Creating Abbreviated Rating Scales to Monitor Classroom Inattention-Overactivity, Aggression, and Peer Conflict: Reliability, Validity, and Treatment Sensitivity

Robert J. Volpe
Northeastern University

Kenneth D. Gadow
Stony Brook University

Abstract. Rating scales developed to measure child emotional and behavioral problems typically are so long as to make their use in progress monitoring impractical in typical school settings. This study examined two methods of selecting items from existing rating scales to create shorter instruments for use in assessing response to intervention. The psychometric properties of two sets of abbreviated rating scales derived from the IOWA Conners Teacher Rating Scale and the teacher-completed Peer Conflict Scale were examined and compared to the longer original versions of these scales. The rating scales were evaluated using data from a randomized, placebo-controlled, crossover trial of immediate release methylphenidate involving a sample 65 children between 6 and 12 years old who met research diagnostic criteria for attention deficit hyperactivity disorder and either chronic motor tic disorder or Tourette’s disorder. Specifically, the abbreviated and original versions of the rating scales were examined for internal consistency, temporal stability, concurrent validity, and treatment sensitivity. Results indicate that there were few significant differences between versions of the scales, which support the use of abbreviated rating scales for use in progress monitoring. Implications for practice and future research are discussed.

A problem-solving approach to treating child emotional and behavior problems involves continuous monitoring of student behavior in response to intervention for the purpose of assessing progress toward predetermined goals. Unfortunately, efforts to develop

This study was supported, in part, by a research grant from the Tourette Syndrome Association, Inc., and Public Health Service Grant MH 45358 from the National Institute of Mental Health. The authors thank their dedicated colleagues Jeffrey Sverd, MD, Joyce Sprafkin, PhD, Jayne Schneider, PhD, and Linda Volkersz for conducting clinic evaluations, and Michele and Michael De Angelis for providing pharmacy services. The authors are particularly indebted to the families who made this study possible.

Correspondence regarding this article should be addressed to Robert Volpe, Northeastern University, Department of Counseling and Applied Educational Psychology, 202 Lake Hall, Boston, MA 02115; E-mail: r.volpe@neu.edu

Copyright 2010 by the National Association of School Psychologists, ISSN 0279-6015, which has nonexclusive ownership in accordance with Division G, Title II, Section 518 of P.L. Law 110-161 and NIH Public Access Policy
feasible methods for monitoring emotional and behavior problems have been limited by the lack of a general outcome measure of behavioral health akin to those available to assess academic progress (Shinn, 2007). This raises the following question: “What are the appropriate targets of measurement when assessing responsiveness to interventions for emotional and behavior problems?”

Targets of measurement should be aligned with the objectives of intervention, and should include both short-term performance objectives and broader long-term general objectives (Kratochwill & Bergan, 1990). Two possible long-term objectives are the severity of symptomatology related to one or more emotional or behavioral disorders and levels of functional impairment such as those exhibited in the areas of social, academic, and family functioning (Pelham, Fabiano, & Massetti, 2005). Specific behaviors targeted for intervention should be imbedded into one of these broad categories, although other target behaviors are typically of interest. There are myriad rating scales that assess the symptoms of the most common Diagnostic and Statistical Manual, 4th edition (DSM-IV; American Psychological Association, 1994) disorders affecting children and adolescents as well as areas of functional impairment (see Merrell, 2008; Pelham et al., 2005). However, using such measures as the sole indicator of treatment response would not be prudent because they would not be as sensitive to change as instruments measuring specific target behaviors (e.g., gets out of seat, yells), particularly in the case of interventions that target a narrow set of target behaviors or when the target behaviors are not key indicators of the construct of interest. Indeed, these scales may not be closely aligned with the referral concern. Conversely, restricting measurement to one or two specific target behaviors precludes assessment of potential broader impacts of intervention. Therefore, a comprehensive progress monitoring system (CPMS) would need to measure both broad domains of emotional and behavior disorders and related areas of impairment and specific behaviors targeted for intervention. To date, such a system has generally not been feasible outside the realm of well-funded research studies (see Volpe, Heick, & Gureasko-Moore, 2005).

In formulating a CPMS, one must balance both psychometric and feasibility concerns (see Briesch & Volpe, 2007). In other words, the ability of the system to provide valid and reliable estimates of change must be weighed relative to the willingness of the assessor and informants to engage in the process. In the following section we briefly review several behavioral assessment methods, focusing on both concerns.

