Psychology in the Schools, Vol. 48(6), 2011
View this article online at wileyonlinelibrary.com/journal/pits

DOI: 10.1002/pits.20577

EXAMINING THE VALIDITY OF OFFICE DISCIPLINE REFERRALS AS AN INDICATOR OF STUDENT BEHAVIOR PROBLEMS

ELISE T. PAS, CATHERINE P. BRADSHAW, AND MARY M. MITCHELL

Johns Hopkins University

Office discipline referral (ODR) data are increasingly used to monitor student behavior problems and the impact of interventions, but there has been limited research examining their validity. The current study examined the concordance of ODRs with teacher ratings of student behavior using data on 8,645 children in 335 classrooms at 21 elementary schools. The results of a variety of analyses (e.g., correlations, multivariate analysis of variance, receiver operating characteristics) suggested that ODRs are moderately valid and reliable. Multilevel analyses revealed that teacher ratings of disruptive behaviors were significantly associated with ODRs, even after controlling for other student-, classroom-, and school-level factors. These findings suggest that ODRs are moderately valid indicators of student behavior problems and may be an efficient source of information for use in school-based research and data-based decision-making. © 2011 Wiley Periodicals, Inc.

Office discipline referrals (ODRs) have been defined as events in which a staff member observes a student violating a school rule and submits documentation of the event to the administrative leadership, who then delivers a consequence to the student (Irvin et al., 2006). ODRs are often used to evaluate the impact of school-based interventions and policies (Sugai, Sprague, Horner, & Walker, 2000), monitor student problem behaviors (Irvin, Tobin, Sprague, Sugai, & Vincent, 2004), and make programmatic decisions regarding student support services (Irvin et al., 2006; Morgan-D'Atrio, Northup, LaFleur, & Spera, 1996; Nelson, Benner, Reid, Epstein, & Currin, 2002). Consistent with a Response to Intervention (RTI) approach, ODRs may also be informative in identifying children in need of more intensive preventive interventions (Hawken, Vincent, & Schumann, 2008). However, some researchers have questioned the validity of ODRs as an indicator of student problem behaviors. For example, school discipline and referral practices vary both across and within schools (Morrison, Redding, Fisher, & Peterson, 2006). Furthermore, the referral process can be somewhat subjective (Irvin et al., 2004) and thus may be influenced by a student's history of behavior problems or race/ethnicity (Skiba et al., 2008).

The current study aimed to address some of the concerns regarding the validity of ODRs by comparing two sources of ODR data to one another and to other previously validated measures of behavior. We also explored potential student-, classroom-, and school-level influences on students' likelihood of receiving an ODR. Having an enhanced understanding of the validity of ODRs will inform the use of ODRs in research, clinical practice, and data-based decision-making (Irvin et al., 2006). This work also has important implications for the potential use of ODRs as a screening tool for children in need of programs and services to address their problem behaviors (Hawken et al., 2008; Sugai et al., 2000).

Prior Research on the Validity of ODRs

Prior research indicates that receipt of an ODR has a moderate association with poor academic achievement ($r = .34$; McIntosh, Flannery, Sugai, Braun, & Cochrane, 2008) and a small to moderate

This study was supported by the Institute of Education Sciences (R305A090307 and R324A07118), the Centers for Disease Control and Prevention (1U49CE 000728-011 and K01CE001333-01), and the National Institute of Mental Health (1 R01 MH67948-1A1 and T32 MH19545-11). The authors would like to thank Philip Leaf for his support of this project.

Correspondence to: Elise T. Pas, Department of Mental Health, Bloomberg School of Public Health, Johns Hopkins University, 624 N. Broadway, Baltimore, MD 21205. E-mail: epas@jhsph.edu

association with a range of later disruptive behavior problems (Tobin & Sugai, 1999), including rebellious behavior, drug and cigarette use, and classroom disorderliness (Nelson & Roberts, 2000; Sprague et al., 2001). ODRs have also been found to predict school dropout rates (Tobin & Sugai, 1999). Other studies have reported low to moderate convergence between ODRs and standardized diagnostic behavioral ratings of borderline and/or clinical impairment (Nelson et al., 2002). For example, a study by Morgan-D'Atrio and colleagues (1996) found that among middle and high school students who received an ODR, 21% to 22% of these students were also found to have social skills deficits. Furthermore, 42% of the high school students with an ODR were rated by teachers as having clinically significant problems with aggression, 25% were rated as displaying delinquent behavior, and 8% were rated as having attention problems. Among middle school students with an ODR, 23% were rated as aggressive, 19% displayed delinquent behavior, and 13% had attention problems (Morgan-D'Atrio et al., 1996).

Findings regarding the validity of ODRs have been less clear when examined in elementary schools. A study of elementary students found low rates of borderline/clinical problems on the Teacher Report Form among students with an ODR (Nelson et al., 2002). Just 3.1% of students who received an ODR were rated as clinically aggressive, 2.1% displayed delinquent behaviors, and only 2.5% were rated as having clinically impaired attention problems. This low level of concordance may be the result of inconsistencies in the procedures for documenting the ODR. A recent study of the concurrent validity of ODRs found a high correlation ($r = .51$) between receiving an ODR and ratings of externalizing behaviors on the Behavior Assessment Scale for Children (McIntosh, Campbell, Carter, & Zumbo, 2009). The associations between the internalizing and adaptive scales with ODRs were nonsignificant. Therefore, additional research is needed on the validity of ODRs among behavior ratings of elementary school students.

Student, Classroom, and School Factors Influencing ODRs

There remain some concerns regarding the potential influence of other factors on a student's receipt of an ODR (e.g., Bear, 2007; Morrison et al., 2006). For example, emerging evidence documents the disproportionality of African American and Latino students in school discipline data (Kaufman et al., 2010; Skiba et al., 2008), which suggests that student race/ethnicity may play a role in teachers' perceptions of different behavior problems or decisions about how best to handle behavior problems (Bradshaw, Mitchell, O'Brennan, & Leaf, 2010). As with other measures of student behavior problems (Koth, Bradshaw & Leaf, 2009), male students and older students (i.e., those in higher grades) are also at greater risk for receiving an ODR (Bradshaw, Mitchell, O'Brennan et al., 2010; Kaufman et al., 2010; Skiba et al., 2008).

