Journal of Autism and Developmental Disorders, Vol. 22, No. 2, 1992

Brief Report: Analysis of the Internal Consistency of Three Autism Scales

Peter Sturmey
University of Birmingham, U.K.

Johnny L. Matson[1] **and Jay A. Sevin**
Louisiana State University

A number of autism assessments have been developed and refined over the last 20 years (Matson & Mulick, 1990). More and more, standardized autism scales are being employed in the diagnostic and assessment process. Currently, the most frequently used instruments include the Autism Behavior Checklist (ABC; Krug, Arick, & Almond, 1980), the Real Life Rating Scale (RLRS; Freeman, Ritvo, Yokota, & Ritvo, 1986), and the Childhood Autism Rating Scale (CARS; Schopler, Reichler, & Renner, 1988). A recent review of measures of autism has shown that, although there are data on interrater reliability and some validity data on these scales, relatively little is known about the internal consistency of these scales (Sturmey & Sevin, in press). Analyses of internal consistency are necessary to demonstrate item homogeneity. Item analyses of subscales are also necessary to demonstrate that item groupings are empirically meaningful. Items within a scale should be more highly correlated with each other than with the entire pool of items. Internal consistency data has important implications for construct validity (Anastasi, 1982).

Coefficient alpha for the CARS is reported to be .94 (Schopler et al., 1988). Garfin, McCallon, and Cox (1988) examined internal consistency for the CARS with both child and adolescent samples. Coefficient alpha was .79 for the children; item-total correlations also appeared adequate (median value across scales = .50; range = -.22 to .78). Similar results were reported for adolescents (alpha = .73; median interitem correlation = .40; range = -.17 to .73). However, internal consistency data for the

[1] Address all correspondence to Johnny L. Matson, Department of Psychology, 236 Audubon Hall, Louisiana State University, Baton Rouge, Louisiana 70803-5501.

0162-3257/92/0600-0321$06.50/0 © 1992 Plenum Publishing Corporation

ABC are limited (Volkmar et al., 1988). Krug et al. (1980) reported a split-half reliability of .87. In a second study with an autistic sample, split-half reliability for the total score of the ABC was reported as .70; split-half reliabilities for the five subscales ranged from .30 to .70 with a median value of .52. No data on the internal consistency of the RLRS were identified.

In this paper, analyses of internal consistency for all three scales are reported. Data were calculated based on completed protocols for a sample of 34 children and adolescents with pervasive developmental disorders. It is our view that a study of this type is important in establishing that a general construct, autism, is being assessed.

METHOD

Subjects

A total of 34 children and adolescents participated in the study. Subjects included all children with pervasive development disorders (PDD) referred to a university psychology clinic within a 15-month period. Subject demographics included the following: (a) age (range 2-22 years, M = 7 years 7 months), (b) sex (28 male, 6 female), and (c) race (16 white, 16 black, and 2 Asian). The sample included children with normal IQs (n = 6), mild mental retardation (MR) (n = 9), moderate MR (n = 11), and severe/profound MR (n = 8) according to AAMD criteria (Grossman, 1983).

Differential diagnoses of subjects were made by two of the authors based on parent interviews and direct and/or videotaped observations of each subject. The mean duration each subject was observed for diagnostic purposes was 3.5 hr, with a minimum duration of 2 hr. Most subjects were observed in multiple settings including home, school, and clinic. In addition, for reliability purposes, independent diagnoses for each subject were made by at least one professional not associated with the study (e.g., school psychologist, medical doctor). Diagnostic agreement between the authors and additional professionals was 100%.

Twenty-seven subjects met DSM-III-R criteria for autism. Seven of the subjects failed to meet DSM-III-R cutoff requirements (i.e., 8 of the 16 criteria). These seven subjects all exhibited symptoms in the three core areas (social, language, and sameness) and met at least 6 of the 16 criteria for autism. These individuals were classified as having pervasive developmental disorders not otherwise specified (PDD NOS).

