THE JOURNAL OF SPECIAL EDUCATION VOL. 29/NO. 3/1995/pp. 337–354

THE ACCURACY OF NEW NATIONAL SCALES FOR DETECTING EMOTIONAL DISTURBANCE IN CHILDREN AND ADOLESCENTS

Paul A. McDermott
University of Pennsylvania
Marley W. Watkins
Pennsylvania State University
Amy F. Sichel
Abington School District
Abington, Pennsylvania
Ellen M. Weber
Joseph T. Keenan
Anne M. Holland
Nanette M. Leigh
University of Pennsylvania

Through a series of multivariate classification studies, we examined the sensitivity, specificity, and overall accuracy of the Adjustment Scales for Children and Adolescents (ASCA; McDermott, Marston, & Stott, 1993) in detecting emotional disturbance. The ASCA, a teacher rating scale normed on a stratified national sample of 5- through 17-year-old youths, was first applied to a sample of 150 nondisabled subjects and 150 subjects who were socially or emotionally disturbed (SED) matched for age, gender, race, and grade level. Discriminant analyses, including single- and split-sample cross-validations, produced sensitivity, specificity, and overall accuracy ratings of approximately 80%, a level significantly above chance expectancy. The discriminant solution maintained similar accuracy when applied independently to preadolescent, adolescent, male, female, White, and African American children. Accuracy also sustained when the scale was used to differentiate the SED subjects from samples of students who were learning disabled, communication impaired, or gifted/talented, and from a mixture of 1,843 nondisabled subjects and 537 others with learning, mental, and physical disabilities. The scale's performance was evaluated in light of classification accuracy research on other popular instruments.

Behavior rating scales and observation schedules have gained widespread acceptance among child specialists. Developed originally as a means to codify lists of symptoms found in clinical settings, the devices are now commonly applied in school settings for the identification and placement of students who are socially or emotionally disturbed (SED) (Costenbader & Keller, 1990; Mattison, Bagnato, Mayes, & Felix, 1990; Tharinger, Laurent, & Best, 1986).

Address: Paul A. McDermott, Graduate School of Education, University of Pennsylvania, 3700 Walnut St., Philadelphia, PA 19104.

Popularity notwithstanding, the legitimate application of rating scales has been limited by a number of factors. Scales that rely on parent informants are particularly problematic because parents themselves play formative roles in observed behavior pathology. Indeed, compelling new evidence suggests that the ratings and observations provided by parents are often as or more indicative of parent pathology as they are of disturbance in the children being rated (Banez & Compas, 1990; Brody & Forehand, 1986; Campbell, March, Pierce, Ewing, & Szumowski, 1991; Ferguson, Lynskey, & Horwood, 1993; Phares, Compas, & Howell, 1989; Richters, 1992; Schaughency & Lahey, 1985; Watson, Henggeler, & Whelan, 1990). Children's self- and peer-reports tend to be unreliable (Edelbrock, Costello, Dulcan, Kalas, & Conover, 1985; Loeber, Green, & Lahey, 1990) and riddled with problems related to semantic distortion, cognitive immaturity, social conformity, deception, and limited literacy (Anastasi, 1988). Alternatively, observations made by classroom teachers afford relatively unobtrusive assessments in natural social situations. Teachers, although imperfect informants, tend to provide more consistency in their evaluations (Brandon, Kehle, Jenson, & Clark, 1990; Campbell & Steinert, 1978) and are able to benefit from the comparative experience with many different children across time and varied social contexts—learning, recreation, confrontation, and so forth (Achenbach, 1988; McDermott, 1986).

Clearly the foremost impediment to proper application of rating scales has been the absence of representative national normative data on the base rates for youth psychopathology (Fantuzzo & McDermott, 1992; Garfield, 1978; Kline, 1988). Until recently, all available normative information has pertained exclusively to local geographic areas and has derived from samples of children that diverge widely, and often in unknown ways, from the general youth population in the United States.

Within this framework, the Adjustment Scales for Children and Adolescents (ASCA; McDermott, Marston, & Stott, 1993) were developed and standardized nationally on a large and representative sample of American school children, age 5 through 17 years. ASCA applies a format that differs from most other empirical observation scales. Most scales present lists of problem behavior or symptoms and infer pathology from the number of related symptoms observed or the observer's general impression of the frequency or intensity of symptoms. Such information does not clarify whether the symptomatic behavior is isolated to specific circumstances (as, for instance, when peers are physically aggressive or when authorities impose expectations for learning) or whether it is pervasive across varied circumstances. Lacking confirmation of the contextual ethos for observed behavior, it is difficult to determine motivation and suggest remedial action. Alternatively, ASCA defines behavior pathology through its *multisituational* occurrence. Respondents indicate which specific behaviors typify given circumstances, and the verity and severity of pathology are drawn from its pervasiveness across contexts. This approach is consistent with that proposed by Horn, Wagner, and Ialongo (1989) and by Breen and Altepeter (1991), favoring the diagnostic import of pervasive behavior, and with the view that behavior emergent only in isolated contexts is far more likely to be random or reactive than it is pathognomonic.

As part of the criterion validation process, ASCA was used to evaluate subsamples of children from across the nation who were independently classified

as SED for special education purposes and subsamples of comparable children who were nondisabled in terms of emotional adjustment. This article describes the assessment procedures and the relative accuracy of ASCA for distinguishing between children who are disturbed and those who are emotionally nondisabled.

METHOD

Subjects

The primary sample contained 300 children age 5 through 17 who were drawn from a larger cohort of 2,530 subjects obtained during national standardization of ASCA. Sample construction began with identification of 150 subjects who had been classified by multidisciplinary child study teams as SED and placed in special education. These children came from 12 sites distributed across the four geographic regions of the country and constituted a representative subset of the 154 public school districts and 47 private school sites involved in the national project. Criteria for identification and placement, although locally implemented, were commensurate with federal guidelines (*United States Code*, 1977) and were applied prior to use of the ASCA scales, with 146 of the children placed in self-contained classroom environments and 4 routinely alternating between general class and resource room attendance.

