Psychology in the Schools, Vol. 39(4), 2002
Published online in Wiley InterScience (www.interscience.wiley.com).

DOI: 10.1002/pits.10043

# INTERRATER AGREEMENT FOR DISCRIMINANT CLASSIFICATIONS FOR THE ADJUSTMENT SCALES FOR CHILDREN AND ADOLESCENTS

GARY L. CANIVEZ

*Eastern Illinois University*

MARLEY W. WATKINS AND BARBARA A. SCHAEFER

*The Pennsylvania State University*

Investigation of interrater agreement for the Adjustment Scales for Children and Adolescents (ASCA) discriminant classifications is reported. Two teaching professionals or paraprofessionals working in the same classroom for a minimum of 1 hour per day provided independent ratings of the same child using the ASCA. A total of 119 students ranging in age from 7 to 18 years were independently rated on the ASCA. Results indicated significant and moderate to substantial interrater agreement for the discriminant classifications. © 2002 Wiley Periodicals, Inc.

Current psychological practice indicates a growing preference for objective assessment techniques that help facilitate a link between assessment and intervention (Reschly & Ysseldyke, 1995; Piacentini, 1993). Among applied psychologists, standardized behavior rating scales and checklists have achieved popularity (Hart & Lahey, 1999; Merrell, 1994a) and are the most frequently used instruments in assessing youths' emotional and behavioral difficulties in school (Stinnett, Havey, & Oehler-Stinnet, 1994). Behavior rating scales are considered to be "one of the most efficient, sound, and effective ways . . . to identify a referred student's behavioral strengths and weaknesses . . ." (Knoff, 1995, p. 857). McConaughy and Ritter (1995) suggested that the use of behavior rating scales from multiple sources (e.g., teacher, parent, youth) in concert with interviews, observations, and other assessment tools is "best practice" when assessing and diagnosing emotional and behavioral disorders in children.

Behavior rating scales offer relatively unobtrusive evaluations of students' behaviors in the natural social settings of school, home, and community. While parents inform evaluators of youths' behavior at home, teachers are natural observers and useful informants in the school environment because they have the comparative experience of observing many students across time and varied social contexts. They also appear to take a normative perspective in rating children's behaviors (Piacentini, 1993). Consequently, teachers are considered to be among the most accurate adult raters of child behaviors (Kamphaus & Frick, 1996), and have demonstrated an absence of expectation and practice effects (Brandon, Kehle, Jenson, & Clark, 1990).

Like all tests, behavior rating instruments must demonstrate acceptable psychometric properties before they can be validly applied in practice. Edelbrock (1983) reported that existing behavior rating scales differed across a number of psychometric dimensions. One of the most critical psychometric properties of any instrument that relies upon third-party raters is the degree to which two or more informants agree on the presence or absence of behaviors (Suen & Ary, 1989). Commonly referred to as interrater or interobserver agreement, this measures the degree to which conclusions drawn from an instrument vary as a function of the rater rather than the student being rated.

---

Correspondence to: Gary L. Canivez, Department of Psychology, 600 Lincoln Avenue, Charleston, Illinois 61920–3099. E-mail: glcanivez@eiu.edu

Subjectivity of raters is the primary source of error in rating scale data (Martin, Hooper, & Snow, 1986). 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 levels of behavior on a rating scale. If they do not, results of the scale would not generalize to other raters. Differential results could be the result of instrument or rater error rather than behavior of the student. If raters do agree, then scores can be generalized to other raters and, in a theoretical sense, represent the scores of all raters for that student.

McDermott (1988) pointed out that there are two major dimensions of consistency in interrater agreement for interval scale data: pattern and level. Pattern agreement is often called interrater reliability and is typically quantified via a Pearson product–moment correlation coefficient. This statistic provides a measure of the consistency of agreement regarding the pattern (or direction) of ratings. Tests of mean differences between two or more raters are used to assess the level of interrater agreement. Information on pattern and level are both necessary to determine interrater agreement (McDermott, 1988).

