Journal of Psychoeducational Assessment
2000, *19*, 123-136

CROSS-CONTEXT AGREEMENT OF THE ADJUSTMENT SCALES FOR CHILDREN AND ADOLESCENTS

Barbara A. Schaefer
Marley W. Watkins
The Pennsylvania State University

Gary L. Canivez
Eastern Illinois University

Interobserver agreement of children's problem behavior was assessed using two samples of special education students ages 5 to 18 years. The first sample had observers from the same setting ($N = 71$); the second sample ($N = 182$) had observers from different settings with no concurrent observation. Regular and special education teachers and aides completed the Adjustment Scales for Children and Adolescents (ASCA; McDermott, Marston, & Stott, 1993). Inter- and intraclass correlations were generally significant for both samples, with some exceptions. Substantial interobserver agreement was found for the same-setting sample; however, agreement coefficients were lower for the different-setting sample and some level effects were noted. Overall, interobserver agreement for the ASCA was supported in common settings, but rating variability was evident across classrooms and appears indicative of contextual influences on behavior.

School classrooms are unique contexts in which teachers work to facilitate student learning. Components of this environment include the physical arrangement of the class, implementation of appropriate classroom management procedures to maximize time engaged in learning, and intervention to prevent and address problem behaviors in the classroom (Evertson, Emmer, Clements, & Worsham, 1997). A classroom ecology develops that both reflects the teacher's expectations and tolerance and encompasses the physical, learning, and social environments (Cohen & Spenciner, 1998). The structure and ecology of the classroom and participants' engagement may influence variability in students' classroom behavior (Doyle, 1986). These behaviors, in turn, influence academic outcomes as shown by the behavioral contribution to the prediction of students' academic achievement beyond that attributable to cognitive ability alone (Schaefer & McDermott, 1999).

Problem behaviors in schools range from brief inattention to minor or infrequent inappropriate behavior to major behavioral problems that significantly interfere with learning to crises that pose serious safety concerns (Evertson et

Correspondence concerning this article should be addressed to Barbara A. Schaefer, Department of Educational and School Psychology and Special Education, The Pennsylvania State University, 227 CEDAR Building, University Park, PA 16802-3108. Electronic mail may be sent via Internet to bas19@psu.edu.

al., 1997). Beyond influences on learning, constellations of behavioral and emotional difficulties may reflect child psychopathology. These patterns of maladaptive behaviors are often the subject of concern and a source of referral for evaluation and special services (Lloyd, Kauffman, Landrum, & Roe, 1991; Salvia & Ysseldyke, 1998). A thorough evaluation assesses multiple domains: intellectual, academic achievement, social-emotional, and behavioral functioning (Gresham, 1983; Sattler, 1992). Psychometrically sound standardized assessment tools are a foundation of evaluations because they provide normative information about individuals in comparison to their peers.

Evaluation of student behavior in naturally occurring situations is useful for assessing social-emotional and behavioral difficulties among school-aged students. Direct observation of a child's behavior, reports from parents and teachers, and the child's self-report can provide such behavioral information (Reid, Patterson, Baldwin, & Dishion, 1988). Using behavior rating scales, observers can report the presence, absence, or frequency of adaptive or maladaptive behaviors, and these observations may reflect the situational nature of the behavior (Barkley & Edelbrock, 1987). In schools, teacher-completed behavior rating scales are efficient and effective tools that rely upon objective evaluation by professionals familiar with normative behaviors of the students they instruct (McDermott, 1986). Teachers' unobtrusive observations avoid reactivity induced by unfamiliar observers in the classroom.

For rating scales to prove useful in evaluations, they must demonstrate sufficient score reliability and validity. Of particular importance is interobserver, or interrater, agreement (American Psychological Association, 1985; Anastasi, 1988). A meta-analysis by Achenbach, McConaughy, and Howell (1987) assessed the consistency of ratings between various raters (e.g., parents, teachers, students, mental health workers) based on results from 119 studies. Evidence of observer agreement was substantial for observers within the same setting. Correlations ranged from .40 to .84 (M_r = .64) for teacher ratings and from .18 to .73 (M_r = .59) for parent ratings. Correspondence of ratings from observers in different settings (e.g., home vs. school) was mixed, however, with teacher-parent correlations ranging from -.12 to .61 (M_r = .27). Notably, this meta-analysis relied on correlational evidence of the relationship between observations but *did not* take into consideration possible *level* differences. Omission of such analyses is problematic because correlational analyses alone would not detect whether one set of observers consistently rate students higher or lower than other observers.

