

Evidence-Based School Behavior Assessment of Externalizing Behavior in Young Children

Daniel M. Bagner

Department of Psychology

Florida International University, Miami

Stephen R. Boggs and Sheila M. Eyberg

Department of Clinical & Health Psychology, University
of Florida, Gainesville

Abstract

This study examined the psychometric properties of the Revised Edition of the School Observation Coding System (REDSOCS). Participants were 68 children ages 3 to 6 who completed parent-child interaction therapy for Oppositional Defiant Disorder as part of a larger efficacy trial. Interobserver reliability on REDSOCS categories was moderate to high, with percent agreement ranging from 47% to 90% ($M = 67\%$) and Cohen's kappa coefficients ranging from .69 to .95 ($M = .82$). Convergent validity of the REDSOCS categories was supported by significant correlations with the Intensity Scale of the Sutter-Eyberg Student Behavior Inventory-Revised and related subscales of the Conners' Teacher Rating Scale-Revised: Long Version (CTRS-R: L). Divergent validity was indicated by nonsignificant correlations between REDSOCS categories and scales on the CTRS-R: L expected not to relate to disruptive classroom behavior. Treatment sensitivity was demonstrated for two of the three primary REDSOCS categories by significant pre to posttreatment changes. This study provides psychometric support for the designation of REDSOCS as an evidence-based assessment procedure for young children.

KEY WORDS: evidence-based assessment, school observation, disruptive behavior, psychometrics, parent-child interaction therapy, preschool children.

The importance of evidence-based assessment (EBA) in clinical child and adolescent psychology has received recent attention (Mash & Hunsley, 2005). Without psychometrically sound assessment procedures, it is difficult to ensure that children benefit from treatment

This study was funded by the National Institute of Mental Health (RO1 MH60632). The data were previously presented at the Sixth Annual Parent-Child Interaction Therapy Conference in Gainesville, FL. We thank the members of the Child Study Laboratory for their contributions to this research.

Correspondence to Daniel M. Bagner, Ph.D., Department of Psychology, Florida International University, 11200 S.W. 8th St., Miami, FL 33199; e-mail: dbagner@fiu.edu.

(Kazdin, 2005). For young children with externalizing behavior problems, accurate early identification and assessment-driven treatment planning is essential in view of the risk for later delinquency and the high societal costs associated with undetected and untreated disorders (McMahon & Frick, 2005).

Behavioral observation is the hallmark of behavioral assessment (Bagner, Harwood, & Eyberg, 2006). It is particularly important for the assessment of externalizing behavior disorders because of its objectivity (McMahon & Frick, 2005; Pelham, Fabiano, & Massetti, 2005). School observation provides a unique opportunity to observe children in a natural environment with expectations for appropriate social behavior. School observation may also be necessary for accurate diagnosis of young children with Attention-Deficit Hyperactivity Disorder (ADHD), which is highly comorbid with disruptive behavior disorders (Angold, Costello, & Erkanli, 1999; Newcorn et al., 2001) and significantly worsens the prognosis for young children (Lahey & Loeber, 1997).

Several school observation systems have been developed to assess disruptive classroom behaviors, such as the Direct Observation Form (DOF; Achenbach & Rescorla, 2001), the Classroom Observation Code (Abikoff, Gittelman, & Klein, 1980; Abikoff & Gittelman, 1985), and the Student Observation System (SOS; Reynolds & Kamphaus, 2004). The DOF, designed for 5- to 14-year-olds, involves observation during 10 minute periods in the classroom during group activities. An independent observer writes a narrative of the child's behavior across the 10-minute period and then completes a 96-item rating scale. Scores are averaged for up to six observation sessions. The DOF has been shown to have good reliability and validity (Achenbach & Rescorla, 2001).

The Classroom Observation Code, designed for children ages 6 to 12, involves coding the presence or absence of hyperactive behaviors in three different categories (i.e., Interference, Solicitation, and Off-Task Behavior). The Classroom Observation Code has been found to discriminate hyperactive from normal children (Abikoff, Gittelman, & Klein, 1980).

The SOS uses a time sampling procedure involving 3-second intervals during which an independent observer codes a range of adaptive and maladaptive behaviors, such as positive peer interaction and repetitive motor movements, of children between the ages of 2.5 and 18 years (Reynolds & Kamphaus, 2004). The SOS has had limited psychometric study and is not recommended for practice at this time (McMahon & Frick, 2005). To our knowledge, there are currently no evidence-based school behavior coding systems for the assessment of

disruptive behavior in preschool-age children.

The Revised Edition of the School Observation Coding System (REDSOCS; Jacobs et al., 2000) is an interval coding system designed to assess young children's externalizing behavior in the classroom. Jacobs and colleagues (2000) examined the psychometric properties of the REDSOCS with 182 non-referred children and provided reference point data based on observations of these children. In addition to high interobserver reliability and concurrent validity with teacher-report measures, their study demonstrated the discriminative validity of the REDSOCS by finding that all categories showed significant differences between non-referred children, clinic-referred children without classroom behavior problems, and clinic-referred children with parent-reported school behavior problems. A modified version of the REDSOCS (i.e., combining noncompliance and inappropriate behavior into a single code) was also found to have high convergent and divergent validity with teacher-report measures as well as the ability to discriminate between disruptive and typical children (Filcheck, Berry, & McNeil, 2004). These studies provided a preliminary psychometric evidence base for the REDSOCS, although further study is needed to establish the REDSOCS as an EBA.

