INTERRATER AGREEMENT OF THE ADJUSTMENT SCALES FOR CHILDREN AND ADOLESCENTS

Marley W. Watkins, *Pennsylvania State University*
Gary L. Canivez, *Eastern Illinois University*

Standardized behavior rating scales and checklists offer unobtrusive evaluations of students' behavior in natural social environments. This study investigated the interrater agreement of the Adjustment Scales for Children and Adolescents (ASCA), a behavior rating scale used in school settings. Participants were 71 students enrolled in a variety of special programs who were rated by 29 observers in 24 classrooms. Resulting interrater reliability coefficients were substantial, and level differences, although significant, were not clinically meaningful. It was concluded that the ASCA produced acceptable levels of interrater agreement when educational professionals and paraprofessionals observed exceptional students within a common environment.

Modern psychoeducational research and practice reveals an increasing preference for objective, rather than inferential, assessment methods that can facilitate a link between assessment and intervention (Power & Ikeda, 1996; Reschly & Ysseldyke, 1995). In the socioemotional and behavioral realm, standardized behavior rating scales and checklists have achieved popularity among school and clinical psychologists (Merrell, 1994). Among school psychologists they are the most frequently used instruments in assessing emotional and behavioral difficulties of youth (Stinnett, Havey, & Oehler-Stinett, 1994). Knoff (1995) stated that behavior rating scales are "one of the most efficient, sound, and effective ways ... to identify a referred student's behavioral strengths and weaknesses..." (pp. 857). McConaughy and Ritter (1995) noted that use of behavior rating scales is "best practice" in assessing emotional and behavioral disorders.

Behavior rating scales offer, among other advantages, unobtrusive evaluations of students' behavior in such natural social settings as schools, classrooms, and homes. Within the school and classroom, teachers are natural observers and informants because they have the comparative experience of observing many students across time and varied social contexts. As such, they appear to take a normative perspective in rating diffi-

Correspondence regarding this article should be addressed to Marley W. Watkins, Pennsylvania State University, 227 CEDAR Building, University Park, PA 16802.

DIAGNOSTIQUE 22(4), 205-213(1997)

culties in children. Consequently, teachers have sometimes been considered to be among the more accurate adult raters of child behavior (Kamphaus & Frick, 1996).

Regardless of the informant, behavior rating scales, like all tests, must demonstrate acceptable psychometric properties before they can be validly applied in practice. Edelbrock (1983) reported that existing behavioral rating scales differ across a number of psychometric dimensions. One critical technical property of any instrument that relies upon informant reports is the degree to which two informants, or raters, agree. This interrater, or interobserver agreement, measures the extent to which conclusions drawn from the instrument vary as a function of the rater, not the student being rated. This is an important distinction because Martin, Hooper, and Snow (1986) have reported that the subjectivity of raters is the primary source of error in rating scale data. For example, when assessing a student's emotional and behavioral adjustment, two teachers observing the same student in the identical classroom environment should report similar types and level of behavior on a rating scale. If they do not, results of the scale do not generalize to other raters and could be the result of instrument or rater error rather than differences between students. If they do agree, the scores can be generalized to other raters and, in a theoretical sense, represent the scores of all raters for that student.

There are two major dimensions of interrater agreement, direction and level. Directional agreement is typically quantified via a Pearson product-moment correlation coefficient and is often called interrater reliability. This statistic provides a measure of consistency of agreement regarding the direction of ratings. In essence, it quantifies the tendency of raters to rank order cases similarly. Tests of mean differences are often used to assess *level* of interrater agreement. These two dimensions of interrater agreement can produce three situations that are indicative of poor agreement. First, two raters could consistently agree upon the direction of their ratings but disagree on the level of their ratings. A second indication of poor agreement is shown when raters agree on level but not direction. The third indication of poor agreement is seen in raters who disagree on both direction and level. Raters can demonstrate acceptable agreement only by producing relatively equivalent ratings across both direction and level dimensions (McDermott, 1998; Reid & Maag, 1994).

