

Gathering ratings from two teachers, both of whom interact with and observe children in the same classroom, may help us better understand the extent to which important adults with the same role view children's externalizing behavior. Although work has investigated multiple teachers' ratings from different classrooms and found that teachers' ratings are more highly correlated than teacher and parent ratings (Achenbach, McConaughy, & Howell, 1987), few studies have examined ratings from teachers within the same classroom. In many preschool classrooms, a teacher aide (TA) works with the teacher during the school day, helping with classroom responsibilities and providing extra support to children (Essa, 2010). However, relatively little work has examined the utility of the TA's perspective in understanding preschool externalizing behavior, and few studies have investigated how TA and teacher perspectives align. In a study that examined TAs' reports of elementary school students with an intellectual disability (ages 5–12, M = 9.0), TAs' reports were more highly correlated with observed classroom behavior than lead teacher reports (Miller, Fee, & Jones, 2004). This suggests that although TAs interact with children in the same classroom, they may have different perspectives than teachers. However, given the particular population and wide age range of the sample used in the Miller et al.'s (2004) study, the extent to which TA ratings align with other measures of preschoolers' behavior remains relatively unexplored.

Challenges of Integrating Discrepant Reports

Given the strengths and limitations of various raters and methods, using information from multiple methods is most likely to result in a fair and accurate assessment of a child (Carter et al., 2004). However, the results produced from various sources are likely to be discrepant, thereby making it challenging to integrate conflicting reports (Achenbach, 2006; De Los Reyes, Henry, Tolan, & Wakschlag, 2009). Although discrepancy among raters has often been conceptualized as measurement error, recent work has examined the correlates and meaning of informant differences and has focused on context as a contributor to rater disagreement (De Los Reyes, 2011). De Los Reyes et al. (2009) examined whether parent and teacher discrepancies were associated with preschoolers' observed behavior in different contexts and found that the variability among parent and teacher reports of externalizing behavior was related to how children behaved in an observed lab task with a parent versus an unfamiliar adult (De Los Reyes et al., 2009). Similarly, Strickland, Hopkins, and Keenan (2012) demonstrated that the agreement between parent and teacher ratings of preschoolers' externalizing behaviors was higher when raters were instructed to report on children's behaviors in the same context.

In addition to relating to context, disagreement among informants has been hypothesized to relate to the severity of dysfunction (De Los Reyes et al., 2009). In cases in which impairing behavior is severe, informant agreement is likely to be stronger because raters might be more apt to notice the manifestations of severe behaviors across contexts and situations (De Los Reyes et al., 2009). Conversely, agreement would be expected to be lower when behaviors occur at moderate levels (De Los Reyes et al., 2009). Although this hypothesis makes sense conceptually (Wakschlag, Tolan, & Leventhal, 2010), we found no studies that explicitly tested this theory within the preschool population.

Present Study

The present study attempted to better understand the assessment of preschool externalizing problems by investigating rater agreement and examining the relation among multiple assessment methods. This study adds to the literature and explores the role of context in behavioral assessment

by utilizing the TA report and observational measures in combination with lead teacher ratings. First, we asked to what extent teachers and TAs agreed in their ratings of preschoolers' externalizing behaviors. We hypothesized that teacher ratings would be associated with TA ratings at similar levels to those cited in previous studies that have investigated rater agreement between individuals in similar contexts (average $r = .60$; Achenbach, 2006). Second, we asked whether agreement between teachers and TAs was related to the severity of behavior as evaluated by the lead teacher. We hypothesized that agreement between teachers and TAs would be greatest for children who were rated as evidencing either low or high levels of externalizing behaviors, as these children would either display few symptoms of externalizing behaviors, or their behavior would be severe and would occur consistently across contexts and caregivers, leading the teacher and TA to rate these behaviors similarly. Third, we asked whether teacher–TA agreement predicted observed externalizing behavior over and above a single teacher's rating. Although one would not expect teachers and TAs to rate children identically, we hypothesized that agreement would be associated with observed externalizing behavior in the classroom over and above a single informant's report.

We also investigated associations among teacher ratings, TA ratings, and observational measures. Because little work has examined TA report and because a previous study explored this issue with a sample of children with an intellectual disability (Miller et al., 2004), we were interested in whether teacher and TA ratings would be differentially related to observed externalizing behavior for children in preschool. Given that the Miller et al.'s (2004) study used a sample of teachers and TAs who likely have different roles and responsibilities than teachers in general education preschool classrooms, we were unsure if TAs' reports would be more aligned with observational measures. In addition, we investigated whether including two teachers' ratings, rather than using a single teacher's rating, would be more highly associated with observed externalizing behavior. Here, we speculated that teachers and TAs may observe children in different contexts within the preschool classroom (e.g., a teacher may observe all children's behavior as she leads circle time, whereas the TA may be getting materials ready for the next activity or may be attending to a single challenging child during this period), and therefore using both their ratings together would be more aligned with an independent observers' rating of a child's behavior across classroom activities. Given the limited information about the relation between teacher ratings, TA ratings, and observed externalizing behavior, examining agreement and associations among different types of assessment approaches allows greater understanding of how various methods may be combined to gather a more complete picture of the externalizing behaviors children display in the classroom.