Thus, the sample included children from multiple age groups, representing a wide range of intellectual functioning and severity of autistic symptoms. Sex ratio (approx. 4.5:1) and percentages of subjects with MR closely approximated national estimates for these figures (Ritvo & Freeman, 1978; Schreibman & Mills, 1983).

Assessment Procedures

Assessment sessions consisted of two parts, parent interview and direct child observation. Assessment sessions were conducted in experimental/interview rooms in a psychology clinic. Clinical psychology doctoral students acted as interviewers and behavioral raters. Raters were blind to the study's purpose.

Parent interviews and child assessments were conducted simultaneously in separate rooms. During interview sessions with parents, several behavioral rating scales were completed, including the ABC. In all cases, interview respondents were primary caregivers. During client observations, RLRS and CARS were completed for subjects, based on observations of children during a 30-min free play period.

Autism scales were completed in the context of a larger project concerning the assessment of autism. More detailed descriptions of assessment procedures, including training of raters and checks of the integrity of these procedures, are presented in Sevin, Matson, Coe, Fee, and Sevin (1991).

Analyses

For each instrument, item analyses were conducted for each subscale and for the instrument total score. Item analyses consisted of Cronbach's alpha, item–total (minus item) point-biserial correlations, and an analysis of interitem correlations. In addition, the number of "rogue" items, items that adversely affect internal consistency, for each scale and subscale were recorded. Using a conservative estimate, items with item–total (minus item) point-biserial correlations of zero or less were identified as rogue items.

RESULTS

Results are summarized in Table I. The ABC total score proved satisfactory having a high value of Cronbach's alpha and only 3 out of 57 items with negative item–total correlations. While the ABC total proved satisfactory, the subscales of the ABC proved less satisfactory. Only the

Table I. Internal Consistencies of the ABC, RLRS, and CARS and Their Respective Subscales[a]

Scale (no. of items)	Cronbach's alpha	Median (range) item–total (minus item) point-biserial correlation	Mean (range) interitem correlation	No. of "rogue" items
ABC				
Total (57)	.873	.320 (–.102 to .667)	.120 (–.503 to .673)	3
Sensory (9)	.466	.291 (–.240 to .381)	.130 (–.271 to .490)	1
Relating (12)	.690	.395 (–.031 to .55)	.161 (–.297 to .535)	1
Body and Object Use (12)	.790	.408 (.258 to .734)	.264 (–.172 to .673)	0
Language (13)	.383	.044 (–.161 to .426)	.044 (–.402 to .673)	3
Social and Self-Help (11)	.423	.232 (–.255 to .568)	.120 (–.503 to .472)	2
RLRS				
Total (47)	.841	.263 (–.235 to .706)	.090 (–.446 to .860)	5
Motor (7)	.419	.228 (.021 to .370)	.074 (–.118 to .345)	0
Social (9)	.677	.418 (.050 to .731)	.185 (–.200 to .785)	0
Affect (5)	.493	.300 (–.001 to .544)	.127 (–.147 to .498)	1
Sensory (16)	.645	.271 (.040 to .665)	.100 (–.258 to .615)	0
Language (10)	.629	.483 (–.179 to .636)	.168 (–.446 to .861)	1
CARS				
Total (15)	.851	.460 (.291 to .709)	.283 (–.123 to .709)	0

[a] Conventionally, values of Cronbach's alpha should exceed .6 and the median item–total (minus item) point biserial correlation should exceed .30.

Body and Object Use subscale had acceptable alpha and item–total correlations with no rogue items. The Relating scale was adequate (coefficient alpha ≥ .60, few rogue items). The three reamining scales, Sensory, Language, and Social and Self-Help, were unacceptable (low coefficient alphas, item–total, and interitem correlations). Rogue items for all scales are listed in Table II.