Prior research also suggests that there may be classroom or school factors that influence a student's risk of receiving an ODR. For example, the consistent implementation of classroom rules, overall management in the classroom, and classroom-level disorder (e.g., large class size, high level of disruptive behavior) likely relate to the receipt of ODRs (Morrison & Skiba, 2001). Characteristics of the teacher, such as having advanced training, may also influence students' receipt of ODRs (Alvarez, 2007). At the school level, it is likely that indicators of disorder (Birnbaum et al., 2003), such as a high faculty turnover, large school size, or high concentration of children with low socioeconomic status, would be associated with an increased risk of receiving ODRs. Additional empirical research is needed to examine whether the teacher ratings of behavior are related to the receipt of ODRs after controlling for classroom- and school-level factors, as previous studies have not accounted for contextual factors beyond the student level.

Overview of the Current Study

To address the gaps in the existing research, the current study examined the following research aims. Our first aim was to examine the reliability of ODRs by comparing two data sources: one from a school-wide administrative data system and the second from teacher report. Examining both sources of ODR data allowed us to determine which source is most closely associated with students' behavior problems. Using teacher reports and a school-managed data source of ODRs also allowed us to examine the validity of ODRs as they are generally administered, thereby enhancing the clinical significance of this study. We build on earlier studies that have incorporated formal (i.e., administrative data; as in McIntosh et al., 2009) and informal (i.e., teacher report; as in Nelson et al., 2002) referral procedures in multimethod assessments of validity. Our second aim was to examine convergent validity with conceptually similar measures (i.e., ratings of aggressive/disruptive behavior and attention problems) and divergent validity with dissimilar measures (i.e., prosocial behavior). We examined prosocial behaviors because divergent validity is largely lacking from previous studies of ODRs. The third aim was to determine whether teacher ratings of student disruptive behaviors are associated with the receipt of ODRs when other student-, classroom-, and school- level contextual factors are controlled for in a multilevel modeling approach. The use of a multilevel framework advances the prior work on this topic, which has only examined the impact of factors at one level (e.g., student- or classroom-based risk factors). Together, these findings will provide an enhanced understanding of the validity of ODRs and, in turn, will inform researchers' and psychologists' use of ODR data.

Method

Sample

The data for this study come from a sample of 8,645 children nested within 335 general education classrooms (K–5) at 21 elementary schools. There were 9,397 students enrolled in the schools at the time of data collection. Data were received for 95.3% of students ($n = 8{,}951$); complete data were received for 8,737 students (97.6% of students for which data were submitted). Ninety-two students in special education classrooms were excluded from the analyses. All general education teachers in the schools were included in this study. Student, teacher, and school demographics are provided in Table 1. The schools were implementing school-wide Positive Behavior Intervention and Supports (PBIS; Sugai & Horner, 2006), which is a universal prevention model that aims to promote consistent use of discipline practices. All of the schools had received training in school-wide PBIS and had implemented the model for 4 consecutive years. At the time of the data collection, all 21 schools were implementing school-wide PBIS with high fidelity, as indicated by scores of 80% or higher on the School-Wide Evaluation Tool (SET; Sugai, Lewis-Palmer, Todd, & Horner, 2001). The SET is a research-based measure of PBIS implementation fidelity (for additional details, see Bradshaw, Koth, Thornton, & Leaf, 2009; Bradshaw, Mitchell, & Leaf, 2010). This study was approved by the Institutional Review Board at the researchers' university.

Measures

Office Discipline Referrals. ODR data were obtained from two sources. The first source was the School-Wide Information System (SWIS; May et al., 2003), which is a commonly used Internet-based ODR data management system. A number of data elements are recorded contemporaneously and entered into this Internet-based system for each referral (e.g., student's name and grade, referring staff member's name, date, time, problem behavior). Preselected categories are present on the SWIS interface for several fields (e.g., grade, time, location, problem behavior), which decreases the risk of inaccurate data entry. Consistent with school-wide PBIS implementation procedures (Sugai &

Table 1
Descriptive Information Regarding Demographic Characteristics of the Participating Students (N = 8,645), Teachers/Classrooms (N = 335), and Schools (N = 21)

Student Characteristics	*M*(*SD*)	%
Male	–	52.7
Ethnicity		
African American	–	47.7
Hispanic	–	3.4
Caucasian	–	43.1
Other	–	5.8
Concentration Problems (TOCA-C)	2.66 (1.20)	
Disruptive Behavior (TOCA-C)	1.93 (.84)	–
Prosocial Behavior (TOCA-C)	4.78 (1.04)	
Percent of Children with Medium/High (2 or more) Teacher-Reported ODRs	–	7.2
Percent of Children with Medium (2 to 4) ODRs from SWIS		5.2
Percent of Children with High (5 or more) ODRs from SWIS		2.1
Teacher/Classroom Characteristics		
Female		93.4
Ethnicity	–	
Caucasian	–	89.0
Minority	–	11.0
Graduate Degree (Master's or Doctoral)	–	51.6
Grade Taught (4th and 5th)	–	35.2
Class size	20.47 (3.97)	–
Average Disruptive Behavior Rating for Class	1.90 (.40)	–
Positive Behavior Supports (Less Than 80% Score)	–	38.2
School Characteristics		
Percent of Faculty Turnover	14.54 (12.19)	–
Percent of Students Receiving FARMs	43.14 (19.37)	–
School Enrollment	472 (1.35)	–

Note. TOCA-C subscales (i.e., concentration, disruption, and prosocial) range from 1 to 6 point scales. *N* varies slightly by analysis due to missing data. The percent column represents the percent of students or teachers in each grouping (e.g., 52.7% of students were male, 47.3% were female).

Horner, 2006), teachers received training from school administrators in the fall of the school year on how to complete a SWIS ODR form. The referrals were entered daily into the online SWIS program by the school secretarial staff. Based on the work of May et al. (2003), we created a trichotomous SWIS ODR variable for each student to indicate whether, during that school year, the student had received 0 or 1 major ODR (e.g., fighting, defiance, abusive language, bullying/harassment; May et al., 2003), placing them in a "low" category of ODRs (or a "green zone" student); 2 to 4 ODRs, placing them in a "medium" (or "yellow") category for students; or 5 or more ODRs, placing them in a "high" (or "red") category. We also collapsed the SWIS ODR data into a dichotomized variable for low (0 or 1 ODR) and medium/high (2 or more) to align it with the teacher-reported ODRs (see next section).

The second source of ODR data was from teacher report. At the end of the school year, each teacher indicated on a student demographic form whether or not the child had been referred to the principal's office during that school year. Teachers were only asked to report whether the student had received 0, 1, or 2 or more ODRs. Therefore, the teacher-reported ODR data were dichotomized

into a "low" referral rate for 0 or 1 ODR (0) and a combined "medium/high" referral rate for 2 or more ODRs (1). These data are denoted in this article as teacher-reported ODRs. Teacher-reported ODR data were available for 8,520 (90.7%) of the eligible student sample.