Method

Participants

The participants for the present study were drawn from a sample of teachers and children who participated in an efficacy trial testing an early intervention to improve behavioral outcomes for preschool children displaying elevated levels of externalizing behavior. The intervention was not of interest and all data used in the current study were collected at baseline before the intervention took place. The larger 1-year intervention occurred with 471 children across three different cohorts in two mid-Atlantic states in a total of 173 classrooms with 183 lead teachers ($n = 37$ from Cohort 1, $n = 103$ from Cohort 2, and $n = 43$ from Cohort 3; the replacement procedure for the study meant that some lead teachers were replaced throughout the year, which is why there are slightly more lead teachers than total number of classrooms). The larger trial included teachers and children residing in mostly urban and suburban areas. A variety of preschool programs participated with the following composition of classrooms: 27% Head Start, 26% state-funded public, and 47% private (not-for-profit and for-profit) programs that served 3-, 4-, and/or 5-year-old children for 5 days a week. Of the initial 173 classrooms selected for the larger intervention trial, 122 classrooms had both a participating teacher and TA and are included in the present study. Of the classrooms that had both teacher and TA ratings, 25 participated from Cohort 1, 51 participated from Cohort 2, and 46 participated from Cohort 3. Classrooms with participating TAs ($n = 122$) were compared with the classrooms of teachers who did not participate, either because the TA did not fill out rating scales or the classroom did not have a TA ($n = 51$). No significant differences were found among the percentage of students who were African American, Hispanic, or White, or the percentage of children who were 3-, 4-, or 5 years old. Participating classrooms had a lower percentage of Asian students ($M = 0.02$), $t(148) = –2.77$, $p = .006$, than non-participating classrooms ($M = 0.06$). Non-participating classrooms were significantly more likely to have fewer students ($M = 12.84$), $t(154) = 4.975$, $p < .001$, than participating classrooms ($M = 15.96$). In addition, non-participating classrooms had a significantly lower percentage of boys ($M = 0.48$), $t(152) = 2.09$, $p = .038$, than participating classrooms ($M = 0.53$). Note that missing data on classroom demographic survey questions ranged from 11% to 17%.

Teachers. The 122 lead teachers were mostly female (97.5%) and were on average 42 years of age (range = 22–67). The sample was fairly ethnically diverse, with 53% of lead teachers identifying themselves as White, 42% as Black, and 5% as Hispanic, Native American, Asian, Multiracial, or Other ethnicity. The average education of teachers was 15.6 years (range = 13–18). The average experience teaching preschool was 10 years (range = 0–38).

The 122 TAs who completed demographic data were 97.8% female and on average 35 years old (range = 18–68). They were also ethnically diverse, with 50.0% of TAs identifying themselves as Black, 39% as White, 4.5% as Hispanic, and 6.5% as Native American, Asian, Multiracial, or Other ethnicity. The average years of education was 14.22 years (range = 12–18), and the average years of experience teaching preschool was 8 years (range = 0–33).

Children. The children in the study were part of the sample selected for a larger efficacy trial designed to improve behavioral outcomes for children. Of the 471 children selected for the larger intervention (2 children selected per classroom in Year 1; 3 children selected per classroom in Years 2 and 3), 360 children had TA ratings in addition to teacher ratings. Of note, there are more total children in the sample than would be expected given the 122 participating classrooms due to the replacement procedure used in the study (if the child withdrew or moved classrooms, the next highest ranked child in the classroom was substituted and selected for the intervention). Children were 68% male and the average age was 48 months, ranging from 30 to 66 months. Participants were racially diverse, with 39% of children identified by caregivers as White, 41% as Black, 9% as Hispanic, 9% as multi-racial, and 2% as Native American, Asian, or Other ethnicity. The average years of maternal education was 14 (range = 11–20) and the average income to needs ratio was 1.86 (SD = 1.56, range = 0.20–6.07).

Procedures

Recruitment. Recruitment of preschool classrooms was consistent across the three sites. Program directors were contacted through paper mailings, emails, and physical visits and were asked permission to recruit their teachers for participation. Once a program agreed, teachers and TAs were invited to participate in the study. Teachers who agreed to participate gave their informed consent for classroom observations, completed personal and classroom demographic surveys, and assisted with obtaining consent from parents regarding their children's participation. TAs who agreed to participate sent in ratings of children's behavior and completed a personal demographic survey. Once full informed consent was secured from lead teachers, they facilitated in obtaining parental consent from all children in the classroom. The parental consent rate for the study averaged 76%.

Selection of children. Teachers and TAs completed the ADHD Rating Scale (ADHD-RS-IV; DuPaul, Power, Anastopoulos, & Reid, 1998) and ODD Rating Scale (ODDRS; Anastopoulos, 1998—see measures section for additional details) on all children in their classroom. The participating teachers filled out a total of 2,379 behavior rating scales (10 of which had omitted items that prevented the generation of a summed score) on all students in their classrooms, and ultimately 471 children were selected to participate for the larger intervention. The three children (2 of those selected were boys, and 1 was a girl) with the highest levels of lead teacher-reported disruptive behaviors were selected to be in the study (for site 1, only 2 children were selected per classroom). Children selected to participate in the intervention demonstrated significantly more externalizing behavior (M = 28.48) than children who were not selected for the intervention (M = 10.82) based on lead teachers' summed scores of all items from the ADHD-RS-IV and the ODDRS, $t(2,369) = 24.831$, $p < .001$. Only those children who were selected for the intervention were observed in the classroom, and thus this group of children comprises the sample for the current study. As stated previously, children in the current study's sample had both teacher and TA ratings (n = 360) and are thus a subsample of the children selected for the larger intervention.

Direct observation training. Data collectors were required to complete an extensive 2-day training from a certified trainer on the observation measure used in the study (Individualized Classroom Assessment Scoring System [inCLASS]; Downer, Booren, Hamre, Pianta, & Williford, 2011) prior to data collection. The training included a review of the content of the measures and required data collectors to code, watch, and discuss training clips. At the end of the training, data collectors were required to reliably code five clips by scoring within 1 point of a mastercode on 80% of the scheme's dimensions. Across the sites, 43 data collectors conducted the observations. Data collectors' initial reliability scores ranged from 80% to 94%. Following successful training but before data collection began, data collectors practiced the coding scheme on children in preschool classrooms that were not part of the study with a master trainer. In addition, data collectors watched and coded practice videos to ensure that they had not drifted from initial reliability. Interrater reliability was estimated by double coding 20% of all field observations; these estimates are provided in the description of the inCLASS measure below.