Information on interrater agreement is infrequently reported for behavior rating scales (Barkley, 1990). When data are reported, only pattern agreement is typically provided. For example, as reported in the manual of the Devereux Behavior Rating Scale-School Form (DBRS-SF; Naglieri, LeBuffe, & Pfeiffer, 1993), teachers and teachers' aides in a psychiatric hospital were informants and the average interrater reliability coefficient for the four DBRS-SF subtests was .51. The manual for the Revised Behavior Problem Checklist (Quay & Peterson, 1996) reports that the average correlation between pairs of teacher ratings of developmentally disabled youths was .52 to .85. Similarly, the Child Behavior Checklist-Teacher Report Form (CBCL-TRF; Achenbach, 1991) reports an average correlation coefficient for two independent teacher raters of .54. However, mean score differences for these instruments were not reported so level of agreement is unknown. Although the pattern of agreement appeared to be significant, it is possible that substantial differences existed in the level or magnitude of behavior problems reported by the raters.

Achenbach, McConaughy, and Howell (1987) reviewed the research on interrater agreement of behavior rating scales and found considerable variability. They reported a mean teacher–teacher interrater reliability coefficient of .64. However, their review did not examine level of agreement between raters so interrater agreement for many behavior ratings scales remains uncertain. As Kratochwill and McGivern (1996) noted, empirical scale approaches to assess psychopathology may be questionable due to limited convergence between raters.

The Adjustment Scales for Children and Adolescents (ASCA; McDermott, Marston, & Stott, 1993) is a behavior rating scale designed to assess youth psychopathology in school settings through teacher ratings. Evidence of significant interrater agreement concerning syndrome *T* scores was reported in the ASCA manual (McDermott, 1994) based upon a small sample ($N = 22$) of students with emotional disabilities who were rated by their special education classroom teachers and their teacher aides. Watkins and Canivez (1997) also investigated the interrater agreement of the ASCA for 71 students enrolled in a variety of special programs that were rated by 29 observers within 24 classrooms. Pattern agreement was strong (median syndrome *T* score $r = .72$, median global adjustment scale *T* score $r = .84$) and level differences, although statistically significant, were not clinically meaningful due to small effect sizes and group means differing by less than .5 raw score points.

Two multivariate options are available to facilitate interpretation of resultant ASCA *T* scores: Syndromic Profile Interpretation and Discriminant Classification Interpretation (McDermott, 1994). Syndromic profile interpretation relies upon the normative typology developed by McDermott and Weiss (1995) using the ASCA norm sample of 1400 youth. This approach determines the normal, at-risk, or maladjusted status of youth based on the extent to which an individual's behav-

ior profile most closely matches 1 of 22 healthy or maladaptive behavior profiles. Canivez and Watkins (in press) found significant and moderately high levels of interrater agreement for the 22 ASCA syndrome profiles as well as reductions of these into 5-, 3- and 2-level broad classifications suggested in the ASCA manual. Canivez, Perry, and Weller (2001) also found significant stability for the 22 ASCA syndrome profiles and the 5-, 3-, and 2-level broad classifications over a 90-day interval.

The second method of multivariate differential diagnosis (classification) prescribed in the ASCA manual is discriminant classification. Discriminant function analyses (DFA) conducted by McDermott et al. (1995) assessed the degree to which the ASCA was able to correctly classify normal from socially/emotionally disturbed (SED) youths. McDermott et al. compared 150 students classified as SED to a group of 150 normal students matched for age, gender, and ethnicity. The SED students were also compared to 1783 normal children, 360 students with learning disabilities (LD), 29 students with speech/language disability, 60 students who were gifted, and 2,530 normal students from the ASCA standardization and validity samples. The DFA for the 300 matched subjects and the redistribution based on the DFA yielded a better quadratic solution than linear solution (by an average of 2%) due to reported heterogeneity of within classification covariance. Diagnostic efficiency statistics (Kessel & Zimmerman, 1993) yielded classification accuracy ranging from 76.9% to 86.2% for all groups compared. More importantly, positive predictive power estimates exceeded a recommended standard (.75) for diagnostic tests proposed by Milich, Widiger, and Landau (1987) for clinical practice.