Information on interrater agreement is infrequently reported for behavior rating scales (Barkley, 1990). Further, even when reported, only directional agreement may be provided. For example, teachers and teachers' aides in a psychiatric hospital were respondents for an interrater reliability study reported in the manual of the Devereux Behavior Rating Scale-School Form (DBRS–SF; Naglieri, LeBuffe, & Pfeiffer, 1993). The average interrater reliability coefficient for the four DBRS–SF subtests was .51; however, the mean scores were not reported so that the level of agreement is unknown. Research on the interrater reliability (directional agreement) of behavior rating scales has revealed considerable variability. Achenbach, McConaughy, and Howell (1987) reviewed the extant research on rating scales and calculated a mean teacher-teacher interrater reliability coefficient of .64. Their review did not analyze level of agreement between raters so that this type of interrater agreement for many behavior ratings scales remains uncertain. Given this situation, Kratochwill and McGivern (1996) noted that there may be limited convergence between raters and questioned empirical scale approaches to assessment of psychopathology.

The Adjustment Scales for Children and Adolescents (ASCA; McDermott, Marston, & Stott, 1993) is a relatively new behavior rating scale designed to assess psychopathology in school settings. Evidence of interrater agreement reported in the ASCA manual (McDermott, 1994) was based upon a small sample of students with emotional disabilities who were rated by self-contained special education classroom teachers and their aides. Given the potential diagnostic applications of the ASCA, a broader assessment of its interrater agreement is needed. Consequently, the purpose of this study was to examine the interrater agreement of the ASCA for a more diverse sample of students enrolled in disparate special programs when they were rated by a variety of educational professionals and paraprofessionals.

Method

Participants

Students were recruited from school districts in two states. Both districts were located in suburban areas of major cities, one in the Southwest and the other in the Midwest. Classrooms where two adults simultaneously worked with students were surveyed and 71 students were identified whose classroom behaviors were jointly observed for at least one hour each day by professionals or paraprofessionals who were willing to participate in this study. These criteria were applied to locate 29 raters in six schools whose job classifications included special education teacher, special education aide, remedial reading teacher, science teacher, and classroom teacher. The most frequent rating pair was a special education teacher and a special education aide in a self-contained, special education setting (58%).

Other observer pairs were classroom teacher/special education teacher (38%) and classroom teacher/remedial reading teacher (48%). In total, there were 29 raters comprising 71 pairs within 24 classrooms in six schools.

Students' racial/ethnic background, as reported by parents on school enrollment forms, was 80% white, 10% Hispanic, 7% Black, and 3% other. Gender distribution was 66% male and 34% female. Students ranged from 7 through 17 years of age with a median age of 11 years and a mean age of 11.1 years. They were enrolled in grades 1 through 10. Participating students received a variety of special services: 44% in Learning Disability; 29% in Emotional Disability; 19% in Severe Language Impairment; and 8% in Mild Mental Retardation programs.

Materials

The Adjustment Scales for Children and Adolescents (ASCA) is a standardized behavioral assessment instrument that was normed on a representative national sample of 1,400 youth, blocked according to gender, age, and grade level. It was also stratified proportionately according to national region, community size, race/ethnicity, parent education, family structure, and handicapping condition. ASCA contains 96 scorable items that are assigned to one of six core syndromes that are universal across race/ethnicity, gender, and age. The six ASCA core syndromes are: (a) Attention-Deficit Hyperactive - inattentive, attention seeking, or restless behavior: (b) Solitary Aggressive (Provocative) - intimidating and overly confrontative behavior; (c) Solitary Aggressive (Impulsive) - impulse-ridden or habit-driven behaviors; (d) Oppositional Defiant - irascible, defiant, and manipulative behaviors; (e) Diffident - timid and fearful behavior; and (f) Avoidant - unusually withdrawn, aloof, and uncommunicative behavior. The core syndromes are combined to form two composite indices: the Attention-Deficit Hyperactive, Solitary Aggressive (Provocative), Solitary Aggressive (Impulsive), and Oppositional Defiant syndromes create the Overactivity scale whereas the Diffident and Avoidant syndromes combine to form the Underactivity scale. As noted previously, the reliability data published in the manual were from a small restrictive sample (see Table 1 for coefficients). However, extensive reliability and validity evidence for the ASCA has been published elsewhere (McDermott, 1993; 1996; McDermott, et al., 1996). In general, psychometric characteristics of the ASCA meet standards for both group and individual decision making (Salvia & Ysseldyke, 1995).