Observation protocol. Observations were scheduled during the first 6 weeks of school within 2 weeks of receiving the lead teachers' rating scales. During this time, teachers and children were observed in the classroom setting. Each observation day lasted approximately 4 hr from the start of the day until mid-day and occurred over multiple days. Data collectors observed the selected children in a series of alternating cycles starting at the beginning of the school day; each cycle consisted of observing a child for 10 min and then coding the observation for 5 min. Data collectors shifted their observation across the two or three selected children (i.e., they observed Child 1, Child 2, then Child 3,

and began again with Child 1; on the next day children were observed in a different repeating order, such as Child 2, Child 3, then Child 1), with the goal of collecting at least eight cycles per child across 2 days. The participants in the current study were observed for approximately 9 cycles ($M = 8.66$, $SD = 1.79$) across 3 days ($M = 3.11$, $SD = 0.84$).

Measures

Teacher and TA ratings of externalizing behavior. Teachers and TAs completed a survey that contained all items from the ADHD-RS-IV (DuPaul et al., 1998) as well as the ODDRS (Anastopoulos, 1998; Hommersen, Murray, Ohan, & Johnston, 2006) to screen for externalizing behavior to select children for the larger intervention. Both are psychometrically sound behavior rating scales often used in clinical research with preschool-aged children (Barkley et al., 2000; DuPaul, McGoey, Eckert, & VanBrakle, 2001; Johnston, Hommersen, & Seipp, 2009) and have been shown to be valid and reliable in the preschool population (McGoey, DuPaul, Haley, & Shelton, 2007). The ADHD-RS-IV is an 18-item scale based upon the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM-IV*; American Psychiatric Association, 1994) symptoms of ADHD. Nine items correspond to symptoms of inattention and 9 items measure hyperactivity/impulsivity. Modeled after the ADHD-RS-IV, the ODDRS contains 8 items corresponding to the *DSM-IV* criteria for ODD. Scores were summed for each individual dimension of externalizing behavior, resulting in a score for Hyperactivity/Impulsivity (HI, $\alpha = .93$) Inattention (IA, $\alpha = .92$), and ODD ($\alpha = .93$). In addition, scores across all dimensions were summed to create a Total Externalizing score ($\alpha = .96$). To demonstrate concurrent construct validity for this measure in the current project, correlations between the ADHD-RS-IV and ODDRS rating scales and the Sutter-Eyberg Student Behavior Inventory–Revised (SESBI-R; Querido & Eyberg, 2004) were examined. The SESBI-R is a widely used, comprehensive, behaviorally specific rating scale that assesses disruptive behaviors in the school setting. The correlations of the ADHD-RS and ODDRS summary scores with the intensity score of the SESBI-R were all significant at the $p < .001$ level (HI, $r = .58$; IA, $r = .46$, ODD, $r = .55$; Total Externalizing, $r = .65$) and were comparable in strength to those found in other research (McGoey et al., 2007). Children were selected for the larger intervention study based on lead teachers' Total Externalizing score.

Observed externalizing behavior. The inCLASS (Downer et al., 2011) is an observational assessment of young children's behavior during everyday interactions with teachers, peers, and tasks in preschool classrooms. The inCLASS measures 10 dimensions of children's behavior on a 7-point scale, including (a) positive engagement with teacher, (b) teacher conflict, (c) teacher communication, (d) peer sociability, (e) peer conflict, (f) peer assertiveness, (g) peer communication, (h) engagement within tasks, (i) self-reliance, and (j) behavior control. Trained observers watch children for 10 min, and immediately following this period, they rate children's positive or negative patterns of behavior based on the child's display of clearly defined behavioral indicators that categorize each dimension. Children are observed over multiple cycles (in this case, an average of 9 cycles across 3 days) to estimate a child's typical behavior pattern displayed in the classroom. In validation studies, the inCLASS has shown construct and criterion validity (Downer, Booren, Lima, Luckner, & Pianta, 2010) in addition to predicting children's self-regulation and language and literacy skills (Williford, Maier, Downer, Pianta, & Howes, 2013; Williford, Vick Whittaker, Vitiello, & Downer, 2013).

Only the dimensions that were conceptually linked to teacher-rated externalizing behaviors were used for the present analyses. Behavior control was reverse scored and labeled observed impulsivity. Behavioral indicators for low behavior control/high observed impulsivity include interrupting, going out of turn, shouting out, speaking loudly, engaging in extra movement, intruding on the personal space of others, and often bumping into objects or other children. Engagement within tasks was reverse scored and labeled observed inattention. Behavioral indicators for low engagement/high observed inattention include engaging in activities that are not assigned, receiving frequent redirections from the teacher related to the activity, changing activities often, wandering around the room, spending time watching others rather than engaging in the activity, and engaging in activities lethargically or repetitively. Conflict with teachers and conflict with peers were averaged together and labeled observed conflict. Children with high conflict scores engage in verbal or physical aggression, such as hitting, kicking, pushing, yelling, and name calling; exhibit non-compliance, such as defying request and provoking arguments; show negative affect, such as frowning, grimacing, or folding hands across the chest; and engage in attention-seeking behaviors, such as whining, complaining, or pouting. Observed impulsivity, observed inattention, and observed conflict were averaged together to create an observed total externalizing score.

Interrater reliability was calculated across 20% of all observations with two data collectors independently observing and rating the same children. Intraclass correlations (ICCs), exact agreement, and agreement within 1 point (the measure developer's benchmark for reliability during the training) were as follows: behavior control = .75, 52%, 91%; engagement within tasks = .69, 41%, 85%; teacher conflict = .63, 84%, 98%; and peer conflict = .72, 81%, 97%. Note that the ICCs for teacher and peer conflict were low due to the skewness of the data (conflict is a low

Table 1. Descriptive Statistics: Teacher and TA Ratings, Disagreement, and Observed Disruptive Behavior.