Mash and Hunsley (2005) have stressed the importance of replicated evidence of validity to qualify a measure as evidence-based. The Jacobs et al. (2000) study examined the psychometric properties of the REDSOCS in a nonreferred sample, and the Filcheck et al. (2004) study examined the psychometric properties of a modified version of the REDSOCS in a mixed normal and clinical sample. However, the psychometric properties of the REDSOCS have not yet been established in a clinical sample. Further study with a clinical sample is important because children with externalizing behavior disorders typically show higher frequencies of certain behaviors (e.g., inappropriate behavior), which may affect the interobserver reliability of the coding system. McMahon and Frick (2005) have also suggested that one criterion important for an assessment procedure designated as an EBA is demonstration of its treatment sensitivity. Although the original version of the REDSOCS (i.e., SOCS) identified generalization of treatment gains following parent-child interaction therapy (PCIT; McNeil, Eyberg, Eisenstadt, Newcomb, & Funderburk, 1991), generalization of treatment effects measured by the REDSOCS has not yet been tested.

The purpose of this study was two-fold. First, we examined the reliability and validity of the REDSOCS in a sample of clinic-referred preschool-age children. Both percent agreement and Cohen's kappa were used to assess agreement between trained observers. We investigated convergent validity by examining correlations between

the REDSOCS primary categories (i.e., Inappropriate, Noncompliant, and Off-Task) and rating scale measures representing similar constructs. We predicted that the REDSOCS Noncompliant Behavior category would correlate positively with the Intensity Scale of the Sutter-Eyberg Student Behavior Inventory-Revised (SESBI-R; Eyberg & Pincus, 1999), a teacher-report scale measuring the frequency of disruptive classroom behavior, and the Oppositional subscale on the Conners Teacher Rating Scale-Revised: Long Version (CTRS-R: L; Conners, 1997). We also hypothesized that the REDSOCS Inappropriate Behavior category would correlate positively with the Hyperactive-Impulsive DSM-IV subscale of the CTRS-R: L. Finally, we predicted that the REDSOCS Off-Task Behavior category would correlate positively with the Inattentive DSM-IV subscale of the CTRS-R: L.

To examine divergent validity, we examined correlations between the three primary REDSOCS categories with the Anxious-Shy and Perfectionism scales of the CTRS-R: L. We predicted that the primary REDSOCS categories would not be related to the Anxious-Shy or Perfectionism scales of the CTRS-R: L, which both represent constructs theoretically unrelated to the REDSOCS.

To determine treatment sensitivity of the REDSOCS, we examined changes in its three primary behavior categories from pre to post PCIT, an evidence-based parent training intervention designed to change parent-child interactions and thereby change child disruptive behavior (Zisser & Eyberg, 2009). Improvements in children's compliance to parental commands and parent ratings of children's behavior at home following PCIT have previously been found to generalize to the school as measured by teacher rating scales (Funderburk et al., 1998), and observations of classroom compliance using the original SOCS (McNeil et al., 1991). Consistent with those findings, we predicted that scores for the three primary REDSOCS categories would improve from pre to posttreatment.

Method

Participants

Participants were 68 children between the ages of 3 and 6 years drawn from a larger study examining the efficacy of maintenance treatment following PCIT. Children had been referred for treatment of disruptive behavior by health care providers or were self-referred to a psychology clinic in a large university health sciences center. For study inclusion, children had to meet diagnostic criteria for Oppositional Defiant Disorder (ODD) according to the Diagnostic Interview Schedule for Children (Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000) and had to receive a *T* score > 61 on the Aggres-

sive Behavior Scale on the Child Behavior Checklist (Achenbach, 1991; Achenbach, 1992). Children taking medication for hyperactivity ($n = 23$) were required to have the dosage constant for at least one month before the pretreatment assessment, and families were asked not to change medication status or dosage during treatment. The child and parents had to obtain standard scores of ≥ 70 and ≥ 75 , respectively, on cognitive screening measures - the Peabody Picture Vocabulary Test Third Edition (Dunn & Dunn, 1997) for the child and the Wonderlic Personnel Test (Dodrill, 1981) for the parents. Children with a history of severe sensory or mental impairment (e.g., deafness, autism) or an immediate crisis requiring hospitalization or out-of-home placement were excluded.