Agreement for interval scale data requires the combination of relationship *and* level consistency (McDermott, 1988). Evidence of a statistically significant correlation between ratings and the absence of statistically significant difference in observer means are needed as indications of observer congruence. A more complex approach relies upon ANOVA using a modified intraclass correlation coefficient (ICC; Shrout & Fleiss, 1979) that simultaneously assesses both linear and nonlinear relationships among observer ratings. Because the intraclass coefficient reflects the overall covariation or homogeneity of ratings, it is most properly regarded as an expression of the strength of association rather than agreement per se (McDermott, 1988).

Existing rating scales have demonstrated initial evidence of interobserver reliability for teacher-completed forms based on intercorrelations; however, many do not assess or report level differences. For example, the manual for the Child Behavior Checklist-Teacher Report Form (CBCL-TRF; Achenbach, 1991a, 1991b) reports the scale intercorrelations only for a sample ($N = 207$) of school-aged students. For a large sample of special education students ($N = 635$), correlations and *t*-test results are reported but descriptive statistics are not. For the Behavior Assessment System for Children-Teacher Rating Scales (BASC-TRS; Reynolds & Kamphaus, 1992a, 1992b), the manual reports interrater reliability coefficients and means and standard deviations for ratings on a small sample ($N = 30$) of children; however, analyses of possible mean level differences are omitted.

A mathematical model for agreement or consensus developed by Kenny (1991) encompasses six factors: acquaintance, overlap, shared meaning systems, consistency, extraneous information, and communication. In brief, acquaintance is the amount of information to which a judge is exposed, and overlap is the extent to which judges observe the same set of behaviors. A shared meaning system is the similarity of meaning given to an act by two judges, whereas consistency is the extent to which the target's behavior is the same from one situation to another. Extraneous information is the extent to which judges rate targets on information other than their acts, and, finally, communication is the extent to which judges share impressions of the target with one another (Kenny, 1991). Particularly relevant to this investigation are the factors of overlap and consistency—to what extent do observers in the same setting or in different settings report the same behaviors for selected targets?

The goal of this study was to further investigate interobserver agreement for the Adjustment Scales for Children and Adolescents (ASCA; McDermott, Marston, & Stott, 1993), a teacher-completed behavior rating scale designed to assess psychopathology among school-aged students across various scholastic situations, such as responding to teachers, interacting with peers, handling materials, and playing games. Prior research has demonstrated sufficient interobserver agreement using correlation and mean level differences for observations in the same setting (McDermott, 1994; Watkins & Canivez, 1997). The present study extends earlier work by examining interclass and intraclass correlations for observations within the same classroom setting and across two different classroom contexts. It was hypothesized that (a) mean ratings would not differ significantly for observations in the same classroom or from different classrooms, (b) interobserver agreement as determined by inter- *and* intraclass correlations would be significant for observers in both the same and different classrooms, and (c) observers in the same classroom situation would show greater consistency and agreement than those from different classrooms.

METHOD

Participants

Two samples of school-aged students participated in the study. The first sample consisted of 71 students (24 females, 47 males) attending one school in the

Midwest and five schools in the southwestern regions of the U.S. Participants ranged in age from 7 to 18 ($M = 10.9$, $SD = 2.7$) and attended grades 1 through 10. Participants were classified as Attention Deficit/Hyperactivity Disordered (1.4%), Seriously Emotionally Disturbed (26.8%), Specific Learning Disabled (39.4%), Mentally Retarded (8.5%), Speech/Language Impaired (19.7%), or not categorized (4.2%). The sample was primarily Caucasian (85%) and Hispanic/Latino (12%), with a small percentage (3%) of other race/ethnicities represented.

The second sample of 182 students in kindergarten through sixth grade was drawn from nine schools in the west and midwestern U.S. Of the 182, 63 (35%) were female and 119 (65%) were male. All students were enrolled in special education programs, and all but 2 were classified as having a specific learning disability. Students ranged in age from 5 to 13 ($M = 10.1$, $SD = 1.9$). As with the first sample, most were Caucasian (92%), with some Hispanic/Latino (6%) and Black/African American (2%) students participating.

Instrument

The Adjustment Scales for Children and Adolescents (ASCA; McDermott et al., 1993) is a teacher-completed behavior rating form designed to provide normative comparison information regarding students' classroom behaviors. Normed on a stratified, nationally representative sample of 1,400 students in grades K through 12, the ASCA is comprised of behavioral items (129 problem behavior, 26 positive behaviors) that encompass a variety of school situations, such as coping with new learning, getting along with peers, and interacting with the teacher. The scale is unique in its reliance upon observation of similar problem behaviors across multiple situations, rather than rating the frequency or intensity of symptoms in a checklist format (McDermott, 1994).