Procedure

Independent ratings of the 71 students were collected over a four-week period following standard administration procedures. All raters had more than 40 days' familiarity with the students before completing the ASCA and based their ratings upon their cumulative, independent observations across time. The student's primary teacher was clerically assigned as Rater A and the secondary teacher was designated Rater B. Thus, self-contained special education teachers and regular classroom teachers were placed in the Rater A group whereas special education aides, resource teachers, and reading teachers were assigned to the Rater B group. All ASCA protocols were returned to the authors, who scored them according to standard procedures (McDermott, 1994). Scores were recorded and protocols were returned to participating schools for their use in evaluation and intervention activities.

As recommended by McDermott (1988), a two-step process to assess interval scale agreement was implemented: (a) *direction* of agreement (interrater reliability) was examined by comparing the ASCA standard T and raw scores of Group A raters with those of Group B raters via Pearson product-moment correlation analyses; and (b) *level* of agreement was examined by comparing the mean ASCA standard and raw scores across raters via t tests for correlated groups. Standard T scores were included in all analyses to ensure alignment with field-based practice (Lee, Elliott, & Barbour, 1994).

Results

ASCA interrater reliability coefficients for each ASCA core syndrome and global composite area were substantial in magnitude and are presented in Table 1. The mean syndrome interrater coefficient was .72[1] for T scores and .78 for raw scores. The median syndrome interrater coefficient was .72 for T scores and .77 for raw scores. The mean reliability coefficient for ASCA global composite T scores was .84 and for raw scores was .88. Means and standard deviations for each ASCA core syndrome and global composite area are provided in Table 2.

An alpha level of .05 was selected for mean difference comparisons. Utilizing this criteria for differences between T scores, the mean rating of Group A was significantly higher than Group B on both the Diffident syndrome (t (70) = 2.59, p = .012) and the Underactivity composite (t (70) = 3.62, p = .001). In difference score terms, the

[1] Average correlation coefficients were obtained using Fisher's Z transformation (Guilford & Fruchter, 1978).

mean of Group A exceeded the mean of Group B by less than 2.5 standard T score points and by less than .5 raw score points on the Diffident and Underactivity scales. These score differences equated to average effect sizes for the Diffident and Underactivity scales of .23 and .15, respectively, for T scores and raw scores (Glass, McGraw, & Smith, 1981).

Table 1

Interrater Reliability Coefficients for ASCA Core Syndrome and Global Composite Raw Scores and Standard T Scores

Scale	Raw Scores*	T Scores*	ASCA Manual+
Attention-Deficit Hyperactive	.76	.72	.67
Solitary Aggressive (Provocative)	.83	.80	.85
Solitary Aggressive (Impulsive)	.68	.55	.72
Oppositional Defiant	.77	.72	.69
Diffident	.87	.75	.81
Avoidant	.71	.71	.65
Overactivity	.87	.83	.81
Underactivity	.89	.85	.84

* n = 71, all p < .001
+ n = 22 from McDermott (1994)

Table 2

Standard T Score Means, Standard Deviations, and Differences for ASCA Core Syndromes and Global Composites

	Group A		Group B		Difference*	
Scale	M	SD	M	SD	t	p
Attention-Deficit Hyperactive	55.3	9.9	55.9	10.1	.74	.46
Solitary Aggressive (Provocative)	57.2	12.4	57.6	12.0	.37	.71
Solitary Aggressive (Impulsive)	53.2	10.8	53.7	10.9	.41	.68
Oppositional Defiant	60.3	13.2	59.6	14.5	.54	.59
Diffident	53.5	10.6	51.2	10.8	2.59	.012
Avoidant	51.8	9.8	51.1	10.0	.82	.42
Overactivity	58.3	8.9	58.1	9.3	.29	.77
Underactivity	53.5	10.2	51.0	11.0	3.62	.001

* t-tests for correlated groups

Discussion

The interrater agreement of educational professionals and paraprofessionals who completed the ASCA with a diverse sample of exceptional students was investigated. Results were robust, with interrater reliability coefficients for core syndromes ranging from .55 to .80 (Median r .72) when T scores were compared and from .68 to .87 (Median r = .77) when raw scores were compared. As expected, global composite scale interrater reliability was higher for Overactivity and Underactivity, r's = .83 and .85, respectively, for T scores and .87 and .89, respectively, for raw scores. Level of agreement among raters was also good, with only two scales differing at a statistically significant level. However, because this represented differences of less than .5 raw score points and small effect sizes (Cohen, 1987), they were not considered clinically meaningful.