Teacher Ratings of Student Behavior. Teacher ratings of each student's classroom behaviors were obtained using the Teacher Observation of Classroom Adaptation–Checklist (TOCA-C; Koth et al., 2009), which includes a 7-item measure of concentration problems (e.g., *pays attention, is easily distracted*; Cronbach's alpha [α] = .95), a 9-item measure of disruptive behavior (e.g., *breaks rules, harms property, fights;* α = .90), and a 5-item measure of prosocial behavior (e.g., *is friendly, is liked by classmates;* α = .89). Teachers rated each student in their classrooms on the TOCA-C using the 6-point scale from "never" to "almost always." Both concentration problems and disruptive behaviors were scored such that higher scores reflected greater problems, whereas the prosocial items were rated such that higher scores reflected more prosocial behaviors (unless otherwise noted). There is extensive research documenting the validity of the TOCA (for a review, see Koth et al., 2009), including 4-month test–retest correlations of .75 or higher and alpha coefficients above .80. High scores on the disruptive behavior subscale in elementary school also have been shown to be a strong predictor of violence in adolescence and adulthood (e.g., Petras, Chilcoat, Leaf, Ialongo, & Kellam, 2004). Confirmatory factor analyses on the TOCA-C demonstrated the fit of the three-factor model (see Koth et al., 2009). Subscale scores were calculated for 8,640 students (i.e., 91.9% of eligible students and 99.4% of students with complete data).

Student-Level Demographic Covariates. Select student demographic information (e.g., gender, grade, and ethnicity) was obtained from the school district. Each student's gender (male = 1, female = 0) and ethnicity (represented by one dichotomous variable: African American = 1, all others = 0) were included in the multilevel models. As indicated in Table 1, there were too few students of other ethnic groups to create additional ethnicity categories. Student demographic data were available for 8,520 (90.7%) of the eligible student sample.

Teacher- and Classroom-Level Covariates. Teacher demographic characteristics included gender (female = 1, male = 0), ethnicity (White = 1, not White = 0), having received a higher education degree (master's or doctoral degree = 1, no advanced degree = 0), and grade taught (fourth or fifth grades = 1, kindergarten through third = 0). In addition, teachers completed the classroom-wide subscale of the Effective Behavior Survey (EBS; Sugai, Todd, & Horner, 2000), which is a 12-item self-reported measure of the quality of classroom management (e.g., *having positively worded and clear statements describing rules and expectations*). This measure has previously been used in other studies as an indicator of classroom management (e.g., Safran, 2006). The EBS demonstrates adequate internal consistency (α = .83), and we have confirmed the factor structure. Teachers indicated whether each item was "in-place" within their classroom, and the measure was scored by indicating the proportion of items marked as in place. To facilitate comparisons on this measure with our dichotomous ODR outcomes, we dichotomized classrooms with EBS percentage ratings lower than 80% (i.e., fewer than 80% of the statements rated as in-place) as having poorer classroom management (i.e., poor = 1, good = 0). The 80% cut point was applied because 80% is considered high implementation of positive behavior supports (Horner et al., 2004). Classroom characteristics included class size and average of all student ratings in each classroom on the TOCA-C subscale for disruptive behavior. The teacher response rate for the self-report measure was 89.9%.

School-Level Covariates. School enrollment and the percent of students receiving free or reduced priced meals (FARMs) were obtained for each school from the state department of education. Faculty turnover rates were calculated based on staff rosters provided by the schools.

Procedure

The data were collected at the end of a single school year from the teachers via an individually addressed survey packet. The survey packets were mailed in bulk to the school and distributed to staff. Teacher participation was voluntary, and participants provided written consent. To ensure confidentiality, teachers completed the study materials on their own time and returned the materials directly to the researchers through the U.S. mail. Each teacher questionnaire packet included a small incentive with an approximate value of less than $1.00. Teachers also completed the brief student demographic form regarding each student, on which they indicated whether or not the student had been referred to the principal's office during the current school year. Across all measures, teacher response rates were 90% or higher.

Analyses

We first conducted some basic descriptive analyses, such as frequencies and means, on the ODR and TOCA-C data, respectively. To examine our first research aim, which explored the reliability of ODRs, we determined the concordance between ODR data as reported on the computerized SWIS (May et al., 2003) program and on the end-of-year teacher-report measure. We expected a high level of concordance between the two sources of ODR data as indicated by correlations, cross-tabulations, and kappa coefficients. To address our second research aim, we examined the validity of both ODR data sources. We calculated a series of correlations to determine the level of convergent validity between the ODR data and teacher ratings of student behavior problems and divergent validity between the ODR data and teacher ratings of student prosocial behavior. We then explored differences in mean ratings of disruptive behavior, concentration problems, and prosocial behavior among students with low versus medium/high, low versus medium, low versus high, and medium versus high ODRs to determine whether those with different levels of ODRs had higher ratings of concentration and disruptive concerns, similar to prior studies of the validity of ODRs (e.g., Morgan-D'Atrio et al., 1996; Nelson et al., 2002). These analyses utilized the teacher-reported and both coding systems of SWIS-reported ODRs. We also expected children with ODRs to have lower ratings of prosocial behavior, thereby demonstrating divergent validity. Next, we used receiver operator characteristic (ROC) curves (Streiner & Cairney, 2007) to determine the probability that students with medium/high and only high levels of ODRs would receive higher ratings of disruptive behavior problems than students with low or low/medium levels of ODRs; similarly, we used ROC curves to compute the probabilities for concentration problems and low rates of prosocial behavior.

To address our third research aim, we used multilevel modeling procedures to examine the associations between teacher ratings of behavior problems and students' receipt of ODRs, while adjusting for student-, teacher/classroom-, and school-level characteristics. Each outcome was dichotomous and compared low levels (0-1) of ODRs with medium/high levels (2 or more); therefore, the analyses were logistic hierarchical linear models (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2004). The three-level analyses allowed us to adjust for potential factors at multiple levels that may influence a child's risk of receiving an ODR (Birnbaum et al., 2003).

Results

Descriptive Student Data

Frequencies were conducted on the two sources of ODR data, and descriptive analyses were conducted on the TOCA-C scores (see Table 1). With regard to the ODR data, 7.2% of students were reported to have received 2 or more ODRs by the teacher, whereas 5.2% had 2 to 4 ODRs and 2.1% had 5 or more ODRs entered into the SWIS system for that school year; therefore, a total of 7.3% of

students had 2 or more SWIS-reported ODRs. With regard to the TOCA-C subscale data, students received an average rating of 2.66 for concentration problems ($SD = 1.20$), an average rating of 1.93 for disruptive behavior ($SD = 0.84$), and an average rating of 4.78 for prosocial behaviors ($SD = 1.04$) on the 1 to 6 point ratings on the TOCA-C subscales. Overall, this suggests a relatively low level of concentration problems and disruptive behavior problems. Similarly, the students were generally rated as often displaying prosocial behaviors.