Ninety-seven problem behavior items contribute to six core and two supplemental syndromes determined via factor analytic techniques (McDermott, 1994). Core syndromes include Oppositional Defiant, Solitary Aggressive-Provocative, Solitary Aggressive-Impulsive, Attention Deficit Hyperactive, Avoidant, and Diffident—each named to represent the component behavioral items most heavily loaded on each factor. Two supplemental syndromes are applicable only for certain subgroups: Delinquent for girls over age 11 and boys ages 5 to 17, and Lethargic for students ages 11 and under. When submitted to second-order factor analyses, the core syndromes load on two global summary scales: Overactivity and Underactivity. Each syndrome demonstrates sufficient specificity and invariance across subgroups based on sex, age, and race/ethnicity (McDermott, 1994). Raw scores are converted to T scores ($M = 50$, $SD = 10$), with scores of 67 or higher considered "Maladjusted" and representing behavior more severe than 95% of students; scores between 60 and 66 considered "At Risk" and more extreme than 85% of students; and scores below 60 considered "Adjusted" (McDermott, 1994).

As presented in the ASCA manual (McDermott, 1994), sufficient internal consistency has been demonstrated for each of the core syndromes and global Overactivity and Underactivity scales, with moderately high coefficents ($r_\alpha \geq .70$). Similarly, test-retest stability coefficients were significant and sufficient, and interrater coefficients for a small sample were promising. Subsequent work

with a larger sample by Watkins and Canivez (1997) demonstrated substantial interobserver agreement based on correlation and mean comparisons across observers.

Procedure

For the first sample, two observers who simultaneously observed the student for at least 1 hour per day in the same classroom (e.g., self-contained special education classes or resource room settings) were identified. Each observer was either a professional or paraprofessional willing to participate in the study, and their job classifications included special education teachers and aides as well as classroom and remedial reading teachers. Fifty-eight percent of the observer pairings were special education teachers and special education aides in self-contained classrooms. Classroom teachers paired with special education teachers and remedial reading teachers were also included. In all, 29 raters participated from 24 classrooms.

For the second sample, 137 observers from unique classroom settings were identified and agreed to participate in the study. Unlike sample 1 observers, sample 2 observers did not share concurrent observation of participants within the same classroom environment. Observers included regular education classroom teachers and special education teachers. Most observers rated a single participant, and the most participants rated by any one observer was 15. For both samples, students were rated following sufficient opportunity for teachers to become familiar with the students (i.e., at least 40 to 50 school days) as recommended in the ASCA manual.

Data Analyses

Application of interclass and intraclass approaches determined interobserver agreement. Using the interclass strategy, a two-step process was applied. First, relative ranking and direction of ratings were determined using Pearson *r*s. Then, observer mean level differences were assessed using *t* tests, with failure to reject the null hypothesis of equality of means considered indicative of similarity in ratings (McDermott, 1988). The intraclass approach applied the two-way random effects model of ICC (Hamer, 1990; Shrout & Fleiss, 1979), which is useful when absolute agreement among measurements is the goal regardless of the observer (Buchanan, McDermott, & Schaefer, 1998; Cho, 1981; McGraw & Wong, 1996). Further analyses of absolute score differences between pairs of teacher ratings of the same student were also conducted.

RESULTS

Table 1 presents the means and standard deviations for the ASCA core, supplementary, and global adjustment syndrome ratings by observers for both samples. Notably, although mean scores ranged from 48.1 to 60.3, most are significantly above 50, with some means an entire standard deviation higher, indicative of higher-than-average behavior problems. Significant differences between observer ratings for Underactivity for both simultaneous and independent observers are shown, as well as for Avoidant and Solitary Aggressive (Impulsive) syndromes for cross-context observations.

Table 1
T-Score Means and Standard Deviations for Adjustment Scale Dimensions by Observer

ASCA scale/syndrome	Same classroom[a]		Different classroom[b]	
	Observer A *M* (*SD*)	Observer B *M* (*SD*)	Observer C *M* (*SD*)	Observer D *M* (*SD*)
Overactivity	58.3 (8.9)	58.1 (9.3)	55.2 (9.9)	55.1 (8.8)
Attention-Deficit Hyperactive	55.3 (9.9)	55.9 (10.1)	55.2 (9.7)	55.4 (9.5)
Solitary Aggressive (Provocative)	57.2 (12.4)	57.6 (12.0)	53.4 (11.6)	52.9 (11.0)
Solitary Aggressive (Impulsive)	51.1 (10.2)	49.8 (9.2)	51.8 (9.6)*	49.2 (6.8)*
Oppositional Defiant	60.3 (13.2)	59.6 (14.5)	51.8 (11.7)	52.1 (11.1)
Underactivity	53.5 (10.2)**	51.0 (11.1)**	52.6 (9.9)**	48.3 (10.0)**
Diffident	53.6 (10.5)	51.1 (10.5)	51.1 (9.6)	49.0 (10.0)
Avoidant	53.9 (10.5)	55.2 (11.3)	52.8 (11.3)**	48.1 (8.8)**
Supplemental				
Delinquent	54.6[c] (12.6)	56.1[c] (13.0)	51.0[d] (10.9)	48.4[d] (8.6)
Lethargic	58.8[e] (11.4)	56.5[e] (11.1)	54.9[f] (11.1)	52.7[f] (10.5)

Note.—ASCA = Adjustment Scales for Children and Adolescents. The Bonferroni correction (Dunn, 1961) was applied to account for family-wise Type I error.
[a]$N = 71$. [b]$N = 182$. [c]$n = 56$. [d]$n = 139$. [e]$n = 36$. [f]$n = 143$.
*$p < .05$. **$p < .01$.