Most of the children were male (68%) and Caucasian (81%). The mean Hollingshead (1975) score was 39.85 ($SD = 14.24$) indicating that, on average, families fell in the lower-middle income range of socioeconomic status. Sixty-three percent of the children were reported by their parents to have behavior problems at school on a *yes-no* question from the intake questionnaire. Children with parent-reported school behavior problems were disproportionately male (72%); $\chi^2(1) = 4.42$, $p = .035$, and obtained higher scores on the CTRS-R: L scales labeled Inattentive, $t(60) = -3.70$, $p < .001$; Hyperactive-Impulsive, $t(61) = -6.35$, $p < .001$; and Anxious, $t(61) = -3.70$, $p < .001$, than other children. These children were also observed to have a higher frequency of Inappropriate Behavior on the REDSOCS than remaining children, $t(61) = -2.75$, $p = .008$. All children were included in this study to obtain representative scores for young children with diagnosed ODD. The demographic characteristics of the 68 children are summarized in Table 1.

Diagnostic Screening Measures

Child Behavior Checklist (CBCL; Achenbach, 1991; Achenbach, 1992). The CBCL is an empirically derived behavior rating scale containing factor-analytically derived subscales representing various categories of child psychopathology. The Aggressive Behavior subscale was used as one index of ODD required for study inclusion. Parents of children ages 4 through 6 completed the CBCL/4-18, which consists of 118 behavior problem items rated on a 3-point scale from (0) *not true*, to (2) *very true* or *often true*. Mean one-week test retest reliability has been reported at .91 for the Aggressive Behaviors subscale (Achenbach, 1991). The CBCL/2-3, administered to parents of 3-year-old children, is similar in format to the CBCL/4-18 but contains 99 items appropriate for younger children. Test-retest reliability of the CBCL/2-3 has been reported at .85 for the Aggressive Behavior subscale over a three-week period (Koot, Van Den Oord, Verhulst, & Boomsma, 1997).

Table 1
Demographic Characteristics (N = 68)

Characteristic	<i>M</i> / %	<i>SD</i>
Child age (years)	4.46	1.04
Child sex (% male)	81	--
Family SES ^a	39.85	14.24
Child medication (%)	34	--
Child ethnicity/race		
% African American	9	--
% Asian	1	--
% Bi-racial	7	--
% Caucasian	81	--
% Hispanic	2	--
Reported school problems (%)	63	--

^aSES = Socioeconomic status measured by the Hollingshead (1975) Four Factor Index of Social Position, which yields a score based on parents' education, occupation, sex, and marital status

Diagnostic Interview Schedule for Children-Fourth Edition-Parent Version (DISC-IV-P; Shaffer et al., 2000). The DISC-IV-P, a structured diagnostic interview administered to parents, includes separate modules for each of the common mental disorders of children included in the DSM-IV. One-week test-retest reliability for administration to parents of 9 to 17 year old children has been reported at .54 for ODD (Shaffer et al.). Although the DISC-IV-P was developed for use with children ages 6 years and older, it has been used successfully with preschool-age children (Speltz, McClellan, DeKlyen, & Jones, 1999). The DISC-IV-P ODD module was used, along with the CBCL Aggressive Behavior subscale, to screen children for inclusion based on the presence of ODD.

Outcome Measures

Revised Edition of the School Observation Coding System (REDSOCS; Jacobs et al., 2000). The REDSOCS is an interval coding system for recording disruptive classroom behaviors of preschool- and elementary-age children. Behaviors are coded in 10-second intervals, and the following eight categories were coded: Inappropriate, Appropriate; Noncompliant, Compliant, No Command Given; Off-task, On-task, and Not Applicable.

Inappropriate Behavior is defined as behavior that may be annoying or disruptive to the teacher or other children, such as

whining, crying, yelling, destructive behavior, and aggressive behavior. Appropriate Behavior is the absence of all Inappropriate Behavior for the entire 10-second interval. Because Appropriate Behavior is the inverse of Inappropriate Behavior, only outcomes for Inappropriate Behavior are reported.

Noncompliant Behavior occurs when the target child makes no movement toward obeying a direct or indirect teacher command during a 5-second period following the command (Wruble, Sheeber, Sorensen, Boggs, & Eyberg, 1991). Compliant Behavior occurs when the target child begins or attempts to obey within 5 seconds of a direct or indirect teacher command. No Command Given is coded when there is no teacher command issued within the 10-second interval.

Off-task Behavior is coded if at any point during the 10-second interval the target child does not attend to the material or task expected in the classroom and includes behaviors such as getting out of seat, talking out, and daydreaming. On-task Behavior is coded in the absence of any Off-task behavior for the entire 10-second interval, whereas Not Applicable is coded when there is no readily identifiable task that the target child is expected to perform.

In the Jacobs et al. (2000) study, Cohen's kappa reliability coefficients for all behavior codes ranged from .78 (Noncompliant Behavior) to .94 (No Task), and the three primary behavior categories were significantly correlated with scores on the Sutter-Eyberg Student Behavior Inventory (Eyberg, 1992) as well as a shortened version (Goyette, Conners, & Ulrich, 1978) of the original 39-item Conners Teacher Rating Scale (Conners, 1969).