Reliability Analyses

To examine the reliability of the teacher-reported ODR data compared with the SWIS data, we computed correlations in SPSS 16.0 between the two sources. The analyses indicated that the dichotomized teacher-reported ODRs and SWIS ODRs (coded dichotomously and trichotomously) were significantly and moderately correlated ($r = .57$, $p < .01$). We also computed cross-tabulations to examine the concordance between the (dichotomized) ODR captured through the SWIS system and the teacher-reported ODR. We found that for 40.8% ($n = 254$) of the children for whom the teachers reported medium/high (2 or more) referrals, there were low levels of (0 or 1) ODRs entered into SWIS (i.e., false positive), whereas 96.8% ($n = 7{,}652$) of the children for whom the teachers reported low (0 or 1) ODRs also had low (0 or 1) ODRs in SWIS (i.e., true negatives). Furthermore, 59.2% ($n = 369$) of the children for whom the teacher reported having medium/high (2 or more) ODRs, the same was reported by the SWIS system (i.e., true positives), whereas just 3.1% ($n = 245$) of the students who were rated as low (0 or 1) by the teacher, were rated as medium/high (2 or more) by SWIS (i.e., false negative). Therefore, the teacher-reported and SWIS ODR data were in agreement about whether a student had medium/high referral levels for 94.1% of students ($n = 8,021$). Finally, we calculated a Cohen's kappa coefficient (Cohen, 1960) to describe the agreement between the two items while taking chance into account. The kappa for a SWIS ODR and teacher-reported ODR was .57 ($p < .01$), which indicates moderate agreement between the two sources (Landis & Koch, 1977).

Validity Analyses

Correlations. To examine the concordance between the different sources of ODR data, we computed Spearman correlations in SPSS 16.0 between the three student-level TOCA-C subscale scores and the SWIS ODRs, coded both ways (Table 2). All correlations were significant at $p < .01$ and were .34 for the disruptive behaviors, .27 for concentration problems, and $-.29$ for the prosocial behavior subscale. Similarly, teacher-reported ODRs were moderately correlated with all three of the TOCA-C subscale scores (disruptive behaviors: $r = .38$, concentration problems: $r = .29$, and prosocial behavior: $r = -.31$) and significant at $p < .01$.

Multivariate Tests of Mean on TOCA-C Subscales. The means on the three TOCA-C subscales for the students who had medium/high levels (2 or more) of teacher-reported ODRs, as well as for those who had medium/high levels of SWIS ODRs, were examined using multivariate analysis of variance (MANOVA) in SPSS 16.0. As expected, we found that children with medium/high ODRs were rated by their teachers as having significantly more disruptive behaviors and concentration problems, in addition to fewer prosocial behaviors (Table 3) compared with students with low referral rates (0 or 1 ODR). The effect sizes (i.e., η^2) were in the small to moderate range for teacher-reported ODRs (.25) and SWIS ODRs (.18). The same results emerged when using the trichotomized SWIS ODRs. The effect size for this coding of SWIS data was small (.11).

ROC Curve Analyses. To further explore the validity of the ODR measures, we used receiver operating characteristic (ROC) curves (Streiner & Cairney, 2007; Lasko, Bhagwat, Zou, &

Table 2
Correlations between ODR Data, Recorded through SWIS and Teacher Report, and the TOCA-C Subscales (N = 8,645)

	ODR Source			TOCA-C Subscales	
	SWIS ODRs (0, 1)	SWIS ODRs (0, 1, 2)	Teacher Reported ODRs	Disruptive Behavior	Concentration Problems
Teacher-Reported ODR	.57	.57	–		
SWIS ODRs (0, 1, 2)	.94	–	–		
TOCA-C Subscales					
Disruptive Behavior	.34	.34	.38	–	
Concentration Problems	.27	.27	.29	.62	–
Prosocial Behavior	−.29	−.29	−.31	−.73	−.71

Note. SWIS (0, 1) indicates that the SWIS outcome was dichotomous, where 0 or 1 ODR = 0, and 2 or more ODRs = 1. SWIS (0, 1, 2) indicates that the SWIS outcome was trichotomous, where 0 or 1 ODR = 0, 2 to 4 ODRs = 1, and 5 or more ODRs = 2. All correlations were significant at $p < .01$. N varies slightly by analysis due to missing data.

Table 3
Comparison between Students with Low Versus Medium/High ODRs on Concentration Problems, Disruptive Behavior, and Prosocial Behavior Subscales as Measured by the TOCA-C (N = 8,645)

	TOCA-C Subscale	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Wilks Lambda	*F*	*p*	η^2
Teacher ODRs		Low	Medium/High					
	CP	2.55 (1.16)	3.99 (.96)		.75	951.8	.001	.25
	DB	1.81 (.71)	3.43 (.90)					
	PB	4.88 (.97)	3.44 (.93)					
SWIS ODRs (0, 1)		Low	Medium/High					
	CP	2.58 (1.17)	3.86 (.99)		.82	667.1	.001	.18
	DB	1.83 (.74)	3.21 (.96)					
	PB	4.86 (.99)	3.58 (.95)					
SWIS ODRs (0, 1, 2)		Low	Medium	High				
	CP	2.56 (1.17)	3.73 (.98)	4.18 (.92)	.81	328.0	.001	.11
	DB	1.83 (.74)	3.00 (.87)	3.68 (1.00)				
	PB	4.87 (.98)	3.72 (.95)	3.22 (.88)				

Note. CP = concentration problems, DB = disruptive behavior, PB = prosocial behavior. Results reported are unadjusted MANOVA results. SWIS (0, 1) indicates that the SWIS outcome was dichotomous, where 0 or 1 ODR = 0 and 2 or more ODRs = 1. SWIS (0, 1, 2) indicates that the SWIS outcome was trichotomous, where 0 or 1 ODR = 0, 2 to 4 ODRs = 1, and 5 or more ODRs = 2. Comparisons of students with low versus medium/high, low versus medium and high (separately), and medium versus high ODRs were all significantly different on each of the continuous outcomes (i.e., TOCA-C subscales; $p < .001$). N varies slightly by analysis due to missing data.