Variables	*n*	Minimum	Maximum	*M*	*SD*
Teacher and TA ratings					
Teacher HI	360	0.00	27.00	11.75	6.99
TA HI	360	0.00	27.00	10.58	7.26
Teacher IA	359	0.00	27.00	10.11	6.49
TA IA	360	0.00	27.00	8.41	6.97
Teacher ODD	360	0.00	24.00	6.77	6.20
TA ODD	359	0.00	24.00	5.87	6.40
Teacher total	360	0.00	78.00	28.57	16.39
TA total	359	0.00	75.00	24.86	17.95
Agreement					
HA absolute difference	360	0.00	19.00	5.57	4.41
IA absolute difference	359	0.00	20.00	5.08	4.28
ODD absolute difference	359	0.00	22.00	4.18	4.11
Total absolute difference	359	0.00	46.00	13.03	10.79
Observed externalizing behavior					
Observed impulsivity	338	1.00	5.88	2.57	0.88
Observed inattention	338	1.33	6.00	3.30	0.82
Observed conflict	338	1.00	3.42	1.31	0.36
Observed total externalizing	338	1.34	4.54	2.39	0.57

Note. Higher absolute difference scores reflect lower agreement. TA = teacher aide; HI = hyperactivity/impulsivity; IA = inattention; ODD = oppositional defiant disorder.

occurring event). In addition, internal consistencies for dimensions across observation periods were calculated. Internal consistencies were as follows: Observed Impulsivity ($\alpha = .84$), Observed Inattention ($\alpha = .71$), Observed Conflict ($\alpha = .75$), and Observed Total Externalizing ($\alpha = .81$).

Data Analysis and Missing Data

To test the relation between teacher and TA ratings and their associations with direct observations, hierarchical linear models were used to account for the organizational nature of the data where children (Level 1) were nested in classrooms (Level 2). We used the Type = Complex Command in MPlus, which adjusts standard errors to take into account the fact that children were clustered within classrooms. In addition, we examined ICCs, which for TA ratings ranged from .14 (ODD) to .25 (IA), and for teacher ratings ranged from .24 (ODD) to .37 (Total Externalizing). Descriptive statistics and pairwise correlations were examined in SPSS Version 21, and all other models were analyzed in Mplus Version 6 (Muthén & Muthén, 1998-2011). Absolute difference scores between teacher and TA ratings of HI, IA, ODD, and Total Externalizing scales were calculated to estimate teacher and TA agreement. There are multiple ways of calculating rater agreement including using raw difference scores, standardized difference scores, and residual difference scores (De Los Reyes & Kazdin, 2004); we used absolute raw difference scores because we did not have hypotheses related to the direction of differences, and the distributions for the teacher and TA variables were similar.

Missing data in the current analyses included those children who did not have baseline inCLASS observational outcome data ($n = 22$) and two children for whom survey data were incomplete ($n = 2$) due to a teacher omitting more than two items on the inattention subscale and a TA omitting more than two items on the ODD subscale. We used full information maximum likelihood estimation with robust standard errors to make use of all available data for each case (Enders & Bandalos, 2001).

Results

Descriptive Statistics and Teacher and TA Agreement

Descriptive statistics are presented in Table 1 and Correlations are presented in Table 2. As Table 1 shows, on average across all scales, teachers rated children's externalizing behavior slightly higher than TAs, although the distributions of teacher and TA ratings were similar. Pairwise correlations between teacher and TA ratings of externalizing behavior dimensions indicated that teacher and TA ratings were moderately correlated between 0.52 (HI) and 0.57 (ODD; see Table 2).

Severity of Externalizing Behavior and Agreement

To test whether teacher–TA agreement was greatest at low and high ends of externalizing behavior, we created a

Table 2. Correlations Among Teacher Ratings, TA Ratings, and Observed Externalizing Behavior.

Variables	1	2	3	4	5	6	7	8	9	10	11	12
1. Teacher HI	1.00	.516**	.665**	.343**	.556**	.401**	.900**	.486**	.248**	.104	.205**	.218**
2. TA HI			.386**	.732**	.335**	.665**	.500**	.927**	.258**	.106	.209**	.225**
3. Teacher IA				.546**	.383**	.239**	.824**	.453**	.132*	.151**	.140**	.168**
4. TA IA					.172**	.487**	.428**	.859***	.151**	.176*	.121*	.185**
5. Teacher ODD						.574**	.767**	.406**	.221**	.074	.313**	.213**
6. TA ODD							.482**	.815**	.207**	.104	.203**	.197**
7. Teacher total								.540**	.243**	.133*	.263**	.241**
8. TA total									.238**	.149**	.205**	.235**
9. Observed impulsivity										.533**	.660**	.898**
10. Observed inattention											.395**	.826**
11. Observed conflict												.730**
12. Observed total externalizing												1.00

Note. TA = teacher aide; HI = hyperactivity/impulsivity; IA = inattention; ODD = oppositional defiant disorder.
*$p < .05$. **$p < .01$. ***$p < .001$.

squared term for teacher ratings of HI, IA, ODD, and Total Externalizing behavior. We then ran four separate regressions and regressed teacher ratings of externalizing behavior and the corresponding squared term onto teacher/TA absolute difference scores of HI, IA, ODD, and Total Externalizing scores. The squared term for ODD was significant (β = −.53, *SE* = 0.23, *p* = .025) indicating that teacher–TA agreement was higher at the tails (when lead teachers rated children as displaying either low or high oppositional behavior) compared with when lead teachers rated children as displaying more moderate levels of oppositional behaviors. The coefficient for the squared terms of HI (β = .07, *SE* = .24, *p* = n.s.); IA (β = −.03, *SE* = .17, *p* = n.s.); and Total Externalizing scores (β = −.04, *SE* = 0.09, *p* = n.s.) were all non-significant.