Conners Teacher Rating Scale-Revised: Long Version (CTRS-R: L). The CTRS-R: L (Conners, 1997) is a 59-item teacher rating scale that measures ADHD and comorbid disorders on a 4-point Likert scale that ranges from (0) *Not True At All* to (3) *Very Much True*. The CTRS-R: L has shown test-retest reliability correlations between .47 (Inattention/Cognitive Problems scale) and .88 (Anxious/Shy scale; Conners, Sitarenios, Parker, & Epstein, 1998). The Oppositional subscale and the DSM-IV Inattentive and Hyperactive-Impulsive subscales of the CTRS-R: L were used in this study to assess convergent validity of the REDSOCS; the internal consistency coefficients of these three subscales were .78, .88, and .93, respectively, in our sample. The Anxious-Shy and Perfectionism subscales were used to assess divergent validity with the REDSOCS; the internal consistency coefficients for these two subscales were .70 and .79, respectively, in our sample.

Sutter-Eyberg Student Behavior Inventory-Revised (SESBI-R). The SESBI-R (Eyberg & Pincus, 1999) is a 38-item teacher rating scale of disruptive behavior at school for children between the ages of 2 and

16 years. It contains two scales: The Intensity Scale measures the frequency of children's disruptive behavior on a 7-point scale from (1) *never* to (7) *always*. The *yes-no* Problem Scale measures the teacher's tolerance for the child's behavior. The Intensity Scale, which measures a construct similar to the REDSOCS categories of child behavior, was used in this study to assess the convergent validity of the REDSOCS. The SESBI-R Intensity Scale has shown 1-week test-retest reliability of .80 in a sample of preschool-age children (Querido & Eyberg, 2003). In this study, its internal consistency reliability was .98.

Procedure

Approvals by the Institutional Review Board and the school board of Alachua County, Florida were obtained prior to collection of data. During the pretreatment assessment, consent to contact the child's teacher was obtained from the mother of each child. Graduate and undergraduate research assistants were then assigned to conduct the school observations. Coder training involved initial didactic instruction in observation coding procedures and category definitions, followed by 3 to 4 hours of practice coding of videotaped classroom behavior. Once observers met 80% interobserver agreement on the eight REDSOCS category codes during a 10-minute tape segment, they began live practice coding during ongoing classroom observations until they attained 80% interobserver agreement. Coder training involved approximately 6 to 8 hours, spread over 3 to 4 weeks.

School observations were conducted on three different days within a 2-week period during structured classroom activities (e.g., circle time). The type of activity varied among classrooms, but required only that some expectation for activity was placed on the child. Observations were conducted in 68 different classrooms during the pretreatment assessment and in an additional 27 different classrooms at the posttreatment assessment. Observational data were collected in 10-second intervals for a total of 10 coded minutes (i.e., 60 intervals) per child. In this study, each minute of continuous coding was followed by a 3-minute pause to lengthen the sampling period. Thus, each observation lasted a total of 40 minutes on each of 3 visits. A REDSOCS coding form is shown in Figure 1. To obtain interobserver reliability data, a second observer accompanied the primary observer on one of the 3 days of observation for each child and independently coded the observation using a dual-jack tape recorder to cue the time intervals. The classroom observers were not involved in other aspects of the child's assessment or treatment and were anonymous to the child. Teachers were also asked to complete the SESBI-R and the CTRS-R: L on each child and were compensated \$25 for their time.

School Observation Sheet

Observer: _____
 Date: _____
 Child: _____

Status: *(circle one)* Primary or Reliability
 School: _____

	Minute 1						Minute 2						Minute 3						Minute 4						Minute 5											
Appropriate																																				
Inappropriate*																																				
Comply																																				
Noncomply																																				
No Com. Given																																				
On Task																																				
Off Task																																				
Not Applicable																																				

	Minute 6						Minute 7						Minute 8						Minute 9						Minute 10											
Appropriate																																				
Inappropriate*																																				
Comply																																				
Noncomply																																				
No Com. Given																																				
On Task																																				
Off Task																																				
Not Applicable																																				

* **Inappropriate behavior includes:** whine destructive talks out self-stimulation yell
 tantrum disruptive cheating out demanding cny negativism
 out of area

Notes: _____

Figure 1. REDSOCS Coding Form

Observations were conducted for 68 children enrolled in school at the time of the pretreatment assessment. Children were required to have been in the classroom of the current teacher for at least 5 weeks at the time of assessment, to permit accurate teacher ratings of child behavior at pre and posttreatment. Posttreatment school observations were collected for 34 children who completed treatment, were enrolled in school at the time of pre and posttreatment assessment, and had been in their current classroom for at least 5 weeks at the time of the assessment (a one-month extension after treatment completion was permitted to obtain posttreatment school data). Children who dropped out of treatment ($n = 22$) or who completed treatment during summer months when their school was closed ($n = 12$) were not included in the treatment sensitivity analyses.