Discriminant function analyses have been used to provide evidence of criterion-related validity for other instruments, as well. Glutting, Robins, and de Lancey (1997) found that test session behaviors evaluated using the Guide to the Assessment of Test Session Behaviors (GATSB; Glutting & Oakland, 1993) demonstrated sufficiently high hit rates (overall correct classification) when comparing attention deficit-hyperactivity disordered (ADHD) to control group youths. Ellen (1989) conducted similar research with LD, SED, and normal elementary school boys using the Classroom Adjustment Rating Scale (Weissberg, Gesten, & Ginsburg, 1981); however, classification rates beyond chance were not favorable. McConaughy and Achenbach (1996) utilized discriminant function analyses to assess the contribution of a semi-structured clinical interview and parent and teacher rating scale information in discriminating between normal, SED, and LD youths. Similarly, Zelko (1991) assessed the relative contributions of three parent-completed behavior rating scales in identification of ADHD, psychiatric, and normal boys, and determined that multidimensional scales are preferable for clinical diagnostic purposes. Another approach applied by Kline, Lachar, and Boersma (1993) using the parent-completed Personality Inventory for Children (PIC; Wirt, Lachar, Klinedinst, & Seat, 1984) combined discriminant function analyses (DFA) and profiles of LD, SED, or mentally impaired youths to identify PIC scales contributing to group discrimination. A set of six hierarchical classification rules derived from DFA were developed, and cross validation indicated 90% accuracy in regular versus special education classifications, but only about 50% accuracy for specific problems (LD vs. SED). These discriminant function analysis studies have demonstrated mixed success in accurate classification of youth. Furthermore, none of these studies investigated the level of agreement between discriminant function classifications from two independent raters.

Given the potential diagnostic and interpretive applications of discriminant classifications with the ASCA (McDermott, 1994), a broader assessment of interrater agreement is needed. To date, there have been no investigations of interrater agreement for the ASCA discriminant classifications. Thus, the purpose of the present study was to investigate the degree of agreement on discriminant classifications produced by two different raters observing the same child in the same classroom environment.

## Method

### *Participants*

Participants from the Watkins and Canivez (1997) and Canivez and Watkins (in press) studies of interrater agreement for the ASCA ($n$ = 71) served as participants in this investigation. To increase the overall sample size, additional participants ($n$ = 48) were subsequently obtained using the same method. In total, 45 teachers from three school districts in two states were recruited to complete ASCA rating forms on their students. Two districts were located in suburban areas of major cities: one in the southwest and one in the midwest. The third district was located in a rural area in the midwest. A total of 119 students were identified whose classroom behaviors were jointly observed for a minimum of 1 hour each day by two professionals or one professional and one paraprofessional who were willing to participate in this study. Job classifications of raters included special education teacher, special education teaching assistant, remedial reading teacher, science teacher, and regular classroom teacher. The most frequent rating pair was a special education teacher and a special education teaching assistant in a self-contained, special education setting (75%). Other observer pairs included regular classroom teacher—special education teacher (18%) and regular classroom teacher—remedial reading teacher (3%). In total, there were 45 raters comprising 119 pairs within 32 classrooms in 7 different schools.

Students' racial/ethnic backgrounds, as reported by parents on school enrollment forms, included 91.6% Caucasian, 6.6% Hispanic/Latino, 0.9% Black/African American, and 0.9% Middle Eastern. The student sample included 73% males and 27% females, ranging in age from 7 through 18 years with a mean age of 11.2 years ($SD$ = 3.31). Students were enrolled in grades 1 through 12 and were involved in a variety of special programs for those at risk or disabled: 24% in Learning Disability; 56% in Emotional Disability; 12% in Severe Language Impairment; and 5% in Mild Mental Retardation. The remaining 3% were not classified as disabled.