Table II. Rogue Items for the ABC and RLRS

Scale	Item
	Autism Behavior Checklist
Total	Insists on keeping certain objects with him/her Has "special abilities" in one area Actively avoids eye contact
Sensory	Painful stimuli . . . evoke no reaction
Relating	Often frightened or very anxious
Body and Object Use	(none)
Language	Does not follow simple commands Seldom uses "yes" or "I" Uses at least 15 but < 30 phrases daily
Social and Self-Help	Learns a simple command but forgets quickly Has "special abilities" in one area
	Real Life Rating Scale
Total	Whirls Genital manipulation Other (affect) Lines up objects Noncommunicative use of delayed echolalia
Motor	(none)
Social	(none)
Affect	Other
Sensory	(none)
Language	Noncommunicative use of delayed echolalia

The RLRS total had a high value of Cronbach's alpha and the mean item-total correlation was only a little below the conventionally acceptable value of .30. Three of the five RLRS subscales, Social Relationships, Sensory, and Language, had adequate to good internal consistencies (coefficient alpha ≥ .60; item-total correlations ô .30; few rogue items; etc.).

The CARS total score was internally consistent on all indices with no rogue subscales/items.

DISCUSSION

The present study replicates previous studies demonstrating good internal consistency of the CARS (Garfin et al., 1988; Schopler et al., 1988). Of particular interest, these findings are in contrast to Garfin et al. (1988), who found that Scale 14 (Intellectual Response) of the CARS was negatively correlated with the full scale. This finding was not replicated here.

As noted by our reviewers, one possible explanation for this inconsistency is that Scale 14 may tend to correlate negatively with other scales when the sample predominantly includes intellectually disadvantaged individuals with isolated intellectual skills. Because these "islets of normal functioning" are rare, they may not always be represented in small samples. However, in some samples, this phenomenon may occur with sufficient frequency to yield a negative item–total correlation. Nevertheless, the current data indicate that problems in level and consistency of intellectual response are positively associated with full-scale scores. Based on our limited sample, dropping Scale 14, suggested by Garfin et al. (1988), does not appear warranted.

Adequate full-scale consistency for the ABC and RLRS were also demonstrated. However, the subscales of both the ABC and RLRS were variable in their internal consistency. Poor reliability of the Language scale of the ABC may be related to the scale's equal emphases on both expressive and receptive language items which may be weakly associated in some subjects. Low indices of internal consistency for the Social and Self-Help Scale of the ABC are not entirely unexpected given the wide variety of behaviors included in this scale. Despite the name of the scale, it includes items related to aberrant behaviors (e.g., severe temper tantrums) and intellectual functioning (e.g., Has "special abilities" in one area) as well as social behaviors and self-help skills. Poor reliability for the Motor and Affect subscales of the RLRS are possibly related to low variance associated with these scales for our sample. Many of the items on these subscales were not endorsed for any subjects. Examination of the original papers on the development of the ABC (Krug et al., 1980) and the RLRS (Freeman et al., 1986) show that the subscales of these instruments were constructed in an ad hoc fashion. That is, items were grouped into subscales based primarily on visual inspection or face validity. In the absence of any factor-analytic studies of these measures, and in light of the high internal consistency of their total scores, we recommend researchers and clinicians consider the use of the total scores only.

In conducting an internal consistency study, one is primarily concerned about errors caused by content sampling (Crocker & Algina, 1986). Consistency of responding is an important consideration if an individual's performance on test items is generalized to a larger domain of items. If an individual performs consistently across subsets of items within a test, we can have some confidence that this performance would generalize to other possible items in the content domain (Crocker & Algina, 1986). Based on our data and other psychometric studies, the ABC, RLRS, and CARS may represent useful measures of the construct of autism. However, ABC and RLRS subscale scores may not provide adequate estimates of a subject's performance in *specific behavioral domains* (e.g., language or social functioning on the ABC).