Results

Reliability

Interobserver agreement for the REDSOCS, as calculated by percent agreement on occurrences for each category, ranged from 47% for the Noncompliant Behavior code to 90% for the No Command Given code. Kappa coefficients (Cohen's kappa; Fleiss, 1981) for all codes ranged from .69 for Compliant Behavior and Off-Task Behavior to .95 for the Not Applicable category (see Table 2).

Convergent and Divergent Validity

Correlations between the primary REDSOCS categories and teachers' ratings on the SESBI-R Intensity Scale and selected CTRS-R: L scales were used to evaluate the convergent and divergent validity of the REDSOCS. Experimentwise error rate was controlled using the Dunn-Bonferroni correction, which required $p < .003$ for correlations to be considered significant. Results of these analyses are shown in Table 3. Significant correlations between the SESBI-R Intensity Scale and the three primary REDSOCS categories (i.e., Inappropriate Behavior, Noncompliant Behavior, and Off-Task Behavior) of the REDSOCS provided support for the convergent validity of the REDSOCS, accounting for 21 to 27% of the shared variance. Significant correlations between the CTRS-R: L Oppositional Scale and the Inappropriate and Off-Task Behavior categories of the REDSOCS provided further support for the convergent validity of the REDSOCS, accounting for 18 to 19% of the shared variance. Although correlations between the CTRS-R: L Inattentive subscale and the REDSOCS Off-Task Behavior category and between the CTRS-R: L Hyperactive-Impulsive subscale and the REDSOCS Inappropriate Behavior category were not significant (p

Table 2
Interobserver Reliability of REDSOCS Category Codes

Category Codes	Percent Agreement	Kappa
Inappropriate Behavior	61%	.72
Compliant Behavior	56%	.69
Noncompliant Behavior	47%	.83
No Command Given	90%	.94
On-Task Behavior	85%	.92
Off-Task Behavior	56%	.69
Not Applicable	75%	.95

Note. Reliability averaged across 59 reliability sessions.

Table 3
Correlations Between REDSOCS Category Scores and Teacher Ratings

Teacher-Rating Scale	REDSOCS Behavior Category					
	Inappropriate		Noncompliant		Off-Task	
	r	p	r	p	r	p
Convergent Validity						
Sutter-Eyberg Student Behavior Inventory-Revised						
Intensity Scale	.52*	< .001	.48*	< .001	.46*	< .001
Conners Teacher Rating Scale-Revised: Long Version						
Oppositional Scale	.44*	.001	.42*	.001	.36	.005
DSM-IV Inattentive Scale	.25	.057	.16	.223	.35	.008
DSM-IV Hyperactive-impulsive Scale	.34	.009	.30	.021	.27	.043
Divergent Validity						
Conners Teacher Rating Scale-Revised: Long Version						
Anxious-Shy Scale	.10	.472	.04	.779	.05	.725
Perfectionism Scale	.07	.579	.06	.663	.08	.542

Note. REDSOCS = Revised Edition of the School Observation Coding System.

*p < .003 as the established level of significance according to the Dunn-Bonferroni correction.

= .008 and .009, respectively), the shared variance for the relationship between these variables was 12%. Correlations between the primary REDSOCS categories and the CTRS-R: L subscales Anxious-shy and Perfectionism were not significantly related to any of the REDSOCS categories, as expected (p value range = .472 to .779), and yielded low effect sizes (r range = .04 to .10), supporting the divergent validity of the three primary REDSOCS categories of classroom behavior.

Treatment Sensitivity

Paired-samples t tests were used to examine the treatment sensitivity of the three primary REDSOCS categories for the subset of children ($n = 34$) who completed posttreatment REDSOCS observations. Analyses of pre to posttreatment scores on REDSOCS categories showed significantly improved scores on the Inappropriate Behavior, $t(33) = 3.10$, $p = .004$, and Off-Task Behavior, $t(33) = 2.72$, $p = .01$, categories after treatment. The frequency of Inappropriate Behavior decreased from 26% ($SD = 15\%$) to 18% ($SD = 11\%$), and the frequency of Off-Task Behavior decreased from 31% ($SD = 17\%$) to 23% ($SD = 13\%$) of the observation intervals. Change in the frequency of Non-compliant Behavior, from 20% ($SD = 15\%$) to 19% ($SD = 14\%$), was not significant, $t(33) = .681$, $p = .50$. These changes in REDSOCS categories are illustrated in Figure 2.

Discussion

This psychometric study of the REDSOCS replicated within a clinical sample the earlier evidence of interobserver reliability as well as evidence of convergent and divergent validity. This study also examined, for the first time, the sensitivity of the REDSOCS to changes following parent training in the behavior of young children with diagnosed disruptive behavior. Together with previous studies (Filcheck et al., 2004; Jacobs et al., 2000), the findings of this study support the designation of REDSOCS as an evidence-based assessment procedure, which requires both replication of psychometric findings (Mash & Hunsley, 2005) and demonstration of treatment sensitivity (McMahon & Frick, 2005).