### *Instrument*

The ASCA is an objective behavior rating instrument completed by a student's classroom teacher and designed for use with all noninstitutionalized youths, ages 5 through 17. The ASCA contains 156 behavioral descriptions within 29 specific situations where teachers may observe students' behaviors, rather than a symptom or problem checklist. Of the 156 items, 97 are scored as items assessing psychopathology and, based on factor analyses, singularly assigned to one of six core syndromes (Attention-Deficit/Hyperactive, Solitary Aggressive–Provocative, Solitary Aggressive–Impulsive, Oppositional Defiant, Diffident, and Avoidant) or two supplementary syndromes (Delinquent and Lethargic/Hypoactive). The core syndromes are combined to form two composite indexes (broadband scales): Overactivity (Attention Deficit–Hyperactive, Solitary Aggressive–Provocative, Solitary Aggressive–Impulsive, and Oppositional Defiant syndromes) and Underactivity (Diffident and Avoidant syndromes). Raw scores are converted to normalized $T$ scores based on the nationally representative standardization sample. The ASCA was normed on a representative national sample of 1400 youths, blocked according to gender, age, and grade level and stratified proportionately according to national region, community size, race/ethnicity, parent education, family structure, and handicapping condition.

Extensive reliability and validity evidence is provided in the ASCA Manual (McDermott, 1994). Internal consistency estimates for the total standardization sample ranged from .68 to .86 for the six core syndromes and two supplementary syndromes. Alpha coefficients equaled .92 for the Overactivity scale and .82 for the Underactivity scale. Test–retest reliabilities over a 30-school-day interval ranged from .66 to .91 for the six core syndromes ($N$ = 40), and equaled .75 for the Overactivity scale and .79 for the Underactivity scale. Canivez et al. (2001) replicated the stability

of the ASCA *T* scores over a 90-day interval as well as supporting stability of the ASCA syndrome profiles and discriminant classifications. Watkins and Canivez (1997) replicated the interrater agreement for ASCA *T* scores while Canivez and Watkins (2001) found significant interrater agreement for ASCA syndrome profiles.

Exploratory and confirmatory analyses support the factor structure of the ASCA. Convergent and divergent validity studies comparing the ASCA with the Conners Teacher Rating Scale (CTRS; Trites, Blouin, & Laprade, 1982) and the Child Behavior Checklist (CBCL; Achenbach & Edelbrock, 1983) found significant correlations among similar psychological dimensions (McDermott, 1994). Canivez and Bordenkircher (in press) and Canivez and Rains (in press) found significant convergent and divergent validity for the ASCA Overactivity and Underactivity syndromes when compared to the Preschool and Kindergarten Behavior Scales (Merrell, 1994b). In general, psychometric characteristics of the ASCA are acceptable (Canivez, 2001; Schowengerdt, 2001) and meet standards for both group and individual decision-making (Salvia & Ysseldyke, 1995).

### *Procedure*

Independent ratings of the 119 students were collected following the ASCA standard administration procedures. The student's primary teacher (special education teacher or regular classroom teacher) was designated as Rater 1 while the secondary rater (special education teaching assistant, resource teacher, and remedial reading teacher) was designated Rater 2.

### *Data Analysis*

Unlike the ASCA syndrome and global adjustment scale *T* scores, discriminant classification produces a nominal scale variable (Normal vs. SED). When investigating agreement on nominal scale or categorical variables, nominal scale statistics such as kappa ($\kappa$) (Cohen, 1960; Fleiss, 1981; McDermott, 1988) should be utilized. Kappa provides an index of agreement beyond chance agreement and is interpreted much like a correlation coefficient as it ranges from $-1$ to $+1$.

Discriminant classifications for the ASCA were made according to the ASCA Manual using linear discriminant classification equations (McDermott, 1994, p. 29) provided in an automated scoring template (Canivez, 1996; 1998a). Profiles were classified "Normal" or "SED" based on the regression equation resulting in the highest discriminant score (McDermott, 1994). Interrater agreement was assessed using an automated template (Canivez, 1999) that compared the discriminant classifications resulting from the two independent raters.