A complementary group of 150 emotionally nondisabled subjects was drawn from the national cohort by the method of simultaneous diadic matching to the disturbed subjects in terms of age, gender, race, and grade level (see Note 1). Children classified under other special education categories were excluded from membership in the nondisabled group. Results of the matching process are presented in Table 1 and indicate appreciable success in producing adjusted and maladjusted contrast groups otherwise balanced for pertinent demography.

TABLE 1
CHARACTERISTICS OF NONDISABLED CHILDREN AND CHILDREN WITH SED MATCHED FOR AGE, GENDER, RACE/ETHNICITY, AND GRADE LEVEL

Variable		Nondisabled	Social or emotional disturbance	Total
Age in years	*M*	11.13	11.55	11.33
	SD	3.07	2.89	2.98
Gender				
Male	*n*	114	114	228
Female	*n*	36	36	72
Race/Ethnicity				
White	*n*	122	122	244
African American	*n*	25	25	50
Hispanic	*n*	2	2	4
Other	*n*	1	1	2
Grade level	*M*	5.74	5.93	5.78
	SD	2.86	2.80	2.83

Note. $N = 300$.

The larger cohort of 2,530 subjects, as well as constituent subgroups of students classified as learning disabled, communication handicapped, and gifted or talented, served as secondary samples used in supplementary analyses attempting to discriminate the 150 subjects with SED from respective groups of nondisturbed subjects. The latter, emotionally nondisabled portion of this sample included 1,783 nonclassified subjects, 360 learning disabled (LD), 29 speech or communication handicapped (CH), 60 gifted or talented (GT), and an assortment of children sustaining rarer mental, sensory, or orthopedic disabilities or chronic illness. The sample essentially merged the full national standardization sample (N = 1,400) and all ancillary validity generalization and ethnic bias samples, including oversampling of exceptional children. All exceptional children had been identified and placed by interdisciplinary teams in accord with the federal guidelines previously cited. Inasmuch as males show higher morbidity rates for disabling conditions (U.S. Department of Education, 1989), this full-population sample was 53.4% male and 46.6% female.

Instrumentation

The ASCA national standardization sample (N = 1,400) was configured to specifications of the 1988–1990 U.S. Census (U.S. Department of Commerce, 1988, 1990), with matrix blocking for gender (700 males and 700 females) and approximately equal distributions of subjects and gender at each 1-year age (5–17 years) and grade (K–12) interval. Stratified random sampling was used to fit the sample to census proportions, both marginally and within cells, across the four regions of the country and for major and minor metropolitan and rural residences, race/ethnicity (White, African American, Hispanic origin, and other), and years of formal schooling completed by mother and father (or guardian).

Stratification provided approximate matching to national distributions of special education subjects (as per the U.S. Department of Education, 1989) classified by multi-disciplinary child study teams, including 6.9% disabled (with proper subcategorical distributions of LD, SED, speech impaired, etc.) and 4.0% gifted or talented. Intellectual ability and academic achievement were assessed for 1,260 of the 1,400 through the Differential Ability Scales (Elliot, 1990), a battery individually administered in all cases by psychologists or supervised psychology graduate students. In each performance area, including General Conceptual, Verbal, and Nonverbal Ability, and reading and mathematics achievement, the derived quotient means and standard deviations for the 1,260 subjects were 100 and 15, respectively.

ASCA contains 97 problem and 26 positive behavior indicators, each presented in 1 of 29 specific situations involving authority, peers, smaller or weaker children, play, learning, or confrontation. For example, within the social situation where a child is corrected by a teacher, ASCA provides problem behavior indicators such as "Improves for the moment but it does not last long," "Takes correction badly (sulky muttering, expressions, etc.)," and "Answers back aggressively, makes threats or creates a disturbance." Thus, each of the 123 indicators constitutes a binary item scored 1 (present) or 0 (absent).

The behavioral indicators and situations were designed and revised in field trials to provide clear content meaning for respondents and to reduce or eliminate necessity for respondent inferences concerning children's thoughts or feelings. As an additional step to comply with respondents' preferences, two versions of the scale are applied. The versions, although providing identical items and situations, differ only in use of gender referents (*she/her* vs. *he/him*).

The 97 problem indicators are assigned to eight mutually exclusive syndromes based on exploratory, multiple-group, principal components analyses (i.e., a common equamax and promax solution), item analysis, and confirmatory oblique principal components cluster analysis (Anderberg, 1973; Harman, 1976). Independent random replications and analyses restricted to respective preadolescent, adolescent, male, and female portions of the national sample and to separate samples of White and non-White subjects (including additional minority cases obtained from the national sites) supported the invariance of the syndrome structures and generalizability to the various demographic subgroups. Specifically, Wrigley-Neuhaus factor structure constancy coefficients (Harman, 1976) averaged .88 (range = .97 to .76) for the full normative sample, .88 for 740 preadolescents, .91 for 660 adolescents, .91 for 700 males, .85 for 700 females, .89 for 965 Whites, .85 for 568 heterogeneous non-Whites, and .79 for 251 African Americans.

Six of the syndromes produced adequate internal consistency overall and within each demographic subgroup, as well as appreciable interobserver and test–retest reliability. These core syndromes include (a) Attention-Deficit Hyperactive, with 19 items indicating inattentive, attention-seeking, or restless behavior; (b) Solitary Aggressive (Provocative), with 13 items depicting intimidating and overtly confrontative behavior; (c) Solitary Aggressive (Impulsive), with 9 items describing impulse- or habit-driven offense; (d) Oppositional Defiant, with 12 items describing irascible, often covert, defiance and manipulation; (e) Diffident, with 13 items distinguishing timid and fearful behavior; and (f) Avoidant, with 10 items referring to unusually withdrawn, aloof, and uncommunicative behavior. Syndrome alpha coefficients averaged .77 (range = .86 to .70) for the norm sample, .78 for preadolescents, .75 for adolescents, .77 for males, .76 for females, .77 for Whites, .77 for all non-Whites, and .80 for African Americans.