Agreement and Its Relation to Observed Externalizing Behavior

To assess whether teacher/TA agreement was associated with observed externalizing behavior over and above the lead teacher's ratings, teacher ratings and teacher/TA absolute difference scores were regressed onto corresponding observed externalizing behavior dimensions. Lead teachers' ratings, rather than the TAs', were used in models because lead teacher ratings are commonly used during the assessment process. Four separate models assessing teacher/TA agreement of HI (β = .039, *p* = n.s.); IA (β = −.069, *p* = n.s.); ODD (β = −.036, *p* = n.s.); and Total Externalizing Scores (β = −.035, *p* = n.s.) were run separately. Analyses demonstrated that teacher/TA agreement did not significantly predict observed externalizing behavior for any dimensions; however, teacher ratings alone continued to significantly predict observed externalizing behavior (see Table 3).

Table 3. Teacher Ratings and Teacher/TA Agreement Predicting Observed Externalizing Behavior.

Regressions	β	*SE*
Observed impulsivity		
Teacher HI	.243***	.057
Absolute difference HI	.039	.060
Observed inattention		
Teacher IA	.181*	.067
Absolute difference IA	−.069	.062
Observed conflict		
Teacher ODD	.326***	.057
Absolute difference ODD	−.036	.055
Observed total externalizing		
Teacher total	.256***	.060
Absolute difference total	−.035	.060

Note. TA = teacher aide; HI = hyperactivity/impulsivity; IA = inattention; ODD = oppositional defiant disorder.
*$p < .05$. ***$p < .001$.

Using Teacher and TA Ratings Together

To examine whether teacher and TA ratings entered into a model together were more associated with observed externalizing behavior than using a single informant's report alone, we investigated models that included teacher ratings alone, TA ratings alone, or their ratings entered into a model together to predict each externalizing behavior outcome. We examined the significance of the coefficients in each of these models to better understand whether estimates of children's externalizing behavior using two raters would be more strongly associated with observed ratings as compared with using a single rater's estimate. We also examined R^2 statistics, which were generated in MPlus, to understand how much variance in observed externalizing behavior could be accounted for by teacher and TA ratings

Table 4. Results of Separate Regressions Examining Teacher Ratings, TA Ratings, and Teacher and TAs Together Predicting Observed Externalizing Behavior.

	Teacher ratings only			TA ratings only			Teacher and TA ratings		
Variables	β	*SE*	R^2	β	*SE*	R^2	β	*SE*	R^2
Observed Impulsivity									
Teacher HI	.232***	.060	.053	—	—	—	.161*	.072	.068*
TA HI	—	—		.219***	.058	.048	.137*	.048	
Observed Inattention									
Teacher IA	.147*	.060	.022	—	—	—	.107	.074	.027
TA IA	—	—		.137*	.063	.019	.081	.076	
Observed Conflict[a]									
Teacher ODD	.318***	.052	.101**	—	—	—	.289***	.067	.103**
TA ODD	—	—		.217***	.054	.045*	.050	.068	
Observed Total Externalizing									
Teacher total	.207**	.066	.043	—	—	—	.148	.076	.052
TA total	—	—		.189**	.059	.036	.109	.067	

Note. R^2 values were calculated in Mplus. TA = teacher aide; HI = hyperactivity/impulsivity; IA = inattention; ODD = oppositional defiant disorder.
[a]Teacher and TA coefficients were significantly different based on Wald Tests.
*$p < .05$. **$p < .01$. ***$p < .001$.

alone and in combination. We found that teacher and TA ratings significantly predicted observed externalizing behavior independently, but when entered in the model together, both ratings were significant only for observed impulsivity (see Table 4). Across all models, teachers alone, TAs alone, and their ratings in combination accounted for small amounts of variance in observed externalizing behavior (R^2 = .019–.103).

Although the standardized coefficient (β) was stronger for the teacher's ratings entered alone when compared with the TA's rating entered alone across all models, we used Wald tests to determine whether teachers' ratings were significantly more associated with observed externalizing behavior than TAs' ratings. The Wald test for ODD was significant (Wald = 3.781, p = .05) indicating a difference in teachers' and TAs' β coefficients such that lead teachers' ratings of children's oppositionality were more closely aligned with observed ratings of conflict in the classroom compared with TAs' ratings of oppositionality. Wald tests of coefficient equivalence were non-significant for ratings of HI (Wald = 0.051, p = n.s.); IA (Wald = 0.06, p = n.s.); and Total Externalizing behavior (Wald = 0.159, p = n.s.).

Discussion

The present study sought to better understand rater agreement and investigated how to integrate assessment information gathered from multiple informants and methods. First, we examined teacher and TA ratings of externalizing behavior to see whether two adults—both of whom have comparable roles, interact with children in the same context, and do so for the same amount of time each day—evaluate children's classroom behavior similarly. We also examined whether teacher/TA agreement was greater when lead teachers rated children as evidencing low or high levels of externalizing behavior. We then tested whether agreement between teachers and TAs predicted observed externalizing behavior over and above lead teachers' ratings alone. We further examined how teacher and TA ratings were associated with children's observed classroom behavior by investigating whether teacher, TA, or both of their ratings were better predictors of observed externalizing behavior.

We found that teacher and TA reports of preschoolers' externalizing behavior were moderately correlated between .52 (HI) and .57 (ODD). This is slightly lower but comparable with associations reported by other studies that have investigated agreement between adults with similar roles (average r = .60; Achenbach, 2006). Although the correlations that we found are higher than those between adults with different roles (e.g., parents and teachers; average r = .28; Achenbach et al., 1987), the correlations from our study are still too low to ensure that a similar picture of a child was obtained from each teacher (Achenbach, 2006)—at the most, 32% of the variance in teacher and TA ratings were shared. The differences in teachers and TAs ratings may exist as a function of the context and relationships with which individual teachers interact with children (De Los Reyes et al., 2009). For example, teachers may typically observe all children's behavior during circle time and whole group activities, whereas the TA may be preparing materials for the next activity or working individually with a single challenging child during this period. An alternate

explanation could be that rater characteristics, such as expectations for behavior, or psychological and demographic variables, contribute to teachers' differential assessments of children's behavior (Mashburn, Hamre, Downer, & Pianta, 2006). Ultimately, these results highlight that while one teacher may find a child's behavior particularly problematic, another teacher in the same classroom may perceive a child's behavior quite differently. This reflects the importance of gathering input from multiple individuals within the preschool setting when assessing the severity of a child's behavior.