## Results

Interrater agreement of the discriminant classifications is presented in Figure 1. Discriminant classifications showed significant interrater agreement ($\kappa = .51$, $z = 5.70$, $p < .00001$) at a level considered moderate to substantial (Everitt & Hay, 1992; Landis & Koch, 1977). Of the 119 students, 34 (29%) were classified "Normal" by both raters while 57 (48%) were classified "SED" by both raters. Nineteen students (16%) were classified "SED" by Rater 1 and "Normal" by Rater 2; while nine (8%) students were classified as "Normal" by Rater 1 but "SED" by Rater 2.

## Discussion

This is the first study to investigate the interrater agreement of the discriminant classification interpretation method presented in the ASCA manual (McDermott, 1994). As previous research has demonstrated that accurate identification of emotional disturbance among youth is attainable using the ASCA (McDermott et al., 1995), this study provided additional support for the consis-

| | | Rater 2 | | |
|---|---|---|---|---|
| | | Normal | SED | Total |
| Rater 1 | Normal | **34** | 9 | 43 |
| | SED | 19 | **57** | 76 |
| | Total | 53 | 66 | 119 |

| Results | |
|---|---|
| **Observed Agreement Po =** | **0.7647** |
| **Chance Agreement Pc =** | 0.5151 |
| **Kappa (κ) =** | **0.5147** |
| **SE(κ) =** | 0.0903 |
| **Significance Tests for Kappa:** | |
| $z$ | **= 5.6999** |
| **two-tail test Ho: κ = 0** $p$ | **< 1.2022E-08** |
| **one-tail test Ho: κ = 0** $p$ | **< 6.011E-09** |

FIGURE 1. ASCA Discriminant classification agreement template and results.

tency of such classifications based on ratings from two observers in the same classroom setting. Results indicated that the discriminant classifications demonstrated significant interrater agreement beyond chance based on kappa coefficients. These data also suggested that interrater agreement is moderate to substantial (Everitt & Hay, 1992; Landis & Koch, 1977) and provided an example of the utility of an actuarial decision-making approach (Davidow & Levinson, 1993; Dawes, Faust, & Meehl, 1989).

This is an encouraging and important finding, as one would expect that classifications based on an objective behavioral or psychopathology measure should be similar for two raters observing the same child in the same classroom at the same time. These results are congruent with prior work that found support for the interrater reliability of ASCA syndrome ratings on youths observed in the same setting, but lower levels of agreement were found for youths observed by education personnel in different settings (Schaefer, Watkins, & Canivez, 2001).

As there are no previous studies investigating interrater agreement for discriminant classifications with behavior rating scales, it is difficult to place the present results in a broader perspective. Danforth and DuPaul (1996) found significant interrater agreement for several teacher rating scales designed to assess attention-deficit hyperactivity disorder (ADHD). Using kappa to assess interrater agreement for different clinical cutoffs or cut scores, coefficients ranging from .26 to 1.0 (*Mdn* = .51) were obtained, with coefficients generally highest at the cut-off two standard deviations above the mean. Molina, Pelham, Blumenthal, and Galiszewski (1998) also assessed teacher agreement on students with ADHD behaviors using three rating forms and reported κ coefficients for behavior scales ranging from nonsignificant to significant (.17 to .48, respectively; *Mdn* = .34). Reid and Maag (1994) pointed out the inconsistency for ratings of ADHD and further argued that reliance on cut scores for scales comprised of items using Likert-type frequency descriptors is problematic. The ASCA ratings of situational items are dichotomous (present or absent) and avoid

this frequency rating confound. Further, the current study is not reliant upon a cut-off score of a single scale (e.g., Oppositional Defiant) or individual item endorsement, but rather utilizes a multivariate approach that simultaneously includes all six ASCA core syndromes in the discriminant function regression formulae to classify social/emotional maladjustment.