Ohno-Machado, 2005) to compare the three TOCA-C subscale scores with the dichotomized teacher-reported and SWIS-reported ODRs coded two ways, as low versus medium/high and low/medium versus high. We examined the area under the curve (AUC) statistic, which is defined as "the probability that the test will produce a value for a randomly chosen . . . [identified] subject that is greater than the value for a randomly chosen . . . [nonidentified] subject" (Lasko et al., 2005, p. 407). The AUC is statistically significant if the test (in this case, the TOCA-C subscale score) discriminates

Table 4
AUC Estimates for Each ODR Outcome (N = 8,645)

TOCA-C Subscale	Teacher-Reported ODR	SWIS ODR (Low vs. Medium/High)	SWIS ODR (Low/Medium vs. High)
Concentration Problems	.82 (.01) [.81–.84]	.80 (.01) [.78–.81]	.84 (.01) [.82–.87]
Disruptive Behavior	.92 (.01) [.91–.93]	.88 (.01) [.86–.89]	.92 (.01) [.90–.93]
Prosocial Behavior	.85 (.01) [.83–.86]	.82 (.01) [.81–.84]	.87 (.01) [.85–.89]

Note. Teacher-reported and SWIS ODR low versus medium/high was coded as 0 or 1 ODR = 0, 2 or more ODRs = 1. SWIS ODR low/medium versus high was coded 0 to 4 ODRs = 0, 5 or more ODRs = 1. Standard errors are in parentheses. 95% confidence intervals are in brackets. *N* varies slightly by analysis due to missing data.

between students with low and medium/high levels of ODRs and low/medium and high levels at a rate that is better than chance. Therefore, the null hypothesis for this test is that the AUC will be .50, which indicates that the TOCA-C subscale is no better than chance at predicting the levels of ODRs a student receives (Streiner & Cairney, 2007). We computed AUC statistics in SPSS, with the TOCA-C subscales as the test variable and the ODR measures as the state variable. AUCs ranged from .80 to .84 for concentration problems, .88 to .92 for disruptive behavior, and .82 to .87 for the reverse-coded prosocial behavior scale, which had high values for students who exhibited fewer prosocial behaviors (Table 4). According to Streiner and Cairney (2007), AUCs between .70 and .90 are considered moderate; therefore, our ROC results indicated that there was moderate concordance (and high in one case) between the TOCA-C subscales and both forms of ODR. The cut point for low/medium compared with high levels of ODRs, as reported by the SWIS system, showed the highest AUCs.

Influence of Contextual Factors on ODRs

We conducted three-level analyses in HLM v.6.01 software (Raudenbush et al., 2004) to adjust for the influence of student-, teacher/classroom-, and school-level variables on students' odds of receiving medium to high levels of ODRs. For the two three-level models (i.e., one for each dichotomous ODR type, coded as 0 or 1 ODR = 0 and 2 or more ODRs = 1), we examined the fully unconditional models, which used chi-square statistics to test the amount of variability among teachers/classrooms at level-2 and among schools at level-3. Each of the three levels in each of the multilevel models had a significant chi-square statistic ($p < .01$), which indicated the importance of examining factors at the teacher/classroom- and school-level when explaining variance in the ODR outcomes. Because the outcomes were dichotomous, the analyses were conducted using logistic hierarchical linear models, and the results are reported in terms of an adjusted odds ratio (AOR). The trichotomous coding of SWIS ODRs was not examined in HLM.

Next, we examined the significance of covariates at all three levels for each outcome and found fairly consistent results across both models (Tables 5 and 6). We found that higher ratings on the disruptive behavior TOCA-C subscale were associated with an increase in the odds of receiving medium/high levels of ODRs (covariate AORs ranged from 5.59–9.22, $ps < .01$), even after adjusting for other student-, classroom-, and school-level covariates. For example, for every 1-point increase in the TOCA-C disruptive behavior rating, students had more than a five-fold increase in the odds of receiving 2 or more SWIS-reported ODRs (AOR = 5.59, $p < .01$). Also, males (AORs ranged from 3.03–3.73, $ps < .01$) and African Americans (AORs ranged from 1.27–1.40, $ps < .05$) had significantly increased odds of receiving 2 or more ODRs; however, the association between race/ethnicity and SWIS reported ODRs in the three-level model was only marginally significant

Table 5
Multilevel Analyses Examining the Influence of Student, Teacher/Classroom, and School Characteristics on Receipt of 2 or More Teacher-Reported ODRs

Predictor Variables	Model 1		Model 2		Model 3	
	Odds Ratio	CI	Odds Ratio	CI	Odds Ratio	CI
Student Level						
Gender (Male = 1)	2.84***	2.19–3.69	2.97***	2.23–3.95	3.03***	2.32–3.95
Ethnicity (African American = 1)	1.35**	1.05–1.73	1.36**	1.04–1.77	1.40**	1.06–1.85
Disruptive Behavior	7.07***	6.12–8.16	8.85***	7.47–10.48	9.22***	7.80–10.80
Teacher/Classroom Level						
Gender (Female = 1)			0.71	0.46–1.08	0.69	0.38–1.28
Ethnicity (White = 1)			0.82	0.49–1.37	0.78	0.46–1.30
Graduate Degree (Yes = 1)			0.90	0.72–1.13	0.87	0.62–1.22
Grade (4th–5th = 1)			2.12***	1.53–2.94	2.13**	1.46–3.11
Class Size			1.02	0.99–1.05	1.03	0.99–1.08
Disruptive Behavior			0.23***	0.14–0.39	0.21***	0.13–0.35
Poor Classroom Management			1.15	0.86–1.54	1.18	0.85–1.65
School Level						
Percent of Faculty Turnover					1.01	0.99–1.04
Percent of Students with FARMs					1.01	1.00–1.03
Enrollment (per 100 Students)					0.83	0.64–1.08
Variance Components		Variance	*df*	X^2		*p*
Level-2 Error		0.97	287	468.13		.00
Level-3 Error		0.41	17	34.40		.01
Slope for Taught		0.31	20	23.10		.28

Note. CI = confidence interval. Chi-square statistics indicated that there was significant variance at each of the three levels in each of the multilevel models ($p < .01$). Graduate degree indicates attainment of master's and/or doctoral degree. Grade indicates grade currently teaching (4th or 5th grade). Poor classroom management indicates less than 80% score on the Effective Behavior Supports (EBS) measure (i.e., less than 80% on the EBS measure was coded 1 and considered poor classroom management). $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

($p = .07$) (see next page). At the classroom-level, students from fourth- and fifth-grade classrooms had increased odds of receiving 2 or more ODRs (AORs ranged from 2.02–2.13, $ps < .01$) compared with students in kindergarten through third grade, after adjusting for student-, other classroom-, and school-level covariates. Those students in classrooms with greater levels of disruptive behavior had significantly reduced odds of receiving 2 or more ODRs (AORs ranged from .21–.28, $p < .01$). Also at the classroom-level, students in poorly managed classrooms had increased odds of receiving 2 or more SWIS-reported ODRs (AOR = 1.63, $p < .05$), but not teacher-reported ODRs. None of the three school-level covariates were associated with either ODR outcome.