Our tests of whether teacher/TA agreement was greater when children were rated as evidencing low or high levels of externalizing behavior by the lead teacher were significant only for ratings of ODD. Our findings regarding the relation between the severity of teacher ratings and agreement for inattention, hyperactivity/impulsivity, and total externalizing behavior ran counter to hypotheses proposed in the literature around the severity of behavior and rater agreement (e.g., De Los Reyes, 2011). At the same time, given that ratings of ODD followed the pattern we expected, it may be that elevated levels of defiance and aggression are more salient and less normative in preschool children when compared with hyperactivity/impulsivity or inattention, which are prevalent in this age group (Campbell, 2002). As such, teachers' definitions of acceptable or typical levels of hyperactivity and inattention may be quite variable. On the other hand, demonstrations of oppositionality and aggression are more likely to be noticed and identified as problematic by multiple individuals; although displays of aggression are not unusual during this period, it is rare for a child to demonstrate these behaviors consistently (Wakschlag et al., 2010).

We found that after controlling for lead teachers' ratings of externalizing behavior, teacher/TA agreement did not predict observed externalizing behavior. This finding was contrary to our expectations, as we had hypothesized that agreement would be indicative of a more objective evaluation of what was occurring in the classroom. There may be several explanations for these null findings. First, we found low associations between observed externalizing behavior and teacher and TA ratings in general. It may be that these measures are tapping different constructs (Thomas et al., 2011), and therefore agreement between two teachers would not necessarily relate to observational measures. Second, as stated previously, agreement may be a product of similarities between raters (e.g., expectations for acceptable behavior, interaction styles; De Los Reyes et al., 2009) rather than being indicative of what is captured by independent observer.

The finding that agreement does not relate to observational measures points to the importance of gathering both observational assessments and using teacher report, as even when raters agree about the severity of children's behavior, this does not appear to be indicative of the information captured by an independent observer. While agreement in this study was not concurrently predictive of observational measures, future research should examine whether agreement between multiple teachers is related to important outcomes, similar to the work that has been done with parent and self-report (for a review, see De Los Reyes, 2011). As an example, if teacher/TA disagreement does in fact reflect teachers' differing views of preschoolers' externalizing behavior, these differences may impact the consistency with which adults in the classrooms implement interventions aimed to improve externalizing behaviors. This may ultimately affect the extent to which preschoolers' externalizing classroom behavior improves or worsens over time.

In examining whether using both teacher and TA ratings provided a better estimate of observed externalizing behavior compared with using the teacher or TA alone, we found that for all dimensions except HI, using both informants was not a better estimate than using a single informant's report only. We also found that the amount of variance accounted for in observed behavior by both raters individually and in combination was small. Furthermore, even though both teacher and TA ratings were significant when entered into the same model for HI, practically speaking, their reports still accounted for little variance in observed impulsivity. These results were consistent with the literature (e.g., Carter et al., 2004; Thomas et al., 2011) and suggest that indirect ratings appear to be tapping into a different construct than observational measures.

Finally, we found no significant differences in the relation between teachers' ratings and observed externalizing behavior and TAs' ratings and observed externalizing behavior, except for ODD. The relation between teacher ratings of ODD and observed conflict was significantly stronger compared with the relation between TA ratings and observed conflict. This difference may be a product of the varying roles that teachers and TAs play in the classroom. Studies of TAs in primary and secondary school classrooms indicate that TAs are increasingly interacting with children in small group settings, as well as supporting children who have special needs (Rubie-Davies, Blatchford, Webster, Koutsoubou, & Bassett, 2010). If TAs are in fact working with children in smaller, more individualized settings in preschool, and lead teachers tend to be involved in directing whole group instruction and transitions, children may be more likely to display difficult behaviors given the high attentional and behavioral demands of whole group and transition activities. More research will be important in understanding and testing this hypothesis.

The present study has several limitations. First, rating scales were collected at the beginning of the school year, and therefore teachers were still becoming familiar with children as they adjusted to the classroom environment. Second, rating scales were clinical in nature and corresponded to

DSM-IV symptoms of ADHD and ODD, and the applicability of these disorders for preschool children has been debated (e.g., Carter et al., 2004). However, the ADHD-RS-IV has been shown to provide useful diagnostic information for preschool age children (Purpura, Wilson, & Lonigan, 2010) and the stability of externalizing behaviors has been documented in this population (e.g., Egger & Angold, 2006). Still, because rating scales were designed for clinical purposes, they may have been less appropriate for use in a community sample with preschool teachers. One could also argue that the small amounts of variance explained by our ratings could have been related to the observational approach used in this study. However, a strength of our study was the amount of time each child was observed (~90 min) across multiple days and a variety of activity settings. Even so, it should be noted that some of the behaviors assessed in this study are low base rate behaviors (non-compliance, aggression, oppositionality) even for children who are identified as challenging to manage in the classroom (Pelham et al., 2005). Finally, the selection of children for the study—which was based on teacher reports of externalizing behavior only—may have influenced levels of agreement between teachers and TAs. Selecting children using alternative criteria could have led to greater agreement between teachers and TAs and a stronger relation between teacher/TA agreement and observed behavior. Despite this limitation in our selection, the fact remains that using lead teachers' reports of problematic behavior is practically meaningful. Lead teachers are often involved in the expulsion of preschool children and their referral for services (Gilliam, 2005), highlighting the critical associations between lead teachers' perceptions and important child outcomes. Thus, gaining multiple perspectives on the externalizing behavior of children who are identified by the lead teacher has important implications in understanding how assessments and agreement align for these children.