The present levels of agreement compare quite favorably to those found in test–retest stability (median parent $\kappa = .58$, median child $\kappa = .49$) and interrater agreement (computer vs. clinician $\kappa = .23$) studies of structured interviews for psychiatric diagnoses (Fisher et al., 1997; Hodges & Zeman, 1993). Diagnostic agreement of ASCA discriminant classifications also was consistent with kappas (.54 to .59) reported for the DSM-IV field trials for disruptive behavior disorders (Lahey et al., 1994). The present level of agreement for discriminant classifications was higher than that obtained in a 90-day test–retest stability study of the ASCA ($\kappa = .35$; Canivez et al., 2001).

Although these results are encouraging, caution should be exercised in interpretation as they are based on a relatively small, nonrandom sample of students who were not representative of the population at large. Generalizability may also be impaired as the present study employed a limited number of raters, classrooms, and geographic locations. Future studies should continue to investigate the interrater agreement of the ASCA in a similar manner and incorporate larger and more diverse and representative student and teacher samples. Replication within regular education settings is particularly needed as behavior rating scales are frequently used within these settings for screening as well as for initial evaluations for determining psychopathology and disability classification. Unfortunately, it is difficult to find regular education classrooms where there are two teachers present at the same time.

Another limitation is that discriminant classification for the ASCA relates to a classification of "Normal" or "SED" based on the six ASCA core syndromes where "SED" students were based on students placed in special education programs and classified SED. Such students are most frequently those with externalizing symptoms. Finer distinctions between more homogeneous groups (i.e., attention deficit-hyperactivity disorder, oppositional defiant disorder, conduct disorder, etc.) would be also be helpful.

Best practice also encourages utilization of multiple sources of information in assessment and decision-making regarding emotional and behavioral disorders (McConaughy & Ritter, 1995). However, Reid and Maag (1994) have warned, "rater variance can impact upon the results of multimethod assessment as much as or more than trait variance" (p. 375). While agreement for ASCA's discriminant function is supported for concurrent observations, discriminant classification agreement for observers across settings has not been evaluated. Observers from different classrooms have shown lower levels of ASCA syndrome score agreement which points to likely situational specificity of behavior (Schaefer et al., 2001).

Nonetheless, the present results supported the integrity and consistency of ASCA discriminant classifications. Given that the discriminant classification is categorical (i.e., "SED" or "Normal"), it is important to remember that an underlying continuum of symptoms exists ranging from well adjusted to psychopathological behavior. Other models may be beneficial in assessing similar data, assuming the data is generated by continuous latent variables, to permit estimation of trait, method, and error variance from observation ratings (Fergusson & Horwood, 1989). Further, researchers may apply probability modeling techniques to integrate multiple observers' ratings to obtain an "optimal classification" (Uebersax, 1988, p. 410) for each case. In applied practice, however, this may be prohibitively complicated. Application of discriminant classification can be easily accomplished, however, and this study demonstrated the utility and consistency of an empirically based interpretive approach in classifying youths with SED that can be integrated as part of a complete psychological evaluation.

## References


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

Achenbach, T.M., & Edelbrock, C. (1983). Manual for the Child Behavior Checklist and Revised Child Behavior Profile. Burlington, VT: University of Vermont.

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, 213–232.

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

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.

Canivez, G.L. (1996). Automated Syndromic Profile and Discriminant Classification Analyses for the Adjustment Scales for Children and Adolescents (ASCA), v2.0. Microsoft® Excel™ spreadsheet template for the Apple® Macintosh™ microcomputer. (Available from G.L. Canivez, Psychology Department, Eastern Illinois University, 600 Lincoln Street, Charleston, IL 61920–3099)

Canivez, G.L. (1998a). Automated syndromic profile and discriminant classification analyses for the Adjustment Scales for Children and Adolescents. Behavior Research Methods, Instruments, & Computers, 30, 732–734.