Reliability of the REDSOCS was assessed by percent agreement and Cohen's kappa. Percent agreement statistics yielded low to acceptable interobserver agreement, whereas Cohen's Kappa yielded acceptable to good interobserver agreement. Percent agreement has the advantage of describing the specific agreement of the occurrence of each behavior category. However, in this study percent agreement was influenced by the relatively low rate of occurrence of some observed behaviors. With low-rate behaviors, only one or two

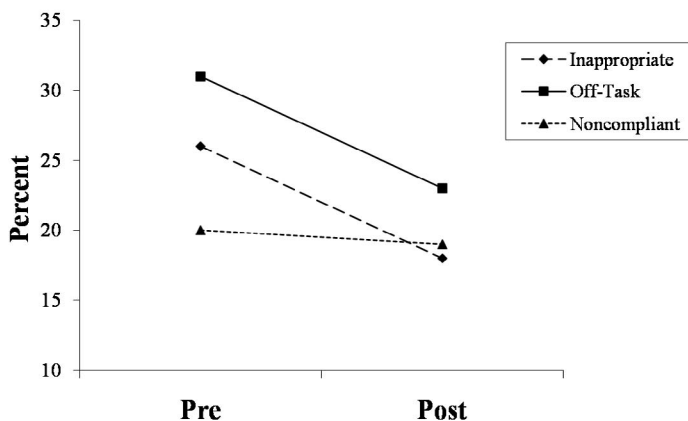


Figure 2. Changes in Negative Behavior Categories

disagreements between observers during an observation session can result in extreme percent agreement reliability estimates. However, Cohen's kappa corrects for such problems by taking the frequency of occurrence of the targeted behaviors into account when computing the agreement between two observer's reports.

For example, if event agreement for noncompliant behavior (a low base rate behavior) occurs on 3 of 60 observed intervals and disagreement on the event occurs in 2 of 60 observed intervals, percent occurrence agreement (agreements/[agreements + disagreements] or 3/5) is a "low" 60%. By contrast, using the formula for kappa ($[(Po - Pe) / (1 - Pe)]$), where Po is the proportion of observed agreement between raters and Pe is the proportion of agreements expected by chance alone, the estimate would be .76, higher than percent reliability and a substantial kappa estimate. Po in the equation would be .97 because although the raters disagreed on 2 of 5 noncompliant behaviors occurring, they agreed on 55 intervals where the noncompliant behavior did not occur, yielding a total reliability estimate of .97 (58/60). Pe was calculated based on the probability that each observer rated the noncompliant behavior as occurring ($[(5/60) * (3/60)]$) plus the probability that each observer rated the noncompliant behavior as not occurring ($[(55/60) * (57/60)]$), yielding a value of $[(.004) + (.870)] = .874$. Therefore, substituting the values for Po and Pe in the kappa formula would result in an estimate of .76.

In comparison to the Jacobs et al. (2000) study of normative preschool and elementary school children, percent agreement was

consistently lower for this sample of children. Cohen's kappa was also somewhat lower than reported in the Jacobs et al. study for some categories (e.g., Inappropriate, Compliant, and Off-Task). The lower reliability scores in this study may be due to the smaller sample size of diagnosed children (DeVon et al., 2007). However, Kappa reliabilities greater than .40 are considered moderate based on standard benchmarks in the field (Landis & Koch, 1977).

Convergent validity was supported by moderate correlations with the SESBI-R Intensity Scale. All three primary behavior categories of the REDSOCS were significantly related to the SESBI-R Intensity Scale, and both the Inappropriate and Noncompliant categories were significantly related to the CTRS Oppositional Scale. Although the Inappropriate Behavior and Off-Task Behavior categories were not significantly related to the other expected CTRS-R: L subscales (i.e., DSM-IV Inattentive and Hyperactive-Impulsive), the correlations were moderate in magnitude, suggesting that re-examination of these relations in a larger sample will be important in future convergent validity studies of the REDSOCS. The correlations between the REDSOCS categories and the SESBI-R Intensity Scale were consistently higher than the correlations between the REDSOCS categories and the CTRS-R: L scales. The higher internal consistency estimates obtained for the SESBI-R Intensity Scale than the CTRS-R: L subscales in this study may have affected the magnitude of the correlations (Haynes & Lench, 2003). Divergent validity was supported by negligible correlations between REDSOCS categories and the CTRS-R: L scales expected not to relate to disruptive classroom behavior, specifically the Anxious-Shy and Perfectionism scales.

Our final set of analyses was conducted to examine the sensitivity of the REDSOCS to changes following a parent-training intervention for disruptive behavior. Although the intervention did not target school behavior problems, the effects of PCIT have previously been found to generalize to the school setting (Funderburk et al., 1998; McNeil et al., 1991). In this study, both the Inappropriate Behavior category and the Off-task Behavior category of the REDSOCS showed significant change in classroom behavior following parent training. Examination of changes in the REDSOCS categories following intervention directly in the classroom would provide a stronger test of treatment sensitivity and will be important in further psychometric evaluation of the REDSOCS.