Unit-weighted syndrome scores (the sum of component behaviors endorsed) were transformed by area conversion (Thorndike, 1982) into *T* scores based on the national sample. Thus, although distributions maintained the positive skew associated with pathology items, each standard score was associated with the appropriate proportion of cases under the normal curve (see Note 2). Syndrome scores were submitted to second-order principal factoring and variance specificity analysis, revealing 40% common variance among the syndromes (communalities ranging from .49 to .22) as explained by two factors resembling the classic over- versus underreactive maladjustment dichotomy (Quay, 1986). Nevertheless, each syndrome also was found to retain a substantial amount of unique and reliable variance (average = 43.7%; range = 58% to 29%), as well as substantial convergent and divergent validity against the revised Conners Teacher Rating Scale (CTRS; Trites, Blouin, & Laprade, 1982) and Child Behavior Checklist (CBCL;

Achenbach & Edelbrock, 1983). (For detailed presentation of ASCA scale development, item analyses, exploratory and confirmatory components analyses, and reliability and concurrent validity studies, refer to McDermott, 1993).

Procedure

The general goal was to test the ability of ASCA syndrome scores to correctly differentiate the independently classified nondisabled subjects and those with SED. Six basic types of analyses were performed: (a) discriminant analysis for the 300 matched subjects classified as nondisabled or disturbed, with classification accuracy assessed through resubstitution based on the resultant discriminant function; (b) cross-validation for the same matched sample using the repeated $n - 1$ reclassification procedure by Lachenbruch and Mickey (1968) with apparent and unbiased accuracy estimates based on posterior probabilities (Hora & Wilcox, 1982); (c) discriminant analysis for a random, 150-subject partition of the 300 matched subjects (where randomization was keyed so as to draw matched pairs of nondisabled children and those classified as disturbed) and classification accuracy determined through resubstitution; (d) cross-validation through application of the function derived from the foregoing calibration sample to classify members of the remaining 150-subject reserve sample; (e) a series of validity generalization studies applying the discriminant function derived for the full matched sample to classify respective subsamples of males, females, preadolescents (age 5 to 11), adolescents (age 12 to 17), Whites, and African Americans; and (f) a series of differential classification studies to discriminate subjects with SED from those classified as LD, CH, and GT, respectively, and from all others, including nondisabled children and those otherwise classified as exceptional.

Although properly a multivariate discrimination task, given nonzero correlations among syndromes (ranging from –.04 to .49), differentiation also was assessed through a simple cut-score method whereby any child retaining a T score at or above the 95th percentile ($T \geq 67$) was classified as SED. For all classification analyses, accuracy was expressed (as advised by Meehl & Golden, 1982) in terms of sensitivity (percentage of true positives) and specificity (true negatives). In addition, overall accuracy was derived from the ratio of true positives and negatives to all classifications. Because the import of these statistics is dependent on prior or chance probability of classification accuracy (McDermott, 1988; Robins, 1985; Sicoly, 1992), it was considered important to apply an inferential test of significance beyond chance. Thus, Fleiss's (1971) formulae for error variance of nominal scale classifications were applied using the computer program CONGRU (Watkins & McDermott, 1979), yielding kappa coefficients and associated tests of statistical significance under the normal deviate.

RESULTS

Bartlett's modification of the likelihood ratio test (Morrison, 1976) for both the full 300-subject and split 150-subject calibration samples indicated marked heterogeneity of the within-classification covariance matrices (see Note 3). Therefore, quadratic solutions were sought for both samples as specified by Anderson (1984) and Perlman (1980). Analysis for the full sample yielded a successful

solution where Wilks' lambda = .68, multivariate $F(6, 293) = 22.72$, and $p < .0001$ (see Note 4). Table 2 presents syndrome *T*-score means and standard deviations for nondisabled and SED children, and univariate tests of group differences. Significantly higher *T* scores were found for children with SED in oppositional defiance, both provocative and impulsive forms of solitary aggression, and attention-deficit hyperactivity.

Table 3 reports classification accuracy for the quadratic function applied to the full sample, first for simple resubstitution, and then cross-validation. Sensitivity

TABLE 2
T SCORE MEANS, STANDARD DEVIATIONS, AND SIGNIFICANT DIFFERENCES FOR ASCA CORE SYNDROMES ACROSS MATCHED SAMPLES OF NONDISABLED CHILDREN AND CHILDREN WITH SED

	Nondisabled		Social or emotional disturbance		
Core syndrome	*M*	*SD*	*M*	*SD*	Univariate *F*[a]
Attention-Deficit Hyperactive	51.73	11.68	58.85	7.76	38.61*
Solitary Aggressive (Provocative)	52.09	11.39	60.29	12.25	35.99*
Solitary Aggressive (Impulsive)	49.85	7.80	59.32	12.39	62.85*
Oppositional Defiant	50.61	10.83	65.25	13.23	110.00*
Diffident	50.47	10.00	51.88	9.88	1.50
Avoidant	50.48	11.68	50.83	10.56	0.07

Note. $N = 300$. Wilks' lambda = .68, multivariate $F(6, 293) = 22.72$, $p < .0001$.
[a]Univariate $df = 1$ and 298.
*$p < .0001$.