Canivez, G.L. (1999). Automated calculation of nominal scale agreement (Kappa) statistics for ASCA syndromic profile classifications and discriminant classifications. Microsoft® Excel™ spreadsheet templates for the Apple® Macintosh™ microcomputer. (Available from G.L. Canivez, Psychology Department, Eastern Illinois University, 600 Lincoln Street, Charleston, IL 61920–3099)

Canivez, G.L. (2001). Review of the Adjustment Scales for Children and Adolescents. In J. Impara & B. Plake (Eds.), The fourteenth mental measurements yearbook (pp. 22–24). Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska.

Canivez, G.L., & Bordenkircher, S.E. (in press). Convergent and divergent validity of the Adjustment Scales for Children and Adolescents for Children and the Preschool and Kindergarten Behavior Scales. Journal of Psychoeducational Assessment.

Canivez, G.L., Perry, A.R. & Weller, E.M. (in press). Stability of the Adjustment Scales for Children and Adolescents. Psychology in the Schools, 38, 217–227.

Canivez, G.L., & Rains, J.D. (in press). Construct validity of the Adjustment Scales for Children and Adolescents and the Preschool and Kindergarten Behavior Scales: Convergent and divergent evidence. Psychology in the Schools.

Canivez, G.L., & Watkins, M.W. (in press). Interrater agreement for Syndromic Profile Classifications on the Adjustment Scales for Children and Adolescents. Assessment for Effective Intervention.

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

Danforth, J.S., & DuPaul, G.J. (1996). Interrater reliability of teacher rating scales for children with attention-deficit hyperactivity disorder. Journal of Psychopathology and Behavioral Assessment, 18, 227–237.

Davidow, J., & Levinson, E.M. (1993). Heuristic principles and cognitive bias in decision making: Implications for assessment in school psychology. Psychology in the Schools, 30, 351–361.

Dawes, R.M., Faust, D., & Meehl, P.E. (1989). Clinical versus actuarial judgment. Science, 243, 1668–1674.

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

Ellen, A.S. (1989). Discriminant validity of teacher ratings for normal, learning-disabled, and emotionally handicapped boys. Journal of School Psychology, 27(1), 15–25.

Everitt, B.S., & Hay, D.F. (1992). Talking about statistics: A psychologist's guide to data analysis. New York: Wiley.

Fergusson, D.M., & Horwood, L.J. (1989). Estimation of method and trait variance in ratings of conduct disorder. Journal of Child Psychology and Psychiatry & Allied Disciplines, 30, 365–378.

Fisher, P., Lucas, C., Shaffer, D., Schwab-Stone, M., Graae, F., Lichtman, J. et al. (1997). Diagnostic Interview Schedule for Children, Version IV (DISC-IV): Test–retest reliability in a clinical sample. Poster presented at the 44th annual meeting of the American Academy of Child and Adolescent Psychiatry, Toronto.

Fleiss, J.L. (1981). Statistical methods for rates and proportions (2nd ed.). New York: Wiley.

Glutting, J.J., & Oakland, T. (1993). Guide to the Assessment of Test Session Behavior. San Antonio, TX: Psychological Corporation.

Glutting, J.J., Robins, P.M., & de Lancey, E. (1997). Discriminant validity of test observations for children with attention deficit/hyperactivity. Journal of School Psychology, 35(4), 391–401.

Hart, EL., & Lahey, B.B. (1999). General child behavior rating scales. In D. Shaffer, C.P. Lucas, & J.E. Richters (Eds.), Diagnostic assessment in child and adolescent psychopathology (pp. 65–87). New York: Guilford.

Hodges, K., & Zeman, J. (1993). Interviewing. In T.H. Ollendick & M. Hersen (Eds.), Handbook of child and adolescent assessment (pp. 65–81). Boston: Allyn & Bacon.

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

Kessel, J.B., & Zimmerman, M. (1993). Reporting errors in studies of the diagnostic performance of self-administered questionnaires: Extent of the problem, recommendations for standardized presentation of results, and implications for the peer review process. Psychological Assessment, 5, 395–399.