Discussion

Reliability and Validity of ODRs

A primary aim of the current study was to examine the reliability and validity of ODRs captured from two sources compared with behavior ratings on the previously validated TOCA-C teacher rating. We first compared the SWIS system recordings of ODRs to teacher-reported ODRs. Cross-tabulations revealed that there were high rates of true negatives and true positives. Of concern were

Table 6
Multilevel Analyses Examining the Influence of Student, Teacher/Classroom, and School Characteristics on Receipt of 2 or More SWIS-Reported ODRs

	Model 1		Model 2		Model 3	
Predictor Variables	Odds Ratio	CI	Odds Ratio	CI	Odds Ratio	CI
Student Level						
Gender (Male = 1)	3.56***	2.84–4.46	2.73***	2.29–3.26	3.77***	2.93–4.84
Ethnicity (African American = 1)	1.27**	1.05–1.55	1.40**	1.15–1.71	1.27	0.99–1.64
Disruptive Behavior	4.63***	4.13–5.19	5.22***	4.66–5.84	5.59***	4.90–6.37
Teacher/Classroom Level						
Gender (Female =1)			1.04	0.52–2.09	1.04	0.57–0.89
Ethnicity (White = 1)			0.88	0.60–1.31	0.89	0.55–1.45
Graduate Degree (Yes = 1)			0.99	0.66–1.49	0.97	0.63–1.50
Grade (4th–5th = 1)			2.04***	1.46–2.87	2.02***	1.48–2.76
Class Size			0.98	0.95–1.01	0.98	0.94–1.03
Disruptive Behavior			0.29***	0.19–0.44	0.28***	0.18–0.44
Poor Classroom Management			1.62***	1.23–2.15	1.63***	1.19–2.24
School Level						
Percent of Faculty Turnover					0.99	0.96–1.02
Percent of Students with FARMs					1.01	0.99–1.04
Enrollment (per 100 students)					1.03	0.72–1.47
Variance Components		Variance	*df*		X^2	*p*
Level-2 Error		0.94	280		484.00	.00
Level-3 Error		0.69	16		55.89	.00
Slope for Graduate Degree		0.58	19		32.23	.03

Note. CI = confidence interval. Chi-square statistics indicated that there was significant variance at each of the three levels in each of the multilevel models ($p < .01$). Graduate degree indicates attainment of master's and/or doctoral degrees. Grade indicates grade currently teaching (4th or 5th grade). Poor classroom management indicates less than 80% score on the Effective Behavior Supports (EBS) measure (i.e., less than 80% on the EBS measure was coded 1 and considered poor classroom management). $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

the slightly elevated levels of false positives, whereby teachers reported that a student received an ODR that was not recorded in SWIS. It is difficult, however, to determine whether the teachers' report was invalid, whether they observed an ODR that was not entered into SWIS (due to administrative error or oversight), or whether some ODRs were not formally documented on a referral form (due to a mistake in formal procedures). It is also possible that some teachers erroneously recalled a student's receiving 2 or more ODRs, when in fact the student had only received one or no referrals. Furthermore, the kappa indicated moderate agreement between the SWIS-reported ODR and the teacher-reported ODR, which provided additional evidence of an acceptable level of agreement between the two sources of ODR data.

Despite some discrepancies, the teacher-report measures of ODRs identified almost the exact same percentage of students receiving 2 or more ODRs as the SWIS system (7.2% vs. 7.3%). Although the correlation between the two sources of ODR data was in the moderate range ($r = .57$), it suggests some disagreement between the two ODR sources (i.e., for 5.9% of students). Additional research is needed to better understand the potential source of such discrepancies. This finding is promising, given that teacher-reported ODRs are a cheaper and more efficient measure than SWIS.

This may indicate that when school-wide archival systems such as SWIS are unavailable, collecting teacher-reported ODRs may be an acceptable alternative.

When examining the convergent and divergent validity of ODR data by comparing it with an established behavior rating scale (TOCA-C), moderate to high statistically significant correlations were revealed for the disruptive behavior, attention problems, and prosocial behavior subscales. Also, when comparing the TOCA-C subscale scores for those students with low and medium/high levels of ODRs, children with medium/high ODRs from both sources were rated by their teachers as having significantly more disruptive behaviors and concentration problems and significantly fewer prosocial behaviors. Not surprisingly, the associations between ODRs and the TOCA-C subscales were slightly stronger for disruptive behavior than for the other two subscales. Similar findings were reported by Morgan-D'Atrio et al. (1996), who found that fewer students were rated by teachers as displaying attention problems than were rated as aggressive or delinquent. We also found moderate to high associations between the TOCA-C subscales and ODR data in ROC curve analyses. Taken together, results of a variety of analyses (e.g., correlational, MANOVAs, and ROC) provide evidence of the moderate convergent and divergent validity for both sources of ODR data with behavior ratings.

Adjusting for Contextual Influences on Receipt of an ODR

To explore the third aim of the study, we fit three-level models in HLM to determine whether the teacher ratings of disruptive behavior were significantly related to ODRs after controlling for other student-, classroom-, and school-level contextual factors. There was a significant association between disruptive behavior and both ODR outcomes (SWIS and teacher-reported). These multilevel models indicated that not only was this effect not diminished by adding covariates at all three levels, but the relationship actually grew in magnitude when these other covariates were adjusted for in the fully conditional models. These findings suggest that although ODRs may be influenced by teacher perception or classroom management (Irvin et al., 2004; Morrison et al., 2006), student-level disruptive behaviors are highly associated with receipt of ODRs over and above the influence of these other contextual factors.