Results of this study differed from the McNeil et al. (1991) study in not finding significant change in the Noncompliant Behavior category after treatment. This difference in study findings may be due to differences in the children's level of classroom noncompliance at

pretreatment. Presence of school behavior problems was an inclusion criterion in the McNeil et al. (1991) study; Noncompliant Behavior was observed in 46% of the pretreatment observation intervals, in contrast to only 20% in this study. Our lower initial rate of school noncompliance may have left insufficient room for significant change in this category.

In interpreting the findings, three methodological limitations must be considered. First, efforts were not made to keep observers in this study uninformed as to the pre versus posttreatment status of the children, which could have influenced their coding. The larger study was concerned with maintenance treatment, randomizing families after standard treatment completion, and change during standard treatment was not a study question. However, observers were unaware of the hypotheses of the present study. Nevertheless, it will be important to re-examine REDSOCS treatment sensitivity with observers purposefully uninformed as to treatment status. Second, test-retest reliability was not measured and would provide additional psychometric evidence of the stability of the REDSOCS over time. High test-retest reliability would also lend greater credence to the treatment sensitivity findings. However, McNeil et al. (1991) found normal controls and clinic-referred controls demonstrated minimal changes on the Appropriate and Compliance categories following a 12-week wait-list period, suggesting possible test-retest reliability. Third, the sample size was small for examining the psychometric properties of an assessment tool and may have limited the conclusions regarding the convergent validity of the REDSOCS.

Despite these limitations, the results of this study provide evidence sufficient to establish the REDSOCS as an EBA tool for measuring young children's disruptive classroom behaviors. With its evidence base established, this easy-to-code system with relatively brief observation periods has particular potential for use by clinical child and school psychologists as part of routine assessments for school intervention. The Jacobs et al. (2000) study of nonreferred children provides reference point data from multiple classrooms in North Central Florida. Yet, the variability in children's behavior across preschool classrooms suggests the value of collecting "normative" data for an individual child by coding not only the index child's behavior but also the behavior of two or three additional, randomly selected children from the index child's classroom. As described in earlier studies (Funderburk et al., 1998; McNeil et al., 1991), classroom comparison children may be coded during alternating 1-minute observation segments without extending total observation time. For practitioners who routinely conduct school observations, the accumulated data

from classroom control children can quickly provide normative data for children within the practitioner's school district or practice area, which can then provide a target goal for classroom or clinical intervention. The minimal time required initially to learn the REDSOCS system (6 to 8 hours) combined with the brief classroom observation time required to obtain meaningful data support the potential utilization of the REDSOCS for identifying children in need of school intervention and for evaluating the effects of the intervention.

References

- Abikoff, H., & Gittelman, R. (1985). Classroom Observation Code: A modification of the Stony Brook Code. *Psychopharmacology Bulletin*, 21, 901-909.
- Abikoff, H., Gittelman, R., & Klein, D. F. (1980). Classroom Observation Code for hyperactive children: A replication of validity. *Journal of Consulting & Clinical Psychology*, 48, 555-565.
- Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4-18 and 1991 Profile*. Burlington, VT: University of Vermont Department of Psychiatry.
- Achenbach, T. M. (1992). *Manual for the Child Behavior Checklist/2-3 and 1992 profile*. Burlington, VT: University of Vermont Department of Psychiatry.
- Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school forms and profiles*. Burlington: University of Vermont. Research Center for Children, Youth, & Families.
- Angold, A., Costello, E. J., & Erkanli, A. (1999). Comorbidity. *Journal of Child Psychology & Psychiatry & Allied Disciplines*, 40, 57-87.
- Bagner, D. M., Harwood, M. D., & Eyberg, S. M. (2006). Psychometric considerations in child behavioral assessment. In M. Hersen (Ed.), *Handbook of child behavioral assessment* (pp. 63-79). Burlington, MA: Elsevier.
- Conners, C. K. (1969). A teacher rating scale for use in drug studies with children. *American Journal of Psychiatry*, 126, 884-888.
- Conners, C. K. (1997). *Conners' Rating Scales – Revised: Technical Manual*. North Tonawanda, NY: Multi-Health Systems.
- Conners, C. K., Sitarenios, G., Parker, J. D. A., & Epstein, J. N. (1998). Revision and restandardization of the Conners Teacher Rating Scale (CTRS-R): Factor structure, reliability, and criterion validity. *Journal of Abnormal Child Psychology*, 26, 279-291.
- DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., et al. (2007). A psychometric toolbox for