Kline, R.B., Lachar, D., & Boersma, D.C. (1993). Identification of special education needs with the Personality Inventory for Children (PIC): A hierarchical classification model. Psychological Assessment, 5(3), 307–316.

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.

Lahey, B.B., Applegate, B., Barkley, R.A., Garfinkel, B., McBurnett, K., Kerdyk, L. et al. (1994). DSM-IV field trials for oppositional defiant disorder and conduct disorder in children and adolescents. American Journal of Psychiatry, 151, 1163–1171.

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

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., & Achenbach, T.M. (1996). Contributions of a child interview to multimethod assessment of children with EBD and LD. School Psychology Review, 25(1), 24–39.

McConaughy, S.H., & Ritter, D.R. (1995). Best practices in multidimensional assessment of emotional or behavioral disorders. In A. Thomas, & J. Grimes (Eds.), Best practices in school psychology (3rd ed., 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. (1994). National profiles in youth psychopathology: Manual of Adjustment Scales for Children and Adolescents. Philadelphia: Edumetric & Clinical Science.

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

McDermott, P.A., Watkins, M.W., Sichel, A.F., Weber, E.M., Keenan, J.T., Holland, A.M., & Leigh, N.M. (1995). The accuracy of new national scales for detecting emotional disturbance in children and adolescents. The Journal of Special Education, 29, 337–354.

McDermott, P.A., & Weiss, R.V. (1995). A normative typology of healthy, subclinical, and clinical behavior styles among American children and adolescents. Psychological Assessment, 7, 162–170.

Merrell, K.W. (1994a). Assessment of behavioral, social, & emotional problems. New York: Longman.

Merrell, K.W. (1994b). Preschool and Kindergarten Behavior Scales. Brandon, VT: Clinical Psychology Publishing Company.

Milich, R., Widiger, T.A., & Landau, S. (1987). Differential diagnosis of attention deficit and conduct disorders using conditional probabilities. Journal of Consulting & Clinical Psychology, 55(5), 762–767.

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.

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

Piacentini, J. (1993). Checklists and rating scales. In T.H. Ollendick & M. Hersen (Eds.), Handbook of child and adolescent assessment (pp. 82–97). Boston: Allyn & Bacon.

Quay, H.C., & Peterson, D.R. (1996). Revised Behavior Problem Checklist PAR edition professional manual. Odessa, FL: Psychological Assessment Resources.

Reid, 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. Boston: Houghton Mifflin.

Schaefer, B.A., Watkins, M.W., & Canivez, G.L. (2001). Cross-context agreement of the Adjustment Scales for Children and Adolescents. Journal of Psychoeducational Assessment, 19, 123–136.

Schowengerdt, R.V. (2001). Review of the Adjustment Scales for Children and Adolescents. In J. Impara & B. Plake (Eds.), The fourteenth mental measurements yearbook (pp. 24–26). Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska.

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

Suen, H.K., & Ary, D. (1989). Analyzing quantitative behavioral observation data. Hillsdale, NJ: Erlbaum.

Trites, R.L., Blouin, A.G.A., & Laprade, K. (1982). Analysis of the Conners Teacher Rating Scale based on a large normative sample. Journal of Consulting and Clinical Psychology, 50, 615–623.

Uebersax, J.S. (1988). Validity inferences from interobserver agreement. Psychological Bulletin, 104, 405–416.

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

Weissberg, R.P., Gesten, E.L., & Ginsburg, M.R. (1981). Classroom Adjustment Rating Scale (CARS): Manual. Rochester, NY: Center for Community Study.

Wirt, R.D., Lachar, D., Klinedinst, J.K., & Seat, P.D. (1984). Multidimensional description of child personality: A manual for the Personality Inventory for Children. Los Angeles: Western Psychological Services.

Zelko, F.A. (1991). Comparison of parent-completed behavior rating scales: Differentiating boys with ADD from psychiatric and normal controls. Journal of Developmental & Behavioral Pediatrics, 12(1), 31–37.