Table 2 displays the inter- and intraclass correlations, which were identical or very similar for both approaches. For observers who shared at least 1 hour of simultaneous observation of the target participant, coefficients were substantial and significant, ranging from .61 to .85. All are above .60 and are considered sufficient because most variance is not error variance (Widaman, 1993). Agreement coefficients for observers from separate classrooms were largely

Table 2
Interobserver Agreement Coefficients for Adjustment Scale Dimensions

ASCA scale/syndrome	Same classroom[a]		Different classroom[b]	
	Interclass	Intraclass	Interclass	Intraclass
Overactivity	.83***	.83***	.57***	.57***
Attention-Deficit Hyperactive	.72***	.72***	.51***	.51***
Solitary Aggressive (Provocative)	.80***	.80***	.48***	.49***
Solitary Aggressive (Impulsive)	.61***	.61***	.16	.15[†]
Oppositional Defiant	.72***	.72***	.54***	.54***
Underactivity	.85***	.83***[‡]	.41***	.38***[†]
Diffident	.72***	.70***	.31***	.30***
Avoidant	.66***	.66***	.42***	.37***[†]
Supplemental				
Delinquent	.83[c]***	.83[c]***	.21[d]	.20[d]
Lethargic	.69[e]***	.69[e]***	.40[f]**	.39[f]**

Note.—ASCA = Adjustment Scales for Children and Adolescents. Interclass = correlation coefficient *r*. Intraclass = Intraclass Correlation Coefficient (ICC; Hamer, 1990, Shrout & Fleiss, 1979). The Bonferroni correction (Dunn, 1961) was applied to account for family-wise Type I error.
[a]$N = 71$. [b]$N = 182$. [c]$n = 56$. [d]$n = 139$. [e]$n = 36$. [f]$n = 143$.
*$p < .05$. **$p < .01$. ***$p < .001$. [‡]$p < .05$ effect for level.

significant but not as substantial, ranging from a modest .30 to .57, with two nonsignificant exceptions—Solitary Aggressive (Impulsive; .15 and .16) and Delinquent (.20 and .21). Although ASCA ratings were largely comparable in terms of pattern, rank order, and level, some intraclass level effects were found. Level effects for Underactivity in both samples and for Avoidant and Solitary Aggressive (Impulsive) ratings from separate classrooms indicated variability in ratings. All correlation coefficients for same-classroom observations were significantly higher ($p < .05$) than those obtained for the comparable dimension from different-classroom observations.

Further analysis of absolute mean level differences between pairs of observer ratings of the same student was conducted (see Table 3). Mean absolute score differences ranged from 2.55 to 5.03, whereas median differences were minimal (0 or 2 points) for same-classroom observations. In contrast, observations from different classrooms evidenced increased variability with generally greater ranges, means, and medians. Means ranged from 5.40 to 8.80, with the largest median difference of 8 points for Underactivity. Overall, the average mean score differences for same-classroom and different-classroom observations were 4.21 and 7.05, respectively, and the majority of students' average absolute rating differences were 10 points or less (87.3% and 83.5%, respectively).

Table 3
Absolute Difference Score Ranges, Means, and Medians between Pairs of Teacher Ratings on Adjustment Scale Dimensions

	Same classroom[a]			Different classroom[b]		
ASCA scale/syndrome	Range	*M* (*SD*)	Median	Range	*M* (*SD*)	Median
Overactivity	0-17	3.51 (3.97)	2.0	0-24	6.39 (5.90)	5.0
Attention-Deficit Hyperactive	0-23	5.03 (5.61)	2.0	0-28	6.79 (6.72)	4.0
Solitary Aggressive (Provocative)	0-24	3.13 (7.09)	0	0-29	6.07 (9.75)	0
Solitary Aggressive (Impulsive)	0-25	3.90 (7.74)	0	0-28	5.40 (9.76)	0
Oppositional Defiant	0-36	5.78 (8.74)	2.0	0-32	6.97 (8.48)	2.0
Underactivity	0-16	3.96 (4.90)	2.0	0-35	8.80 (7.61)	8.0
Diffident	0-19	4.97 (6.60)	0	0-27	8.39 (8.10)	7.0
Avoidant	0-25	4.65 (7.83)	0	0-31	7.81 (9.12)	1.5
Supplemental						
Delinquent	0-27[c]	2.55 (7.06)	0	0-30[d]	6.32 (10.96)	0
Lethargic	0-27[e]	4.61 (7.80)	0	0-30[f]	7.59 (9.39)	0

Note.—ASCA = Adjustment Scales for Children and Adolescents.
[a]$N = 71$. [b]$N = 182$. [c]$n = 56$. [d]$n = 139$. [e]$n = 36$. [f]$n = 143$.