TABLE 3
ASCA CLASSIFICATION ACCURACY FOR NONDISABLED CHILDREN AND CHILDREN WITH SED MATCHED FOR AGE, GENDER, RACE, AND GRADE LEVEL

		Percentage accuracy		
Method	*N*	Sensitivity (true positives)	Specificity (true negatives)	Overall[a]
Full-sample resubstitution[b]	300	80.7	80.7	80.7*
Full-sample cross-validation[c]	300	78.7	78.0	78.3*
		(83.9)	(81.4)	(82.7*)
Split-sample resubstitution[d]	150	78.7	80.8	79.3*
Split-sample cross-validation	150	86.7	69.3	78.0*

[a]Tests of statistical significance are based on Fleiss's (1971) formulae for error variance of normalized congruence beyond chance (or kappa) as applied through computer program CONGRU (Watkins & McDermott, 1979). Kappa values corresponding to the five reported significance tests are .61, .57, .65, .59, and .56, respectively. Multiplied by 100, each value specifies percentage accuracy beyond chance. Corresponding percentages of false positives and false negatives are 19.3/19.3, 21.3/22.0, 16.1/18.6, 21.3/19.2, and 13.3/30.7. [b]Canonical $R = .56$, $F(6, 293) = 22.72$, $p < .0001$. [c]Parenthetical entries are unbiased accuracy-rate estimates based on posterior probabilities (Hora & Wilcox, 1982). [d]Canonical $R = .52$, $F(6, 143) = 8.82$, $p < .0001$.
*$p < .0001$.

and specificity were quite high and similar within each analysis, with overall accuracy highly significant and generally proximate to 80%. (Note: For all solutions posted in Table 3, chance accuracy was 50%.)

When a quadratic solution was tested for the random half of the children, successful group separation was again achieved, Wilks' lambda = .73, $F(6, 143) = 8.82$, $p < .0001$. Resultant accuracy indices, as shown in Table 3, essentially echoed those for the full sample. The subsequent cross-validation with the reserve sample indicated overall accuracy consistent with prior analyses, with sensitivity and specificity diverging somewhat. Specifically, an increase in efficiency for detecting children with SED (+8.7%) was offset by a corresponding loss in efficiency (–8.7%) for correctly recognizing children without disabilities.

Table 4 illustrates what occurred when the general discriminant solution was applied to each specific demographic subgroup within the larger matched sample. Although the number of subjects involved in each analysis differed, depending on frequencies in the initial sample, the prior proportions of subjects independently classified as normal and disturbed remained constant at 50/50. It will be observed that the accuracy rates remained consistently high (about 80%) and statistically significant overall.

Table 5 shows sustained accuracy for differentiating children with SED from alternative special education categories of children classified as learning or communication impaired and from those deemed gifted or talented. The analysis attempting to discriminate the 150 subjects with SED from 1,843 nonclassified students and 537 students with disabilities (the last entry in Table 5) also confirmed that even when youngsters with SED are mixed within the greater popu-

TABLE 4
GENERALITY OF ASCA CLASSIFICATION ACCURACY FOR NONDISABLED CHILDREN AND CHILDREN WITH SED BY DEVELOPMENTAL LEVEL, GENDER, AND RACE

		Percentage accuracy		
Analysis	*N*	Sensitivity (true positives)	Specificity (true negatives)	Overall[a]
Separate samples by age				
5 through 11 years	144	77.5	84.7	81.1*
12 through 17 years	156	83.5	76.9	80.3*
Separate samples by gender				
Male	228	82.5	79.8	81.1*
Female	72	75.8	83.3	79.2*
Separate samples by race				
White	244	79.5	82.8	81.1*
African American	50	88.0	76.8	82.0*

[a]Tests of statistical significance are based on Fleiss's (1971) formulae for error variance of normalized congruence beyond chance (or kappa) as applied through computer program CONGRU (Watkins & McDermott, 1979). Kappa values corresponding to the six reported significance tests are .62, .61, .62, .58, .62, and .64, respectively. Multiplied by 100, each value specifies percentage accuracy beyond chance. Corresponding percentages of false positives and false negatives are 22.5/15.3, 16.5/23.1, 17.5/20.2, 24.2/16.7, 20.5/17.2, and 12.0/23.2.

*$p < .0001$.

TABLE 5
ASCA ACCURACY FOR DIFFERENTIAL DISCRIMINATION OF CHILDREN WITH SED FROM LEARNING DISABLED, COMMUNICATION DISABLED, GIFTED/TALENTED, AND ALL OTHER CHILDREN

Analysis	*N*	Percentage accuracy		
		Sensitivity (true positives)	Specificity (true negatives)	Overall[a]
Differential of SED from:[b]				
360 learning disabled	510	80.7	73.1	76.9*
29 communication disabled	179	80.7	89.7	85.2*
60 gifted/talented	210	80.7	91.7	86.2*
Full child population[c]	2,530	80.7	79.1	79.2*

[a]Tests of statistical significance are based on Fleiss's (1971) formulae for error variance of normalized congruence beyond chance (or kappa) as applied through computer program CONGRU (Watkins & McDermott, 1979). Kappa values corresponding to the four reported significance tests are .73, .88, .88, .and .79, respectively. Multiplied by 100, each value specifies percentage accuracy beyond chance. Corresponding percentages of false positives and false negatives are 19.3/26.9, 19.3/10.3, 19.3/8.3, and 19.3/20.9. [b]These analyses discriminate 150 children with SED from respective groups of exceptional children. [c]This analysis discriminates 150 children with SED from 2,380 other children, including 1,843 children without disabilities and 537 children with various disabilities.
*$p < .0001$.

lation of children and demography is allowed to vary, ASCA's classificatory efficiency remains invariant and statistically significant. This is a particularly important discovery because, as Hosmer and Lemeshow (1989) evinced through maximum-likelihood estimates and Robins (1985) and Bird et al. (1987) demonstrated empirically in epidemiology, classification sensitivity ordinarily will tend to decrease as the proportion of nonmorbid cases in any population increases. ASCA's sensitivity suffers no such decrement.

In order to study more precisely the relative contribution of each core syndrome (and simultaneously to inform an understanding of the composition of the group regarded as SED), the bivariate and canonical relationships of the syndromes to group discrimination are presented in Table 6. The point-biserial *r*s (which are equivalent to correlations between the ASCA syndromes and the canonical variate representing the classification dichotomy) are statistically significant for the first four syndromes but not the latter two. This pattern is consistent with the univariate tests for group mean differences reported in Table 2. Likewise, the squared semi-partial *r*s indicate that the Oppositional Defiant syndrome is associated with about 27% of group distinction, followed by 17% for the Solitary Aaggressive (Impulsive) syndrome, and 11% each for the Solitary Aggressive (Provocative) and Attention-Deficit Hyperactive syndromes. The two underreactive syndromes, Diffident and Avoidant, play virtually no role in distinguishing children considered disturbed. This information is reflected moreover in the pattern of discriminant loadings (i.e., correlations between the syndromes and their own canonical variate).