In conclusion, our study furthered the work that focuses on understanding multi-informant and multi-method assessment by investigating this phenomenon in an early childhood sample with adult raters who have the same role and interact with children in the same setting. These results have implications both for future research and for early childhood educators. This work highlights that much remains unknown about why discrepancies among teacher, TA, and observational measures occur. For example, are discrepancies measure-specific, such that a different measure might provide more consistent estimates of a child's functioning? Or, in line with other work (e.g., De Los Reyes et al., 2009), does the context of interactions contribute to disagreement, such that each individual teacher and activity setting represents a unique context? Or finally, is agreement linked to rater characteristics, such that teachers with similar beliefs, psychological characteristics, and/or demographic characteristics are likely to evaluate children more uniformly? Although studies have examined factors associated with differences in parents' and teachers' reports of preschoolers externalizing behavior (e.g., Dinnebeil et al., 2013; De Los Reyes et al., 2009), future work should investigate classroom and teacher factors (e.g., demographics, psychological characteristics) that are associated with these differences. Observational classroom measures that explicitly track behaviors that occur with a teacher versus a TA will be helpful in uncovering whether meaningful variations occur in children's classroom behavior depending on whether the child is interacting with a teacher versus a TA.

For early childhood educators, this work shows that it is important for teachers and TAs to refrain from assuming that they view children's externalizing behavior similarly. Even when teachers and TAs do agree about behavior, an independent, impartial observer is likely to provide additional information that may not be captured by teachers' and TAs' reports. Rather than assuming that one teacher's report or a direct observation's results represent "the truth," it may be helpful to gather more information about the underlying reasons for discrepancies from teachers and observers themselves. This, in turn, might illuminate the type of situations and relationships that evoke particular child behaviors. If children's externalizing behavior is less severe in particular contexts or with particular teachers, providing children with linked supports could be helpful in reducing problematic behavior in the classroom. Because children who exhibit externalizing behavior in the preschool classroom are at risk for a variety of negative outcomes, gaining a better understanding of how multiple individuals and methods evaluate children's behavior will be critical for assessment, and in turn, effective early intervention systems for young children.

Acknowledgment

We extend our gratitude to the teachers, parents, and children who invited us into their classrooms.

Authors' Note

The opinions expressed are those of the authors and do not represent views of the funders.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A100215 awarded to the second author and the University of Virginia.

References

Abikoff, H., Courtney, M., Pelham, W. E. Jr., & Koplewicz, H. S. (1993). Teachers' ratings of disruptive behaviors: The influence of halo effects. *Journal of Abnormal Child Psychology*, *21*, 519–533.

Achenbach, T. M. (2006). As others see us. Clinical and research implications of cross informant correlations for psychopathology. *Current Directions in Psychological Science*, *15*, 94–98.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations f or situational specificity. *Psychological Bulletin*, *101*, 213–232.

Adams, G., Tout, K., & Zaslow, M. (2007). *Early care and education for children in low-income families patterns of use, quality, and potential policy implications*. Washington, DC: The Urban Institute and Child Trends.

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

Anastopoulos, A. D. (1998). *Oppositional defiant disorder/conduct disorder rating scale*. Unpublished manuscript, University of North Carolina at Greensboro.

Atkins, M. S., & Pelham, W. E. (1991). School-based assessment of attention deficit-hyperactivity disorder. *Journal of Learning Disabilities*, *24*, 197–204.

Barkley, R. A., Shelton, T. L., Crosswait, C., Moorehouse, M., Fletcher, K., Barrett, S., & Metevia, L. (2000). Multi-method psycho-educational intervention for preschool children with disruptive behavior: Preliminary results at post-treatment. *Journal of Child Psychology and Psychiatry*, *41*, 319–332.

Calkins, S. D., Gill, K., & Williford, A. (1999). Externalizing problems in two-year-olds: Implications for patterns of social behavior and peers' responses to aggression. *Early Education and Development*, *10*, 267–288.

Campbell, S. B. (2002). *Behavior problems in preschool children clinical and developmental issues*. New York, NY: Guilford Press.

Carter, A. S., Briggs-Gowan, M. J., & Davis, N. O. (2004). Assessment of young children's social-emotional development and psychopathology: Recent advances and recommendations for practice. *Journal of Child Psychology and Psychiatry*, *45*, 109–134.

De Los Reyes, A. (2011). Introduction to the special section: More than measurement error: Discovering meaning behind informant discrepancies in clinical assessments of children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *40*, 1–9.

De Los Reyes, A., Henry, D. B., Tolan, P. H., & Wakschlag, L. S. (2009). Linking informant discrepancies to observed variations in young children's disruptive behavior. *Journal of Abnormal Child Psychology*, *37*, 637–652.

De Los Reyes, A., & Kazdin, A. E. (2004). Measuring informant discrepancies in clinical child research. *Psychological Assessment*, *16*, 330–334.

Dinnebeil, L. A., Sawyer, B. E., Logan, J., Dynia, J. M., Cancio, E., & Justice, L. M. (2013). Influences on the congruence between parents' and teachers' ratings of young children's social skills and problem behaviors. *Early Childhood Research Quarterly*, *28*, 144–152.

Downer, J. T., Booren, L. M., Hamre, B., Pianta, R. C., & Williford, A. (2011). *The Individualized Classroom Assessment Scoring (inCLASS)*. Unpublished technical manual, Curry School of Education, University of Virginia, Charlottesville.

Downer, J. T., Booren, L. M., Lima, O. K., Luckner, A. E., & Pianta, R. C. (2010). The Individualized Classroom Assessment Scoring System (inCLASS): Preliminary reliability and validity of a system for observing preschoolers' competence in classroom interactions. *Early Childhood Research Quarterly*, *25*, 1–16.