DISCUSSION

Interobserver agreement of classroom behavior ratings for two samples of students in special education was investigated. Most of the ratings reflected comparable assessments of behavior, with some exceptions. For simultaneous observers, behavior ratings were similar except that a level effect for the

Underactivity syndrome was found. The various combinations of constituent rater pairs for the sample precluded further evaluation of possible differences related to observer status (e.g., professionals vs. paraprofessionals). For observers from different classrooms, ratings of students' problem behaviors such as impulsive aggression and underactivity, particularly avoidant behaviors, were higher in regular education classrooms than in special education or resource rooms. Notably, ratings of students in self-contained classrooms revealed higher mean levels of problem behaviors, such as provocative aggression and oppositional defiance, than among those participants not in self-contained classrooms. This finding comports with the expectation that more seriously behavior disordered students are less likely to participate in the regular education instructional environment. Furthermore, behavior ratings for both samples are generally above what would be expected in the normative population and reflect the tendency of students in special education to manifest greater behavioral deviancy than students in regular education.

Substantial interobserver agreement was supported for the ASCA using both interclass and the more conservative intraclass correlation techniques for ratings from observers in the same setting. Congruence demonstrated in observations provided evidence that ratings were not idiosyncratic to the observer. In contrast, ratings agreement for observations from different classroom settings was lower than those from within the same setting. Although most coefficients were significant, they were only moderate in magnitude. These findings are comparable to Molina, Pelham, Blumenthal, and Galiszewski's (1998) evaluation of rater agreement in an adolescent sample. Using three behavioral rating measures obtained from two to five teachers, Molina et al. similarly found only low to moderate agreement for multiple teacher observations. Similar results were also reported for the CBCL-TRF, with coefficients ranging from .30 to .68 (M_r = .54) for teachers from different classrooms rating students referred for evaluation (Achenbach, 1991b). Given the high level of agreement found for observations in the same setting, the lower levels of agreement from different settings appear to reflect behavior variability related to the distinct settings and contexts in which the students receive instruction.

The lower level of cross-setting agreement may also reflect contextual sensitivity to behavior disturbance in regular and special education environments. Current results revealed that students were rated higher on some internalizing and externalizing scales of abnormal behavior by regular education teachers than by special educators or resource room instructors. Previous work by Ritter (1989) found that regular educators rated students higher than special educators for externalizing behavior, whereas no difference was found for internalizing behavior. This finding is not consistent, however, given that Simpson (1991) found the reverse: special educators rated students higher on behavior problems than did regular educators. Previous work by Safran and Safran (1984, 1985) suggested that regular educators are less tolerant of externalizing types of problem behavior than special educators and concluded that teacher judgments of problem behavior were not independent of the classroom environment but rather reflected the classroom context. Further evidence of contextual influences lies in Brandon, Kehle, Jenson, and Clark's (1990) study of various rater effects using videotaped scenarios and teacher ratings. Teachers

were consistent raters because no effects were found for teacher expectation, regression, or ratings practice; however, a presentation order effect was found. Brandon et al. argued that raters apparently use other students as reference points. Extraneous information such as peer or classmate constellations in school settings may be somewhat influential upon behavior ratings; however, the possible contributions to behavior outcomes could not be isolated in this study.

Relatively few studies of agreement have been conducted with children in special education classes in the school setting (Costenbader & Keller, 1990). One interobserver study of a special education sample by Achenbach (1991b) found that special education teachers rated students higher for 10 of 11 problem behavior scales than did special education teacher aides. Notably, interobserver correlations ranged from .27 to .69 (M_r = .49) and are considerably lower than those from sample 1 in the current study. By choosing a special education sample and focusing on behavior pathology, it may have been more difficult to find rating agreement because agreement has been found to be generally higher for adaptive, normal behavior than for maladaptive, problem behavior (Voelker, Shore, Hakim-Larson, & Bruner, 1997). According to Victor, Halverson, and Wampler (1988), agreement is more readily found for "easy" kids and disagreement for "difficult" ones using intraclass correlation. Temperamentally difficult children exhibit varying degrees of emotional and behavioral lability at home and at school, and children's problem behaviors are often inconsistent, characterized by short histories, change, and variability across settings (Reid et al., 1988; Victor et al., 1988). As current results reveal, behavior may vary even from one school setting to another. Indeed, symptom severity of most forms of childhood psychopathology is affected by situational and contextual factors (Barkley, 1996). For example, fluctuations in behavioral symptoms of ADHD have been documented in school contexts (Barkley & Edelbrock, 1987; DuPaul & Barkley, 1992), and the salience of ADHD differs depending on the setting and situation in which behavior is observed (Milich & Landau, 1988). The present study, however, does not reveal significant variability for ADHD behaviors as reported by teachers.