As supplementary assessment of ASCA's classification accuracy, the simple cut-score approach was used to identify as maladjusted those at or above the 95th

TABLE 6
BIVARIATE, CANONICAL, AND WEIGHTED RELATIONSHIPS OF ASCA SYNDROMES WITH DISCRIMINATION OF NONDISABLED CHILDREN AND CHILDREN WITH SED

Core syndrome	Point-biserial *r*	Squared semi-partial *r*	Canonical discriminant loading	Discriminant weights[a]	
				Nondisabled	Social or emotional disturbance
Attention-Deficit Hyperactive	.34*	.11	.60	.38	.39
Solitary Aggressive (Provocative)	.33*	.11	.58	.15	.14
Solitary Aggressive (Impulsive)	.42*	.17	.74	.26	.32
Oppositional Defiant	.52*	.27	.92	.07	.15
Diffident	.07	.01	.13	.57	.61
Avoidant	.01	.00	.03	.18	.16

Note. $N = 300$.
[a]For convenience, raw discriminant weights are reported for direct weighting of respective ASCA syndrome *T* scores, where –41.39 is the constant for the nondisabled (or non-SED) category and –50.51 for the SED category, and where children are classified to the category associated with the higher sum of weighted *T* scores and constant.
*$p < .0001$.

percentile on any syndrome. Applied to the matched 300-subject sample, sensitivity equaled 79.3% and specificity equaled 56.0% (overall = 67.7%, kappa = .59, $p < .0001$). For the full population of 2,530 children, sensitivity equaled 78.7% and specificity equaled 89.1% (overall = 88.3%, kappa = .77, $p < .001$). This suggests that the cutting-score approach, although producing highly significant results, may be less consistent than the multivariate approach and may tend to vary as a function of the relative distributions of children with SED and children without disabilities. The cutting-score approach also assumes that all markedly deviant youths, including the diffident and avoidant, should be regarded as maladjusted—an assumption apparently at variance with common practice in special education.

DISCUSSION

The evidence consistently supports ASCA's claim of classification accuracy for detecting emotional disturbance among schoolchildren. Whereas apparent accuracy of diagnostic devices often can rely heavily on covariation of important but theoretically irrelevant constructs—particularly age, gender, and race (refer to APA, 1985, Standard 1.10)—ASCA's proficiency sustains whether or not those variables are held constant or whether application is confined to adolescents or younger children, males or females, or White or African American children. As a rule, accuracy tends to remain in the vicinity of 80%, a level that is appreciably beyond chance and of practical consequence. ASCA also effectively separates youths who are emotionally disturbed from those alternatively classified as learning or communication impaired or intellectually gifted or talented.

Comparison of these results to previous research is not a straightforward matter. The concepts of classification sensitivity and specificity, although common conventions in medical and taxometric research (Griner, Mayewski, Mushlin, & Greenlan, 1981; Meehl & Golden, 1982), only recently have been recognized as dual imperatives for psychological assessment. It is only within the last decade that psychological investigators have been inclined to report accuracy in these terms, and when information is reported, it usually does not address the notion of improvement beyond chance or generality for different segments of the population. Classification methods also vary widely, not only by the use of different rating or observation scales—some employing parent respondents, some teachers—but also by the alternate use of monotonic cut-score and multivariate decision rules. This great diversity must be appreciated when contrasting instruments. Nonetheless, it seems reasonable to surmise that the varied research efforts on classification accuracy were intended to test the efficiency of specific instruments applied in ways that would potentially optimize their performance. This common perspective lends value to a comparison of past and present evaluations of rating scale sensitivity and specificity.

In a study similar to those reported here, Stein and O'Donnell (1985) attempted to differentiate a generic group of children with SED from a group of children without disabilities using weighted factors of the revised CTRS (Goyette, Conners, & Ulrich, 1978). Initial overall accuracy was 67.0%, with 60.3% sensitiv-

ity and 80.7% specificity. Upon cross-validation, accuracy dropped to 60.7% overall, with 61.9% sensitivity and 59.6% specificity. Similarly, using teacher ratings on the Devereux Rating Scale–School Form (Naglieri, LeBuffe, & Pfeiffer, 1992), LeBuffe, Hess, Servis, and Waschbusch (1992) showed that, whereas overall accuracy for preadolescents was 75.0% and specificity was 97.0%, sensitivity was only 53.0%. Extended to adolescents, the scale yielded overall accuracy of 69.0%, specificity of 90.0%, and sensitivity of 48.0%. Results more comparable to those for ASCA were found by the same researchers when they applied Achenbach's (1991) Teacher's Report Form (TRF) with preadolescents, with overall accuracy, sensitivity, and specificity all being 78.0%. However, when employed with adolescents, sensitivity lowered to 64.0%, with specificity at 86.0% and overall accuracy at 75.0%.

Costenbader and Keller (1990) used cut scores on the CBCL (Achenbach & Edelbrock, 1983) to differentiate children with SED or LD and children without disabilities. With 58 children with SED and the children with LD and those without disabilities combined to form a group of nondisturbed, classification sensitivity was 56.9% and specificity was 81.3% (overall accuracy of 75.2%). When the CBCL and TRF were used in combination, sensitivity was 50.0% and specificity 81.3% (overall 74.8%). Also based on cut scores (and accuracy indices computed from raw data provided by Kazdin & Heidish, 1984), the CBCL's sensitivity for classifying 72 youths who were maladjusted was 94.4%, with specificity merely 12.2% (64.6% overall). Finally, a large-sample study by Steinhausen and Gobel (1987) that applied the Children's Behavior Questionnaire (Rutter, Tizard, & Whitmore, 1970) to classify more than 1,400 inpatient children produced an average sensitivity level of 27.8%, with specificity at 85.4% (78.5% overall).