Although not a central focus of the current paper, a number of individual and contextual factors were significantly associated with receipt of ODRs in the multilevel analyses. For example, males and African Americans also had increased odds of being referred 2 or more times (see models 1-3 in Table 5, and models 1-2 in Table 6); this finding is consistent with prior research (Bradshaw, Mitchell, O'Brennan et al., 2010; Kaufman et al., 2010; Skiba et al., 2008). Students were approximately 60% more likely to have received 2 or more ODRs in the SWIS system when in classrooms rated by the teacher as having poor classroom management. This effect was not significant for the teacher-reported ODRs. In contrast, students in classrooms marked by high levels of disruptive behavior (i.e., the classroom average rating of all students' disruptive behavior on the TOCA-C) were actually less likely to receive 2 or more ODRs. Although this may seem counterintuitive, a possible explanation is that students with behavior problems stand out from their peers in classrooms in which a high proportion of the students do not engage in disruptive behaviors. However, similar students in a classroom with many students who exhibit behavior problems may not stand out and thus are less likely to be sent to the principal's office (Bradshaw, Mitchell, O'Brennan et al., 2010). Other teacher factors were not significantly associated with the odds of students receiving 2 or more referrals. The school-level covariates were not associated with student referrals.

Limitations and Future Directions

The behavior ratings used to test the validity of ODR data were teacher ratings, not independent ratings of children's behavior. Therefore, the teacher-reported ODR data and TOCA-C ratings

share common method variance. Schools were solely responsible for collecting the SWIS referral data. A more comprehensive, and perhaps "true," assessment of ODRs and behavior from multiple perspectives (e.g., parents, students, observations) may reveal different findings. Similarly, classroom management was a self-report rather than an objective observation.

We dichotomized the receipt of 2 or more ODRs to align with PBIS literature, which suggests that students with 0 or 1 referrals are at low-risk, whereas 2 to 4 ODRs indicate elevated risk, and 5 or more ODRs indicate high risk (May et al., 2003). Using this metric, 85% of the current sample had 0 ODRs, and 93% of had 0 or 1 ODRs, which is a slightly higher proportion than the hypothesized 80%. It is possible that the developmental level of the sample may explain the relatively low rate of ODRs (Bradshaw, Mitchell, & Leaf, 2010). Future studies should examine whether a similar pattern of findings occurs in secondary schools, where the rates of disruptive behavior problems and ODRs tend to be higher. The low rates of ODRs could also be due in part to implementation of school-wide PBIS. For example, the staff received training on how to administer and complete ODR documentation, which may have resulted in greater consistency in ODR procedures and documentation than observed in schools not implementing PBIS. Moreover, this study took place within the context of school-wide PBIS, which previous research has demonstrated significantly alters the school's organizational context (Bradshaw et al., 2009) and reduces ODRs (Bradshaw, Mitchell, & Leaf, 2010).

There also are other contextual factors not assessed within this study that should be examined in future studies. For example, we did not have information on teachers' mood, which may have influenced a teacher's likelihood of referring the student to the office. Furthermore, variation in school rules and administrative leadership regarding the use of ODRs may have influenced teachers' use of reactive management practices (Morrison & Skiba, 2001). Finally, although all the data are nested within classrooms and schools, only the multilevel analyses accounted for the nesting. Further methodological developments are needed to calculate reliability and validity statistics (e.g., ROC curves and kappa) while accounting for the correlated error terms.

Whereas the current study examined factors that predict the receipt of ODRs in a single school year, future research should examine the predictive validity of ODRs across multiple years with regard to other outcomes, such as dropout, suspension, school failure, or other behavioral or mental health problems. As noted in the introduction, there is some research exploring the predictive validity of ODRs for subsequent problems such as academic problems, substance use, and deviant behavior (e.g., McIntosh et al., 2008; Nelson & Roberts, 2000). However, additional research is needed to determine the utility of ODRs as a screening tool to predict subsequent problems. Consistent with an RTI approach to behavior problems, ODRs could be informative in identifying children not responding to the universal system of support and therefore in need of other interventions (Hawken et al., 2008; Sugai et al., 2000).

Implications

These findings extend prior research, which has primarily focused on the correspondence between ODR data and children's social skills, academic problems, and clinical symptoms (Morgan-D'Atrio et al., 1996; Nelson et al., 2002). The results indicate that ODRs have moderate concordance with teacher ratings of student problem behaviors and divergence from prosocial behaviors. Not surprisingly, the correlational data suggested that ODRs are more highly associated with disruptive behavior than with concentration problems or prosocial behavior. It is promising to see that ODRs appear to be reserved for students with disruptive behavior problems and do not appear to be associated with more general problems (e.g., concentration problems). However, an ODR may signal a risk for academic as well as behavior problems (McIntosh et al., 2008).

These results also suggest that ODRs may be helpful as a screening measure to identify students who are in need of further intervention and supports, particularly students who should be screened for further follow-up on behavior issues. There is preliminary evidence that using cut points for the number of ODRs received corresponds with differing levels of clinical symptoms and therefore can be effective in guiding decisions regarding intervention supports (McIntosh et al., 2009). Based on our ROC findings, teacher ratings of disruptive behavior, concentration problems, and poor prosocial behaviors best predict the receipt of ODRs at the cut point of 5 or more. This provides additional support for the use of cut points when making decisions regarding problem behaviors. On the other hand, some students in need of support may not be identified by an ODR. This suggests that although ODRs may signal some types of behavioral concerns, there may be other students whose problems may be more difficult to detect (Bradshaw, Buckley, & Ialongo, 2008).

These findings also have implications for program evaluations using ODRs as outcomes. ODRs may be an efficient proxy for teacher ratings of disruptive behavior, which are typically more difficult and expensive to collect in large-scale prevention studies. The multilevel results revealed that student behavior and characteristics were most predictive of referrals, but they also highlighted the potential significance of classroom-level factors. For example, poor classroom management was linked with an increased risk of ODRs. Therefore, professional development and consultation focused on improving classroom management strategies may reduce students' risk of being referred. In closing, these findings suggest that ODRs have a moderate level of validity as an indicator of student problem behaviors and thus may be potentially helpful in identifying students in need of additional support services (Hawken et al., 2008).

References

Alvarez, H. K. (2007). The impact of teacher preparation on responses to student aggression in the classroom. Teaching and Teacher Education, 23, 1113–1126.

Bear, G. G. (2007). School-wide approaches to behavior problems. In B. Doll & J. Cummings (Eds.), Transforming school mental health services: Population-based approaches to promoting the competency and wellness in children (pp. 103–142). Thousand Oaks, CA: Corwin Press.

Birnbaum, A. S., Lytle, L. A., Hannan, P. J., Murray, D. M., Perry, C. L., & Forester, J. L. (2003). School functioning and violent behavior among young adolescents: A contextual analysis. Health Education Research, 18, 389–403.

Bradshaw, C. P., Buckley, J., & Ialongo, N. (2008). School-based service utilization among urban children with early-onset educational and mental health problems: The squeaky wheel phenomenon. School Psychology Quarterly, 23, 169–186.