According to Verhulst and Akkerhuis (1989), agreement for externalizing behavior has been superior to that for internalizing behavior. For the CBCL-TRF, coefficients of .66 and .69 for externalizing and .41 and .44 for internalizing broad-band scales were reported for referred and special education samples, respectively (Achenbach, 1991b). Similarly on the BASC-TRS, the manual presents reliability coefficients of .79 and .69, with externalizing higher than internalizing (Reynolds & Kamphaus, 1992b). Current results reveal similar levels of agreement for same-setting observations of overall Overactivity and Underactivity, with coefficients from .83 to .85; however, these coefficients drop to approximately .57 and .40, respectively, for observations from different settings, reflecting somewhat greater agreement for externalizing behaviors.

The notable exception concerning externalizing behaviors is the nonsignificant agreement coefficients for Solitary Aggressive (Impulsive) and Delinquent. Previous work by Milich and Landau (1988) compared teacher ratings with classroom observations for inattention and aggression among boys in different classroom situations (large group, small group, and individual seat-

work). Teachers' ability to distinguish between hyperactivity and aggression was established, and some preliminary evidence of aggression as a function of setting variability was shown. Some researchers argue that adults' perceptions and students' behavior are "dyad- and situation-specific" (Jensen, Xenakis, Davis, & Degroot, 1988, p. 454). For example, Reid and Patterson (1991) pointed out that interactions with parents, teachers, and peers are the proximal and crucial determinants of aggression in various settings and, therefore, are critical in intervention. According to these authors, the social setting in which disruptive, antisocial, or aggressive behavior occurs encompasses contingencies that are powerful determinants of aggressive behavior. Peer rejection in particular can play an important role in aggressive behavioral responses among youth (see Coie & Lenox, 1994). Indeed, current results indicate that impulsive aggression appears to be particularly sensitive to contextual influences.

Other types of externalizing behavior can be considered on the covert-overt and destructive-nondestructive continua and the categorization of disruptive child behavior as property violations, aggression, status violations, and oppositional as presented by Frick et al. (1993). Whereas ASCA's Solitary Aggressive (Impulsive) items reflect verbal and physical outbursts and lesser property violations, Solitary Aggressive (Provocative) items are primarily overt interpersonal verbal and physical attacks or destruction of property. Greater behavioral consistency across settings may exist for provocative versus impulsive aggression. In contrast, Delinquent items on the ASCA reflect status and property violations that tend to be more covert in nature and rarely occur within the classroom (e.g., substance use or distribution, truancy, association with gangs or troublesome youth, property damage, and carrying weapons; see McDermott & Schaefer, 1996). Differing levels of teachers' awareness of students' activities outside the classroom environment may underlie the low level of agreement. It is plausible that greater communication about students' extracurricular activities occurs between teacher raters in the same setting than would occur between teachers from different classrooms, which may help explain the higher level of agreement in sample 1 and the lack of agreement for this scale for observers with no concurrent observation of the target students in sample 2.

Different informants may have unique perspectives on students' interactions with peers or adults, and students' interpersonal relations may vary in different settings or situations. Not surprisingly, behavioral reports may reflect disagreement as Verhulst and Akkerhuis (1989) suggested that situational variation in behavior is the key question, and the types of behavior (e.g., externalizing or internalizing) as well as the age and other characteristics of the target students must also be considered. Relevant to this investigation are what Merrell (1994) identified as source variance (i.e., possible response biases in how raters respond to the format of the scale) and setting variance (i.e., behavioral "situational specificity" [p. 69] related to differing contingencies present in two environments). Although these are considered types of error variance possible with behavior rating scales, it is not necessarily the case that disagreement in behavior ratings should call into question the verity of teacher reports. Instead, broader sampling of students' behaviors in various settings is encouraged, as is the use of multiple raters in evaluating students, as previously argued by others

(e.g., Achenbach et al., 1987; Molina et al., 1998; Suen, Logan, Neisworth, & Bagnato, 1995).