Although sensitivity and specificity are decidedly the most widely accepted indices of classification accuracy (Kline, 1988; Meehl & Golden, 1982), Milich, Widiger, and Landau (1987) expressed a preference for indices of conditional predictive power when assessing the practical usefulness of different classification methods or rules. According to Dawes (1962), positive predictive power (PPP) indicates the utility of a classification method as an *inclusion* criterion and specifies the conditional probability of disorder (e.g., SED) given that the classification method predicts SED. Inversely, negative predictive power (NPP) indicates the value of the classification method as an *exclusion* criterion and specifies the conditional probability of not being SED given the absence of such a prediction. On the average, ASCA's PPP was 80.6% and NPP 78.0% across the analyses appearing in Tables 3, 4, and 5. In contrast, the classification methods presented above produce average PPP of 44.5% and NPP of 79.0%. Thus, whereas ASCA and the alternative methods function equivalently for avoiding misclassification of children without disabilities, ASCA appears to be superior for identifying children with SED and simultaneously meets the 75% PPP criterion established by Milich et al. (1987) for valuing a classification procedure as useful in clinical practice.

It would appear that high and reasonably equivalent sensitivity, specificity, and predictive power are more the exception than the rule. This is not to fault necessarily the various classification devices. It is quite conceivable that, whereas given empirical rating scales retain substantial integrity, the clinical or special

educational criterion groups to which they are expected to correspond may themselves lack reliability and validity (Kline, 1988; McDermott, 1988; Swets, 1988). With respect to the current study, it should be noted that ASCA's two underactive syndromes (Diffident and Avoidant) played virtually no role in the discrimination process, nor by inspection of canonical loadings was there any conceivable role that they might play in recognizing children otherwise deemed SED (see Note 5). In fact, upon reexamination it was discovered that only one out of six of the children considered SED by school personnel evidenced distinctively underactive or withdrawn behavior, most children having manifested overactive behavior problems.

This is not a unique observation. Schools and clinics tend to give more attention to acting-out disorders than to those of the withdrawn type, the latter viewed as a *low impact* problem (Strauss, Lahey, Frick, France, & Gynd, 1988). The potential danger this signals, however, is that seriously underactive youths will not benefit from the social conditioning or resources afforded those whose problem behavior is overt or interfering. This, in light of new evidence that underactive disturbance tends to increase in prevalence as youngsters move from childhood into adolescence (McDermott, in press), and with very destructive and lasting consequences, including school dropout and serious mental illness (e.g., Banez & Compas, 1990; Compas, Howell, Phares, Williams, & Giunta, 1989; Ollendick, Greene, Weust, & Oswald, 1990), raises a considerably more ominous alarm.

Although underactive disorders appear uncommon in special education samples, it is not incontrovertible that ASCA would prove highly useful for identifying such children. Clearly, ASCA's Diffident and Avoidant syndromes express behavior problems viewed as critical by teachers and that relate significantly to important academic and social problems (see McDermott, 1993). But ASCA is not intended as an exhaustive screen for all possible types of symptomatic behavior. ASCA is not a measure of behaviorally indistinct pathology such as anxiety or depression—constructs requiring inferences about feelings and emotions. (Note: Whereas there is strong evidence that adults typically are better informants about behavioral and most other phenomena, the generalization may not extend to subjective phenomena such as mood, feelings, etc. For a review, see Piacentini, Cohen, & Cohen, 1992.) It also is worth emphasizing that the behavioral domains assessed, although quite comprehensive, are nevertheless limited to what is detectable within the social context of common schooling. There is no attempt to address problems unique to children in residential care or whose difficulties relate primarily to eating, sleeping, or physiologic conditions that only a physician or parent might confirm. However, within the social milieu of common schooling, ASCA can provide valid and specific information, even when applied for children disabled scholastically, cognitively, or physically.

Optimally, useful measures of behavior pathology should be fairly uniform across developmental levels, gender, and ethnicity. They also should perform with comparable efficiency across such demographic groups. This position flows not only from the theoretical implausibility associated with psychological constructs that differ within species but also from the limited consequential validity (Messick, 1989) for measures that are developmentally discontinuous. ASCA's core syn-

dromes do afford appreciable and generalizable accuracy to children and adolescents, males and females, and Whites and African Americans. In addition to distinguishing children with SED from children without disabilities, they effectively distinguish children with SED from other exceptional youth, including those who are learning disabled, communication impaired, and gifted or talented, as well as from groups of youths with mixed disabilities.

Although ASCA provides a variety of applications and interpretation schemes (including syndromic profile clustering and logistic risk estimation), we limit the discussion to the current study. Practitioners will obtain best results if the discriminant functions procedure is applied to determine the relative probability of a child's membership in SED versus alternative classifications, including nondisabled. The procedure can weight ASCA's six core syndrome scores as based on the national data (this would be regarded as the most generalizable approach; see Note 4) or alternative weights can be derived from local populations. This method has the special advantage of considering automatically the idiosyncratic and differential relationships among the syndromes within SED and other groups.

The cut-score method, although less accurate, is much simpler and straightforward. The practitioner need only examine the syndrome *T* scores and consider any child having any *T* score above 66 (the 95th percentile) as a likely member of an SED classification. This approach essentially regards all six syndromes with equal weight, a procedure that in the long run tends to avoid confusion (the rules remain constant over local sites) and enhance accuracy (see Piacentini et al., 1992, on benefits of equal weighting). It should be understood, however, that the discriminant functions approach depends on the validity of the SED group upon which weights are based (Tharinger et al., 1986), and the utility of both discriminant and cut-score approaches rests on the reasonable balance of other contextual assessment influences (teacher competence, other information sources, etc.) and integrity of treatments that children with SED ultimately experience (Barnett & Macmann, 1992; McDermott, 1981).