Several additional issues should be considered. First, an item may decrease the internal consistency of a scale while *improving* predictive validity. Restated, an item can have important diagnostic significance yet lower the value of coefficient alpha. Thus, before a rogue item is dropped, the effects of the change on the validity of the scale should be assessed. When an item contributes to the validity of an instrument but adversely affects subscale internal consistency, one solution is to simply remove the item from the otherwise homogeneous subscale while still retaining it as a separate predictor. Second, at present, computing total score alphas is acceptable; the assumption that there is a single latent variable underlying these instruments is reasonable in the absence of alternative data. However, subjecting these instruments to factor analyses in future studies would provide us with important additional information regarding scale dimensionality. Overall alphas can be high even if scales are multifactorial. In these cases, high alphas may be misinterpreted as evidence that scales are unidimensional. Third, discussions of internal consistency or homogeneous item samples should not be confused with discussions of homogeneous subtypes of pervasive developmental disorders. Good internal consistency of items of an autism scale should not be cited as evidence that autism is a behaviorally homogeneous disorder. It is possible for the internal consistency of a scale to hold up across essentially different diagnostic subgroups within autism (Sevin, Matson, Coe, Love, Matese, & Benevidez, 1991).

This and previous studies have generally employed small subject samples. Future studies using larger and more heterogenous populations might find that results are more robust. Future research could also focus on comparing current ad hoc derived subscales with subscales derived using more empirical techniques such as factor analysis.

REFERENCES

Anastasi, A. (1982). *Psychological testing*. New York: Macmillan.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston, Inc.

Freeman, B. J., Ritvo, E. R., Yokota, A., & Ritvo, A. (1986). A scale for rating symptoms of patients with the syndrome of autism in real life settings. *Journal of the American Academy of Child and Adolescent Psychiatry, 25*, 130-136.

Garfin, D. G., McCallon, D., & Cox, R. (1988). Validity and reliability of the Childhood Autism Rating Scale with autistic adolescents. *Journal of Autism and Developmental Disorders, 18*, 367-378.

Grossman, H. J. (1983). *Classification in mental retardation*. Washington DC: American Association on Mental Retardation.

Krug, D. A., Arick, J., & Almond, P. (1980). Behavior checklist for identifying severely handicapped individuals with high levels of autistic behavior. *Journal of Child Psychology and Psychiatry, 21*, 221-229.

Matson, J. L., & Mulick, J. A. (1990). *Handbook of mental retardation* (2nd ed.). New York: Pergamon.

Ritvo, E. R., & Freeman, B. J. (1978). National Society for Autistic Children Definition of the Syndrome of Autism. *Journal of Autism and Developmental Disorders, 8*, 162-170.

Schopler, E., Reichler, R. J., & Renner, B. R. (1988). *The Childhood Autism Rating Scale*. Los Angeles: Western Psychological Services.

Schreibman, L., & Mills, J. I. (1983). Infantile autism. In T. A. Ollendick & M. Hersen (Eds.), *Handbook of child psychopathology*. New York: Plenum Press.

Sevin, J. A., Matson, J. L., Coe, D. A., Fee, V. E., & Sevin, B. M. (1991). A comparison and evaluation of three commonly used autism scales. *Journal of Autism and Developmental Disorders, 21*, 417-432.

Sevin, J. A., Matson, J. L., Coe, D., Love, S. R., Matese, M. J., & Benevidez, D. A. (1991). Empirically derived subtypes of pervasive developmental disorders: A cluster analytic study. Unpublished manuscript.

Sturmey, P., & Sevin, J. A. (in press). Defining and assessing autism. In J. L. Matson (Ed.), *Autism in children and adults: Etiology, assessment, and intervention*. Sycamore, IL.: Sycamore Press.

Volkmar, F., Cicchetti, D. V., Dykens, E., Sparrow, S. S., Leckman, J. F., & Cohen, D. J. (1988). An evaluation of the Autism Behavior Checklist. *Journal of Autism and Developmental Disorders, 18*, 81-97.