Present results provide preliminary support for situational sensitivity in the behavioral responses of special education students; however, this study is limited by several factors. First, it includes only special education students from geographically limited areas and participating students were primarily preadolescents, so results may not be generalizable to other populations. Second, different samples of target students and teacher observers were used to assess interobserver agreement in the same and separate classrooms. Ideally, collection of independent ratings of the same target students by two observers in each setting would be helpful; however, very few, if any, schools have multiple instructors or paraprofessionals in each classroom. Future research using more ethnically, socioeconomically, and geographically diverse samples of regular and special education students across the entire school-age range would be beneficial. Additional work exploring the relative impact of school context and situations, particularly peer and teacher-student interactions, might be helpful. Perhaps most important, empirical assessment of the contribution of multiple teacher ratings to the prediction of student outcomes would be beneficial.

In conclusion, this study found high levels of interobserver agreement using inter- and intraclass correlation coefficients for observers in the same setting but lower levels of agreement for ratings from different settings. Together, these findings provide indications of contextual influences on students' behavior. As Achenbach and McConaughy (1987) noted, behavior can be affected by intrapersonal intrinsic and environmental factors, both of which are variable and likely to affect variability in emotional and behavioral responses across situations and time. Assessment of students' behavior in multiple settings and by multiple raters will continue to be beneficial in the evaluation of children and students being considered for special services.

REFERENCES

Achenbach, T. M. (1991a). *Child Behavior Checklist-Teachers' Report Form.* Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (1991b). *Manual for the Teachers' Report Form and 1991 Profile.* Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M., & McConaughy, S. H. (1987). *Empirically based assessment of child and adolescent psychopathology: Practical applications.* Newbury Park, CA: Sage.

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

American Psychological Association. (1985). *Standards for educational and psychological testing.* Washington, DC: Author.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Barkley, R. A. (1996). Attention-Deficit/Hyperactivity Disorder. In E. J. Mash & R. A. Barkley (Eds.), *Child psychopathology* (pp. 63–112). New York: Guilford.

Barkley, R. A., & Edelbrock, C. (1987). Assessing situational variation in children's problem behaviors: The Home and School Situations Questionnaires. In R. J. Prinz (Ed.), *Advances in behavioral assessment of children and families* (Vol. 3, pp. 157–176.). Greenwich, CT: JAI Press.

Brandon, K. A., Kehle, T. J., Jenson, W. R., & Clark, E. (1990). Regression, practice, and expectation effects on the Revised Conners Teacher Rating Scale. *Journal of Psychoeducational Assessment, 8,* 456–466.

Buchanan, H. H., McDermott, P. A., & Schaefer, B. A. (1998). Agreement among classroom observers of children's stylistic learning behaviors. *Psychology in the Schools, 35*(4), 1–7.

Cho, D. W. (1981). Inter-rater reliability: Intraclass correlation coefficients. *Educational and Psychological Measurement, 41,* 223–226.

Cohen, L. G., & Spenciner, L. J. (1998). *Assessment of children and youth.* New York: Addison-Wesley Longman.

Coie, J. D., & Lenox, K. F. (1994). The development of antisocial individuals. In D. C. Fowles, P. Sutker, & S. H. Goodman (Eds.), *Progress in experimental personality & psychopathology research* (pp. 45–72). New York: Springer.

Costenbader, V. K., & Keller, H. R. (1990). Behavioral ratings of emotionally handicapped, learning disabled, and nonreferred children: Scale and source consistency. *Journal of Psychoeducational Assessment, 8*(4), 485–496.

Doyle, W. (1986). Classroom organization and management. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 392–431). New York: Macmillan.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56,* 52–64.

DuPaul, G. J., & Barkley, R. A. (1992). Situational variability of attention problems: Psychometric properties of the Revised Home and School Situations Questionnaires. *Journal of Clinical Child Psychology, 21*(2), 178–188.

Evertson, C. M., Emmer, E. T., Clements, B. S., & Worsham, M. E. (1997). *Classroom management for elementary teachers* (4th ed.). Boston: Allyn & Bacon.

Frick, P. J., Lahey, B. B., Loeber, R., Tannenbaum, L., Van Horn, Y., Christ, M. A. G., Hart, E. A., & Hanson, K. (1993). Oppositional defiant disorder and conduct disorder: A meta-analytic review of factor analyses and cross-validation in a clinic sample. *Clinical Psychology Review, 13,* 319–340.

Gresham, F. M. (1983). Multitrait-multimethod approach to multifactored assessment: Theoretical rationale and practical application. *School Psychology Review, 12*(1), 26–34.

Hamer, R. M. (1990). Intraclass correlations [Computer code]. Cary, NC: SAS Institute.

Jensen, P. S., Xenakis, S. N., Davis, H., & Degroot, J. (1988). Child psychopathology rating scales and interrater agreement: II. Child and family characteristics. *Journal of the American Academy of Child & Adolescent Psychiatry, 27*(4), 451–461.

Kenny, D. A. (1991). A general model of consensus and accuracy in interpersonal perception. *Psychological Review, 98*(2), 155–163.

Lloyd, J. W., Kauffman, J. M., Landrum, T. J., & Roe, D. L. (1991). Why do teachers refer pupils for special education? An analysis of referral records. *Exceptionality, 2,* 115–126.