CONCLUSIONS

Compared to other available scales for detecting child and adolescent maladjustment, ASCA appears to provide several advantages. It employs teacher observations and is able to distinguish social or emotional disturbance with high efficiency, irrespective of developmental level or gender, and with equivalent accuracy for African American and White youths. It effectively separates children regarded as emotionally disturbed from those considered learning disabled, speech or communication disabled, or gifted or talented. It also bases assessments on nationwide norms that stratify for pertinent demographics and disabling conditions, and it validates pathology through the observation of similar types of problem behavior across multiple and distinct situations. More precise appraisal of classification utility for ASCA and other scales ultimately will rest on each scale's ability to differentiate successfully among relevant subtypes of social and emotional disturbance—a task that will necessitate as much concern for the validity of the reference groups as for the scales under review.

Notes

1. Prior to sample selection, the significance of the variability in child behavior accounted for by age, gender, race, social class, geographic region, community size, and grade level was assessed for the ASCA norm sample via MANOVA and partialed canonical redundancy analyses. Age accounted for 1.2% of behavior variation ($p < .001$), gender 1.9% ($p < .001$), race 0.4% ($p < .05$), and grade level 1.2% ($p < .001$), and, consequently, they were applied as control matching parameters in the current study. No other demographic variable achieved statistical significance or accounted for more than 0.2% of behavioral variance.
2. The conversion assigns *T* scores to proper and comparable percentiles across syndromes. It standardizes but neither normalizes nor is intended to normalize scores.
3. This phenomenon is consistent with the differential, although natural, positive skews typifying pathology dimensions across youth populations that are nondisabled and disturbed (see McDermott, 1982).
4. Quadratic discriminant function weights and constants may be obtained by writing to the senior author.
5. As further illustration, when the two underactive syndromes were excluded from classification analyses, overall accuracy decreased only 0.7% to 3.4%, depending on the analysis applied. Analyses that included *only* the two syndromes produced near-chance accuracy.

References

Achenbach, T. M. (1988). Integrating assessment and taxonomy. In M. Rutter, A. H. Tuma, & I. S. Lann (Eds.), *Assessment and diagnosis in child psychopathology* (pp. 300–342). New York: Guilford.

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

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

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

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

Anderberg, M. R. (1973). *Cluster analysis for applications.* New York: Academic Press.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York: Wiley.

Banez, G. A., & Compas, B. E. (1990). Children's and parents' daily stressful events and psychological symptoms. *Journal of Abnormal Child Psychology, 18,* 591–605.

Barnett, D. W., & Macmann, F. M. (1992). Decision reliability and validity: Contributions and limitations of alternative assessment strategies. *The Journal of Special Education, 25,* 431–452.

Bird, H. R., Canino, G., Gould, M. S., Ribera, J., Rubio-Stipec, M., Woodbury, M., Heurtas-Goldman, S., & Sesman, M. (1987). Use of the Child Behavior Checklist as a screening instrument for epidemiological research in child psychiatry: Results of a pilot study. *Journal of the American Academy of Child and Adolescent Psychiatry, 26,* 207–213.

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.

Breen, M. J., & Altepeter, T. S. (1991). Factor structure of the Home Situations Questionnaire and the School Situations Questionnaire. *Journal of Pediatric Psychology, 16,* 59–67.

Brody, G. H., & Forehand, R. (1986). Maternal perceptions of child maladjustment as a function of the combined influence of child behavior and maternal depression. *Journal of Consulting and Clinical Psychology, 54,* 237–240.

Campbell, S. B., March, C. L., Pierce, E. W., Ewing, L. J., & Szumowski, W. K. (1991). Hard-to-manage preschool boys: Family context and the stability of externalizing behavior. *Journal of Abnormal Child Psychology, 19,* 301–313.

Campbell, S. B., & Steinert, Y. (1978). Comparisons of rating scales of child psycho-

pathology in clinic and nonclinic samples. *Journal of Consulting and Clinical Psychology, 46,* 358–359.

Compas, B. E., Howell, D. C., Phares, V., Williams, R. A., & Giunta, C. T. (1989). Risk factors for emotional/behavioral problems in young adolescents: A prospective analysis of adolescent and parental stress and symptoms. *Journal of Consulting and Clinical Psychology, 57,* 732–740.

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,* 485–496.

Dawes, R. M. (1962). A note on base rates and psychometric efficiency. *Journal of Consulting Psychology, 26,* 422–424.

Edelbrock, C., Costello, A. J., Dulcan, M. K., Kalas, R., & Conover, N. C. (1985). Age differences in the reliability of the psychiatric interview of the child. *Child Development, 56,* 265–275.

Elliot, C. D. (1990). *Differential Ability Scales: Introductory and technical handbook.* San Antonio, TX: Psychological Corp.

Fantuzzo, J. W., & McDermott, P. A. (1992). Critical issues in the assessment of family violence involving children. In R. Ammerman & M. Hersen (Eds.), *Assessment of family violence: A clinical and legal sourcebook* (pp. 11–25). New York: Wiley.

Ferguson, D. M., Lynskey, M. T., & Horwood, J. (1993). The effect of maternal depression on maternal ratings of child behavior. *Journal of Abnormal Child Psychology, 21,* 245–269.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76,* 378–382.

Garfield, S. L. (1978). Research problems in clinical diagnosis. *Journal of Consulting and Clinical Psychology, 46,* 596–607.

Goyette, C. H., Conners, C. K., & Ulrich, R. F. (1978). Normative data on revised Conners Parent and Teacher Rating Scales. *Journal of Abnormal Child Psychology, 6,* 221–236.

Griner, P. F., Mayewski, R. J., Mushlin, A. I., & Greenlan, P. (1981). Selection and interpretation of diagnostic tests and procedures: Principles and applications. *Annals of Internal Medicine, 94,* 557–592.

Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press.

Hora, S. C., & Wilcox, J. B. (1982). Estimation of error rates in several-population discriminant analysis. *Journal of Marketing Research, 19,* 57–61.

Horn, W. F., Wagner, A. E., & Ialongo, N. (1989). Sex differences in school-aged children with pervasive attention deficit hyperactivity disorder. *Journal of Abnormal Child Psychology, 17,* 109–125.

Hosmer, D. W., Jr., & Lemeshow, S. (1989). *Applied logistic regression.* New York: Wiley-Interscience.