Bradshaw, C. P., Koth, C. W., Thornton, L. A., & Leaf, P. J. (2009). Altering school climate through school-wide Positive Behavioral Interventions and Supports: Findings from a group-randomized effectiveness trial. Prevention Science, 10, 100–115.

Bradshaw, C. P., Mitchell, M. M., & Leaf, P. J. (2010). Examining the effects of School-Wide Positive Behavioral Interventions and Supports on student outcomes: Results from a randomized controlled effectiveness trial in elementary schools. Journal of Positive Behavior Interventions, 12, 133–148.

Bradshaw, C. P., Mitchell, M. M., O'Brennan, L. M., & Leaf, P. J. (2010). Multilevel exploration of factors contributing to the overrepresentation of Black students in office disciplinary referrals. Journal of Educational Psychology, 102, 508–520.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37–46.

Hawken, L. S., Vincent, C. G., & Schumann, J. (2008). Response to Intervention for social behavior: Challenges and opportunities. Journal of Emotional and Behavioral Disorders, 16, 213–225.

Horner, R. H., Todd, A. W., Lewis-Palmer, T., Irvin, L. K., Sugai, G., & Boland, J. B. (2004). The School-Wide Evaluation Tool (SET): A research instrument for assessing school-wide positive behavior support. Journal of Positive Behavior Interventions, 6(1), 3–12.

Irvin, L. K., Horner, R. H., Ingram, K., Todd, A. W., Sugai, G., Sampson, N. K., et al. (2006). Using office discipline referral data for decision making about student behavior in elementary and middle schools: An empirical evaluation of validity. Journal of Positive Behavior Interventions, 8, 10–23.

Irvin, L. K., Tobin, T. J., Sprague, J. R., Sugai, G., & Vincent, C. G. (2004). Validity of office discipline referral measures as indices of school-wide behavioral status and effects of school-wide behavioral interventions. Journal of Positive Behavior Interventions, 6, 3–12.

Kaufman, J. S., Jaser, S. S., Vaughan, E. L., Reynolds, J. S., Di Donato, J., Bernard, S. N., et al. (2010). Patterns in office referral data by grade, race/ethnicity, and gender. Journal of Positive Behavior Interventions, 12, 44–54.

Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2009). Teacher Observation of Classroom Adaptation-Checklist (TOCA-C): Development and factor structure. Measurement and Evaluation in Counseling and Development, 42, 15–30.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159–174.

Lasko, T. A., Bhagwat, J. G., Zou, K. H., & Ohno-Machado, L. (2005). The use of receiver operating characteristic curves in biomedical informatics. Journal of Biomedical Informatics, 38, 404–415.

May, S., Ard, W., Todd, A., Horner, R., Glasgow, A., Sugai, G., et al. (2003). Schoolwide information system. Eugene: University of Oregon, Educational and Community Supports.

McIntosh, K., Campbell, A. L., Carter, D. R., & Zumbo, B. D. (2009). Concurrent validity of office discipline referrals and cut points used in Schoolwide Positive Behavior Support. Behavioral Disorders, 34, 100–113.

McIntosh, K., Flannery, K. B., Sugai, G., Braun, D. H., & Cochrane, K. L. (2008). Relationships between academics and problem behavior in the transition from middle school to high school. Journal of Positive Behavior Interventions, 10, 243–255.

Morgan-D'Atrio, C., Northrup, J., LaFleur, L., & Spera, S. (1996). Toward prescriptive alternatives to suspensions: A preliminary evaluation. Behavioral Disorders, 21, 190–200.

Morrison, G. M., Redding, M., Fisher, E., & Peterson, R. (2006). Assessing school discipline. In S. Jimerson & M. Furlong (Eds.), Handbook of school violence and school safety: From research to practice (pp. 211–220). New York: Routledge.

Morrison, G. M., & Skiba, R. (2001). Predicting violence from school misbehavior: Promises and perils. Psychology in the Schools, 38, 173–184.

Nelson, J. R., Benner, G. J., Reid, R. C., Epstein, M. H., & Currin, D. (2002). The convergent validity of office discipline referrals with the CBCL-TRF. Journal of Emotional & Behavioral Disorders, 10, 181–188.

Nelson, J. R., & Roberts, M. L. (2000). Ongoing reciprocal teacher-student interactions involving disruptive behaviors in general education classrooms. Journal of Emotional & Behavioral Disorders, 8, 27–37.

Petras, H., Chilcoat, H., Leaf, P., Ialongo, N., & Kellam, S. (2004). The utility of teacher ratings of aggression during the elementary school years in identifying later violence in adolescent males. Journal of the American Academy of Child and Adolescent Psychiatry, 1, 88–96.

Raudenbush, S. W., Bryk, A., Cheong, Y.-F., Congdon, R., & du Toit, M. (2004). HLM Statistical Software (Version 6) [Computer software]. Lincolnwood, IL: Scientific Software International.

Safran, S. (2006) Using the Effective Behavior Supports survey to guide development of schoolwide positive behavior support. Journal of Positive Behavior Interventions, 8, 3–9.

Skiba, R., Horner, R., Chung, C. G., Rausch, M. K., May, S. L., & Tobin, T. (2008, March). Race is not neutral: A national investigation of African American and Latino disproportionality in school discipline. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Sprague, J., Walker, H. M., Stieber, S., Simonsen, B., Nishioka, V., & Wagner, L. (2001). Exploring the relationship between school discipline referrals and delinquency. Psychology in the Schools, 38, 197–206.

Streiner, D. L., & Cairney, J. (2007). What's under the ROC? An introduction to receiver operating characteristics curves. Canadian Journal of Psychiatry, 52, 121–128.

Sugai, G., & Horner, R.H. (2006). A promising approach for expanding and sustaining the implementation of school-wide positive behavior support. School Psychology Review, 35, 245–259.

Sugai, G., Lewis-Palmer, T., Todd, A., & Horner, R. (2001). School-wide evaluation tool (SET). Eugene, OR: Center for Positive Behavioral Supports, University of Oregon.

Sugai, G., Sprague, J. R., Horner, R. H., & Walker, H. M. (2000). Preventing school violence: The use of office discipline referrals to assess and monitor school-wide discipline interventions. Journal of Emotional and Behavioral Disorders, 8, 94–101.

Sugai, G., Todd, A. W., & Horner, R. (2000). Effective Behavior Support (EBS) Survey: Assessing and Planning Behavior Supports in Schools. Eugene, OR: University of Oregon.

Tobin, T. J., & Sugai, G. M. (1999). Using sixth-grade school records to predict school violence, chronic discipline problems, and high school outcomes. Journal of Emotional and Behavioral Disorders, 7, 40–53.