McDermott, P. A. (1986). The observation and classification of exceptional child behavior. In R. T. Brown & C. R. Reynolds (Eds.), *Psychological perspectives on childhood exceptionality: A handbook* (pp. 136–180). New York: Wiley.

McDermott, P. A. (1988). Agreement among diagnosticians or observers: Its importance and determination. *Professional School Psychology, 3,* 225–240.

McDermott, P. A. (1994). *National profiles in youth psychopathology: Manual of Adjustment Scales for Children and Adolescents.* Philadelphia: Edumetric and Clinical Science.

McDermott, P. A., Marston, N. C., & Stott, D. H. (1993). *Adjustment Scales for Children and Adolescents.* Philadelphia: Edumetric and Clinical Science.

McDermott, P. A., & Schaefer, B. A. (1996). A demographic survey of rare and common problem behaviors and common problem behaviors among

among American students. *Journal of Clinical Child Psychology, 25*, 352–362.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46.

Merrell, K. W. (1994). *Assessment of behavioral, social, and emotional problems: Direct and objective methods for use with children and adolescents.* New York: Longman.

Milich, R., & Landau, S. (1988). Teacher ratings of inattention/overactivity and aggression: Cross-validation with classroom observations. *Journal of Clinical Child Psychology, 17*(1), 92–97.

Molina, B. S. G., Pelham, W. E., Blumenthal, J., & Galiszewski, E. (1998). Agreement among teachers' behavior ratings of adolescents with a childhood history of attention deficit hyperactivity disorder. *Journal of Clinical Child Psychology, 27*(3), 330–339.

Reid, J. B., & Patterson, G. R. (1991). Early prevention and intervention with conduct problems: A social interactional model for the integration of research and practice. In G. Stoner, M. R. Shinn, & H. M. Walker (Eds.), *Interventions for achievement and behavior problems.* Bethesda, MD: National Association of School Psychologists.

Reid, J. B., Patterson, G. R., Baldwin, D. V., & Dishion, T. J. (1988). Observations in the assessment of childhood disorders. In M. Rutter, A. H. Tuma, & I. S. Lann (Eds.), *Assessment and diagnosis in child psychopathology* (pp. 156–195). New York: Guilford Press.

Reynolds, C. R., & Kamphaus, R. W. (1992a). *Behavior Assessment System for Children Manual.* Circle Pines, MN: American Guidance Service.

Reynolds, C. R., & Kamphaus, R. W. (1992b). *Behavior Assessment System for Children–Teacher Rating Scales.* Circle Pines, MN: American Guidance Service.

Ritter, D. R. (1989). Teachers' perceptions of problem behavior in general and special education. *Exceptional Children, 55*(6), 559–564.

Safran, J. S., & Safran, S. P. (1985). Teachers' judgments of problem behaviors. *Exceptional Children, 54*(3), 240–244.

Safran, S. P., & Safran, J. S. (1984). Elementary teachers' tolerance of problem behaviors. *The Elementary School Journal, 85*(2), 237–243.

Salvia, J., & Ysseldyke, J. E. (1998). *Assessment* (7th ed.). Boston: Houghton Mifflin.

Sattler, J. M. (1992). *Assessment of children* (Rev. 3rd ed.). San Diego, CA: Author.

Schaefer, B. A., & McDermott, P. A. (1999). Learning behavior and intelligence as explanations for children's scholastic achievement. *Journal of School Psychology, 37*(3), 299–313.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Simpson, R. G. (1991). Agreement among teachers of secondary students in using the Revised Behavior Problem Checklist to identify deviant behavior. *Behavior Disorders, 17*(1), 66–71.

Suen, H. K., Logan, C. R., Neisworth, J. T., & Bagnato, S. (1995). Parent-professional congruence: Is it necessary? *Journal of Early Intervention, 19*(3), 243–252.

Verhulst, F. C., & Akkerhuis, G. W. (1989). Agreement between parents' and teachers' ratings of behavioral/emotional problems of children aged 4–12. *Journal of Child Psychiatry, 30*(1), 123–136.

Victor, J. B., Halverson, C. F. Jr., & Wampler, K. S. (1988). Family-school context: Parent and teacher agreement on child temperament. *Journal of Consulting and Clinical Psychology, 56*, 573–577.

Voelker, S., Shore, D., Hakim-Larson, J., & Bruner, D. (1997). Discrepancies in parent and teacher ratings of adaptive behavior of children with multiple disabilities. *Mental Retardation, 35*(1), 10–17.

Watkins, M. W., & Canivez, G. L. (1997). Interrater agreement of the Adjustment Scales for Children and Adolescents. *Diagnostique, 22,* 205–213.

Widaman, K. F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research, 28,* 263–311.