Kazdin, A. E., & Heidish, I. E. (1984). Convergence of clinically derived diagnoses and parent checklists among inpatient children. *Journal of Abnormal Child Psychology, 12,* 421–435.

Kline, R. B. (1988). Methodological considerations in the evaluation of the convergence of psychiatric diagnoses and parent-informant checklists. *Journal of Abnormal Child Psychology, 16,* 289–298.

Lachenbruch, P. A., & Mickey, M. A. (1968). Estimation of error rates in discriminant analysis. *Technometrics, 10,* 1–10.

LeBuffe, P., Hess, R., Jr., Servis, L., & Waschbusch, D. (1992, August). *Identifying SED students: The Devereux scales vs. the TRF.* Paper presented at the Centennial Convention of the American Psychological Association, Washington, DC.

Loeber, R., Green, S. M., & Lahey, B. B. (1990). Mental health professionals' perceptions of the utility of children, parents, and teachers as informants on childhood psychopathology. *Journal of Clinical Child Psychology, 19*(2), 136–143.

Mattison, R. E., Bagnato, S. J., Mayes, S. D., & Felix, B. C., Jr. (1990). Reliability and validity of teacher diagnostic ratings for children with behavioral and emotional disorders. *Journal of Psychoeducational Assessment, 8,* 509–517.

McDermott, P. A. (1981). Sources of error in the psychoeducational diagnosis of children. *Journal of School Psychology, 19,* 31–44.

McDermott, P. A. (1982). Generality of disordered behavior across populations of normal and deviant school children: Factorial relations analyses. *Multivariate Behavioral Research, 17,* 69–85.

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-Interscience.

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. (in press). A nationwide study of developmental and gender prevalence for psychopathology in childhood and adolescence. *Journal of Abnormal Child Psychology.*

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

Meehl, P. E., & Golden, R. R. (1982). Taxometric methods. In P. C. Kendall & J. N. Butcher (Eds.), *Handbook of research methods in clinical psychology* (pp. 127–181). New York: Wiley.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education, Macmillan.

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

Morrison, D. F. (1976). *Multivariate statistical methods.* New York: McGraw-Hill.

Naglieri, J., LeBuffe, P., & Pfeiffer, S. (1992). *The Devereux School Form.* San Antonio, TX: Psychological Corp.

Ollendick, T. H., Greene, R. W., Weust, M. D., & Oswald, D. P. (1990). The predictive validity of teacher nominations: A five-year follow-up of at-risk youth. *Journal of Abnormal Child Psychology, 18,* 699–713.

Perlman, M. D. (1980). Unbiasedness of the likelihood ratio tests for equality of several covariance matrices and equality of several multivariate normal populations. *Annals of Statistics, 8,* 247–263.

Phares, V., Compas, B. E., & Howell, D. C. (1989). Perspectives on child behavior problems: Comparisons of children's self-reports with parent and teacher reports. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1,* 68–71.

Piacentini, J. C., Cohen, P., & Cohen, J. (1992). Combining discrepant diagnostic information from multiple sources: Are complex algorithms better than simple ones? *Journal of Abnormal Child Psychology, 20,* 51–63.

Quay, H. C. (1986). Classification. In H. C. Quay & J. S. Werry (Eds.), *Psychopathological disorders of childhood* (3rd ed., pp. 1–42). New York: Wiley.

Richters, J. E. (1992). Depressed mothers as informants about their children: A critical review of the evidence for distortion. *Psychological Bulletin, 112,* 485–499.

Robins, L. N. (1985). Epidemiology: Reflections on testing the validity of psychiatric interviews. *Archives of General Psychiatry, 42,* 918–924.

Rutter, M., Tizard, J., & Whitmore, K. (1970). *Education, health and behavior.* London: Longman.

Schaughency, E. A., & Lahey, B. B. (1985). Mothers' and fathers' perceptions of child deviance: Roles of child behavior, parental depression, and marital satisfaction. *Journal of Consulting and Clinical Psychology, 53,* 718–723.

Sicoly, F. (1992). Estimating the accuracy of decisions based on cutting scores. *Journal of Psychoeducational Assessment, 10,* 26–36.

Stein, M. A., & O'Donnell, J. P. (1985). Classification of children's behavior problems: Clinical and quantitative approaches. *Journal of Abnormal Child Psychology, 13,* 269–280.

Steinhausen, H. C., & Gobel, D. (1987). Convergence of parent checklists and child psychiatric diagnoses. *Journal of Abnormal Child Psychology, 15,* 147–151.

Strauss, C. C., Lahey, B. B., Frick, P., France, C. L., & Gynd, G. W. (1988). Peer social status of children with anxiety disorders. *Journal of Consulting and Clinical Psychology, 56,* 137–141.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285–1293.

Tharinger, D. J., Laurent, J., & Best, L. R. (1986). Classification of children referred for emotional and behavioral problems: A comparison of PL 94-142 SED criteria, DSM III, and the CBCL system. *Journal of School Psychology, 24,* 111–121.

Thorndike, R. L. (1982). *Applied psychometrics.* Boston: Houghton Mifflin.

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.

United States Code, 1977, Title 20, Secs. 1401(19), 1414(a)(5).

U.S. Department of Commerce. (1988). *Current population survey, March 1988* [Machine-readable computer file]. Washington, DC: Author.

U.S. Department of Commerce. (1990). School enrollment: Social and economic characteristics of students. *Current Population Reports* (Series P-20, No. 443). Washington, DC: Bureau of the Census.

U.S. Department of Education. (1989). *Eleventh annual report to Congress on the implementation of the Education of the Handicapped Act.* Washington, DC: Author.

Watkins, M. W., & McDermott, P. A. (1979). A computer program for measuring overall and partial congruence among multiple observers on nominal scales. *Educational and Psychological Measurement, 39,* 235–239.

Watson, S. M., Henggeler, S. W., & Whelan, J. P. (1990). Family functioning and the social adaptation of hearing-impaired youths. *Journal of Abnormal Child Psychology, 18,* 143–163.