Dunlap, G., Strain, P. S., Fox, L., Carta, J. J., Conroy, M., Smith, B. J., & Sowell, C. (2006). Prevention and intervention with young children's challenging behavior: Perspectives regarding current knowledge. *Behavioral Disorders*, *32*, 29–45.

DuPaul, G. J., McGoey, K. E., Eckert, T. L., & VanBrakle, J. (2001). Preschool children with attention-deficit/hyperactivity disorder: Impairments in behavioral, social, and school functioning. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*, 508–515.

DuPaul, G. J., Power, T. J., Anastopoulos, A. D., & Reid, R. (1998). *ADHD Rating Scale—IV: Checklists, norms, and clinical interpretation*. New York, NY: Guilford Press.

Egger, H. L., & Angold, A. (2006). Common emotional and behavioral disorders in preschool children: Presentation, nosology, and epidemiology. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *47*, 313–337.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*, 430–457.

Essa, E. (2010). *Introduction to early childhood education*. Belmont, CA: Wadsworth Cengage Learning.

Frede, E., & Barnett, W. S. (1992). Developmentally appropriate public school preschool: A study of implementation of the High/scope curriculum and its effects on disadvantaged children's skills at first grade. *Early Childhood Research Quarterly*, *7*, 483–499.

Gilliam, W. S. (2005). *Prekindergarteners left behind: Expulsion rates in state prekindergarten systems*. New Haven, CT: Child Study Center, Yale University.

Hinshaw, S. P. (1992). Academic underachievement, attention deficits, and aggression: Comorbidity and implications for intervention. *Journal of Consulting and Clinical Psychology*, *60*, 893–903.

Hommersen, P., Murray, C., Ohan, J. L., & Johnston, C. (2006). Oppositional Defiant Disorder Rating Scale: Preliminary evidence of reliability and validity. *Journal of Emotional and Behavioral Disorders*, *14*, 118–125.

Jackson, D. A., & King, A. R. (2004). Gender differences in the effects of oppositional behavior on teacher ratings of ADHD symptoms. *Journal of Abnormal Child Psychology*, *32*, 215–224.

Johnston, C., Hommersen, P., & Seipp, C. M. (2009). Maternal attributions and child oppositional behavior: A longitudinal study of boys with and without attention- deficit/hyperactivity disorder. *Journal of Consulting and Clinical Psychology*, *77*, 189–195.

Mashburn, A. J., Hamre, B. K., Downer, J. T., & Pianta, R. C. (2006). Teacher and classroom characteristics associated

with teachers' ratings of prekindergartners' relationships and behaviors. *Journal of Psychoeducational Assessment*, *24*, 367–380.

McGoey, K., DuPaul, G. J., Haley, E., & Shelton, T. L. (2007). Parent and teacher ratings of attention-deficit/hyperactivity disorder in preschool: The ADHD rating scale-IV preschool version. *Journal of Psychopathology and Behavioral Assessment*, *29*, 269–276.

Miller, M. L., Fee, V. E., & Jones, C. J. (2004). Psychometric properties of ADHD rating scales among children with mental retardation. *Research in Developmental Disabilities*, *25*, 477–492.

Muthén, L. K., & Muthén, B. O. (1998–2011). *Mplus user's guide* (6th ed.). Los Angeles, CA: Author.

Pelham, W. E., Fabiano, G. A., & Massetti, G. M. (2005). Evidence-based assessment of attention deficit hyperactivity disorder in children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, *34*, 449–476.

Purpura, D. J., Wilson, S. B., & Lonigan, C. J. (2010). Attention-deficit/hyperactivity disorder symptoms in preschool children: Examining psychometric properties using item response theory. *Psychological Assessment*, *22*, 546–558.

Qi, C. H., & Kaiser, A. P. (2004). Problem behaviors of low-income children with language delays: An observation study. *Journal of Speech, Language, and Hearing Research*, *47*, 595–609.

Querido, J. G., & Eyberg, S. M. (2004). Psychometric properties of the Sutter-Eyberg Student Behavior Inventory–Revised with preschool children. *Behavior Therapy*, *34*, 1–15.

Rubie-Davies, C., Blatchford, P., Webster, R., Koutsoubou, M., & Bassett, P. (2010). Enhancing learning? A comparison of teacher and teaching assistant interactions with pupils. *School Effectiveness and School Improvement*, *21*, 429–449.

Strickland, J., Hopkins, J., & Keenan, K. (2012). Mother-teacher agreement on preschoolers' symptoms of ODD and CD: Does context matter? *Journal of Abnormal Child Psychology*, *40*, 933–943.

Thomas, L. B., Shapiro, E. S., DuPaul, G. J., Lutz, J. G., & Kern, L. (2011). Predictors of social skills for preschool children at risk for ADHD: The relationship between direct and indirect measurements. *Journal of Psychoeducational Assessment*, *29*, 114–124.

Wakschlag, L. S., Tolan, P. H., & Leventhal, B. L. (2010). Research Review: "Ain't misbehavin": Towards a developmentally-specified nosology for preschool disruptive behavior. *Journal of Child Psychology and Psychiatry*, *51*, 3–22.

Williford, A. P., Maier, M. F., Downer, J. T., Pianta, R. C., & Howes, C. (2013). Understanding how children's engagement and teachers' interactions combine to predict school readiness. *Journal of Applied Developmental Psychology*, *34*, 299–309.

Williford, A. P., Vick Whittaker, J. E., Vitiello, V. E., & Downer, J. T. (2013). Children's engagement within the preschool classroom and their development of self-regulation. *Early Education & Development*, *24*, 162–187.

Wolraich, M. L., Lambert, E. W., Bickman, L., Simmons, T., Doffing, M. A., & Worley, K. A. (2004). Assessing the impact of parent and teacher agreement on diagnosing attention-deficit hyperactivity disorder. *Developmental and Behavioral Pediatrics*, *25*, 41–48.