- testing validity and reliability. *Journal of Nursing Scholarship*, 39, 155-164.
- Dodrill, C. B. (1981). An economical method for the evaluation of general intelligence in adults. *Journal of Consulting & Clinical Psychology*, 56, 145-147.
- Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test—Third Edition: Manual*. Circle Pines, MN: American Guidance Services.
- Eyberg, S. M. (1992). Parent and teacher behavior inventories for the assessment of conduct problem behaviors in children. In L. VandeCreek, S. Knapp, & T. L. Jackson (Eds.), *Innovations in clinical practice: A source book* (Vol. 11; pp. 261-270). Sarasota, FL: Professional Resource Press.
- Eyberg, S. M., & Pincus, D. (1999). *Eyberg Child Behavior Inventory and Sutter-Eyberg Student Behavior Inventory-Revised: Professional Manual*. Odessa, FL: Psychological Assessment Resources.
- Filcheck, H. A., Berry, T. A., & McNeil, C. B. (2004) Preliminary investigation examining the validity of the compliance test and a brief behavioral observation measure for identifying children with disruptive behavior. *Child Study Journal*, 34, 1-12.
- Fleiss, J.L. (1981). *Statistical methods for rates and proportions*. New York: Wiley.
- Funderburk, B. W., Eyberg, S. M., Newcomb, K., McNeil, C. B., Hembree-Kigin, T. & Capage, L. (1998). Parent-child interaction therapy with behavior problem children: Maintenance of treatment effects in the school setting. *Child & Family Behavior Therapy*, 20, 17-38.
- Goyette, C. H., Conners, C. K., & Ulrich, R. F. (1978). Normative data on Revised Conners Parent and Teacher Rating Scales. *Journal of Abnormal Child Psychology*, 6, 221-236.
- Haynes, S. N., & Lench, H. C. (2003). Incremental validity of new clinical assessment measures. *Psychological Assessment*, 15, 456-466.
- Jacobs, J. R., Boggs, S. R., Eyberg, S. M., Edwards, D., Durning, P., Querido, J. G., et al. (2000). Psychometric properties and reference point data for the revised edition of the school observation coding system. *Behavior Therapy*, 31, 695-712.
- Kazdin, A. E. (2005). Evidence-based assessment for children and adolescents: Issues in measurement development and clinical application. *Journal of Clinical Child & Adolescent Psychology*, 34, 548-558.

- Koot, H. M., Van Den Oord, E. J., Verhulst, F. C., & Boomsma, D. I. (1997). Behavioral and emotional problems in young preschoolers: Cross-cultural testing of the validity of the child behavior checklist/2-3. *Journal of Abnormal Child Psychology*, 25, 183-196.
- Lahey, B. B., & Loeber, R. (1997). Attention-deficit/hyperactivity disorder, oppositional defiant disorder, conduct disorder, and adult antisocial behavior: A life span perspective. In D. M. Stoff, J. Breiling, & J. D. Maser (Eds.). *Handbook of antisocial behavior* (pp. 51-59). Hoboken, NJ: Wiley.
- Landis, J. R., & Koch G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159 -174.
- Mash, E. J., & Hunsley, J. (2005). Evidence-based assessment of child and adolescent disorders: Issues and challenges. *Journal of Clinical Child & Adolescent Psychology*, 34, 362-379.
- McMahon, R. J., & Frick, P. J. (2005). Evidence-based assessment of conduct problems in children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, 34, 477-505.
- McNeil, C. B., Eyberg, S. M., Eisenstadt, T., Newcomb, K., & Funderburk, B. W. (1991). Parent-child interaction therapy with behavior problem children: Generalization of treatment effects to the school setting. *Journal of Clinical Child Psychology*, 20, 140-151.
- Newcorn, J. H., Halperin, J. M., Jensen, P. S., Abikoff, H. B., Arnold, E., Cantwell, D. P., et al. (2001). Symptom profiles in children with ADHD: Effects of comorbidity and gender. *Journal of the American Academy of Child & Adolescent Psychiatry*, 40, 137-146.
- 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.
- Querido, J. G. & Eyberg, S. M. (2003). Psychometric properties of the Sutter-Eyberg Student Behavior Inventory-Revised with preschool children. *Behavior Therapy*, 34, 1-15.
- Reynolds, C. R., & Kamphaus, K. W. (2004). *Behavioral assessment system for children* (2nd ed.). Circle Pines, MN: American Guidance Service.
- Shaffer, D., Fisher, P., Lucas, C. P., Dulcan, M. K., & Schwab-Stone, M. E. (2000). NIMH Diagnostic Interview Schedule for Children

- Version IV (NIMH DISC-IV): Description, differences from previous versions, and reliability of some common diagnoses. *Journal of the American Academy of Child & Adolescent Psychiatry*, 39, 28-38.
- Speltz, M., McClellan, J., DeKlyen, M., & Jones, K. (1999). Preschool boys with oppositional defiant disorder: Clinical presentation and diagnostic change. *Journal of the American Academy of Child & Adolescent Psychiatry*, 38, 838-845.
- Wruble, M. K., Sheeber, L. B., Sorensen, E. K., Boggs, S. R., & Eyberg, S. M. (1991). Empirical derivation of child compliance times. *Child & Family Behavior Therapy*, 13, 57-68.
- Zisser, A., & Eyberg, S. M. (2009). Frequent tantrums: Oppositional behavior in a young child, psychotherapeutic perspective. In C. A. Galanter & P. S. Jensen (Eds.) *DSM-IVTR Casebook and Treatment Guide for Child Mental Health* (pp. 380 - 383). Arlington, VA: American Psychiatric Publishing, Inc.