The implications for users of the ASCA are clear. If ASCA ratings are used in diagnostic or intervention decisions with exceptional students, ratings generated by one competent adult observer will be relatively similar in terms of direction and level to those generated by a second rater. This extension to professional and paraprofessional raters is particularly useful for continuous or repeated standardized ratings in special classrooms. Of course, the limited number of raters, students, and classrooms used in this study suggest that these conclusions should be tentative and generalization to other raters and settings should be cautious. Nevertheless, these results are promising and suggest that adequate interrater agreement can be obtained when a behavior scale is psychometrically sound and observers interact with a student within a common environment.

References

Achenbach, T.M., McConaughy, S.H., & Howell, C.T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations of situational specificity. *Psychological Bulletin, 101,* 231–232.

Barkley, R.A. (1990). *Attention deficit hyperactivity disorder: A handbook for diagnosis and treatment.* New York: Guilford.

Cohen, J. (1987). *Statistical power analysis for the behavioral sciences* (Revised Edition). Hillsdale, NJ: Erlbaum.

Edelbrock, C. (1983). Problems and issues in using rating scales to assess child personality and psychopathology. *School Psychology Review, 12,* 293–299.

Glass, G.V., McGraw, B., & Smith, M.L. (1991). *Meta analysis in social research.* Beverly Hills, CA: Sage.

Guilford, J.P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education* (6th ed.). New York: McGraw-Hill.

Kamphaus, R.W., & Frick, P.J. (1996). *Clinical assessment of child and adolescent personality and behavior.* Boston: Allyn and Bacon.

Knoff, H.M. (1995). Best practices in personality assessment. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology–III* (pp. 849–864). Washington, DC: National Association of School Psychologists.

Kratochwill, T.R., & McGivern, J.E. (1996). Clinical diagnosis, behavioral assessment, and functional analysis: Examining the connection between assessment and intervention. *School Psychology Review, 25,* 342–355.

Lee, S.W., Elliott, J., & Barbour, J.D. (1994). A comparison of cross-informant behavior ratings in school-based diagnosis. *Behavioral Disorders, 19,* 87–97.

Martin, R.P., Hooper, S., & Snow, J. (1986). Behavior rating scale approaches to personality assessment in children and adolescents. In H.M. Knoff (Ed.), *The assessment of child and adolescent personality* (pp.309–351). New York: Guilford.

McConaughy, S.H., & Ritter, D.R. (1995). Best practices in multidimensional assessment of emotional or behavioral disorders, *Best practices in school psychology–III* (pp. 865–877). Washington, DC: National Association of School Psychologists.

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

McDermott, P.A. (1993). National standardization of uniform multisituational measures of child and adolescent behavior pathology. *Psychological Assessment, 5,* 413–424.

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. (1996). A nationwide study of developmental and gender prevalence for psychopathology in childhood and adolescence. *Journal of Abnormal Child Psychology, 24,* 53–66.

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

McDermott, P.A., Watkins, M.W., Sichel, A.F., Weber, E. M., Keenan, J.T., Holland, A.M., & Leigh, N.M. (1996). The accuracy of new national scales for detecting emotional disturbance in children and adolescents. *Journal of Special Education, 29,* 337–354.

Merrell, K.W. (1994). *Assessment of behavioral, social, & emotional problems.* NY: Longman.

Naglieri, J., LeBuffe, P.A., & Pfeiffer, S.I. (1993). *Devereux Behavior Rating Scale–School Form.* San Antonio, TX: The Psychological Corporation.

Power, T.J., & Ikeda, M.J. (1996). The clinical utility of behavior rating scales: Comments on the diagnostic assessment of ADHD. *Journal of School Psychology, 34,* 379–385.

Reed, R., & Maag, J.W. (1994). How many fidgets in a pretty much: A critique of behavior rating scales for identifying students with ADHD. *Journal of School Psychology, 32,* 339–354.

Reschly, D.J., & Ysseldyke, J.E. (1995). School psychology paradigm shift. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology–III* (pp. 17-32). Washington, DC: National Association of School Psychologists.

Salvia, J., & Ysseldyke, J.E. (1995). *Assessment* (6th ed.). Boston: Houghton Mifflin.

Stinnett, T.A., Havey, J.M., & Oehler-Stinett, J. (1994). Current test usage by practicing school psychologists: A national survey. *Journal of Psychoeducational Assessment, 12,* 331–350.