There has long been an emphasis on systematic direct observation to assess treatment response, wherein specific behaviors targeted for intervention are operationalized and coded systematically (e.g., Hintze, Volpe, & Shapiro, 2008; Merrell, 2008). Because the interpretation of systematic direct observation data requires little inference and has compelling face validity, it has long been considered the “gold standard” of behavioral assessment. Although systematic direct observation offers highly sensitive assessment of a limited set of target behaviors, its focus is relatively narrow. As systematic direct observations require trained observers and may necessitate several hours of observation to obtain dependable estimates of target behaviors (see Hintze & Matthews, 2004; Volpe, McConaughy, & Hintze, 2009), it is not generally considered to be a feasible strategy for sustained progress monitoring in typical school settings (Chafouleas, McDougal, Riley-Tillman, Panahon, & Hilt, 2005). Furthermore, systematic direct observation is best suited for assessing overt behaviors of moderate frequency (Hintze et al., 2008) and is poorly suited for measuring internalizing problems (Volpe, McConaughy, & Hintze, 2009).

Rating scales are among the most widely used assessment methods in school settings and are available for assessing a wide variety of constructs. Commercially available rating scales are generally well validated and generally exhibit exceptional psychometric characteristics. They also enable the measurement of low-frequency behaviors through cost-efficient sampling of behavior over a long
period of time (Merrell, 2008). In addition, they are easily administered across several settings, and little or no training is required to complete them. One limiting factor of many rating scales for application in a CPMS is their length. Some rating scales are fairly short (e.g., 18 items). However, continuous monitoring makes the burden on informants multiplicative, which is a serious threat to their feasibility in a CPMS (e.g., Volpe et al., 2005). In a school-wide response to intervention model in which theoretically 15% of all students will require progress monitoring, the amount of time required to complete ratings for each child is a significant consideration. For example, if an 18-item rating scale were used for weekly progress monitoring in a school with 500 students, each week teachers would be completing ratings of 1,350 items.

To address feasibility concerns associated with systematic direct observation and commercial rating scales in public school settings, authors have suggested the use of more global measures such as the daily report card (e.g., Pelham, 1993), or the construction of customized rating scales by selecting items from existing measures (Hyman et al., 1998). Both of these approaches have received increased attention in recent years (Chafouleas et al., 2005; Volpe, Gadow, Blom-Hoffman, & Feinberg, 2009).

Daily report cards, which more recently have come to be called direct behavior ratings (DBR), combine elements of both systematic direct observation and rating scales. They require informants to rate operationally defined target behaviors across a prespecified period of time (e.g., separate ratings during each class period) each day they spend in contact with the child. Advantages of DBR include cost-effectiveness, brevity, and adaptability for specific applications. For example, new DBR items can be created for specific situations, but in such cases, nothing is known about the psychometric properties. Heretofore, only a small number of DBR items has been investigated (e.g., Chafouleas et al., 2005; Chafouleas, Christ, Riley-Tillman, Briesch, & Chanese, 2007), and although initial results are encouraging, evidence concerning concurrent validity and treatment sensitivity is limited (Pelham et al., 2005). DBR hold great potential for the assessment of observable target behaviors (Briesch, Chafouleas, & Riley-Tillman, 2010; Riley-Tillman, Chafouleas, Sassu, Chanese, & Glazer, 2008) and may eventually prove useful in the assessment of broader constructs (see Chafouleas et al., 2007).

Selecting items from commercial rating scales to create abbreviated versions capitalizes on the advantages of the former while at the same time minimizing the constraints on feasibility imposed by their length. Several authors of commercial rating scales have created shortened versions of well-established instruments for use as screening measures (see Volpe & DuPaul, 2001, for a review). Although several have been used in medication titration studies (e.g., Gadow, Sverd, Sprafkin, Nolan, & Ezor, 1995; Ullman, Sleator, & Sprague, 1997), they often are considered to be too long for prolonged progress monitoring in typical school settings.

Recent attention has focused on developing brief behavior rating scales specifically for progress monitoring (Gresham et al., 2010; Meier, McDougal, & Bardos, 2008; Sprafkin, Mattison, Gadow, Schneider, & Lavigne, in press; Volpe, McConaughy, & Hintze, 2009). Several methods have been employed, which can be divided into nomothetic and idio- graphic approaches. Nomothetic approaches involve the use of data from large samples to create scales with acceptable psychometric characteristics, which then can be used to assess individual children. One common nomothetic approach involves selecting items from existing rating scales with the highest factor loadings (factor-derived approach). In this approach, only the items that most strongly are associated with the construct of interest are retained. Alternatively, items can be selected based on their relative sensitivity to change. The former approach is purely statistical, but the latter strategy uses both statistical (see Gresham et al., 2010) and more complicated decision rule selection including statistical methods (Meier, 1997; Meier et al., 2008). One key advantage of nomothetic approaches is that once developed, their psychometric
properties can be easily evaluated. There are, however, potential drawbacks. For example, the degree to which results from large samples of children generalize to individual children will vary widely. As for methods of selecting treatment sensitive items, it is reasonable to assume that the degree to which particular items are sensitive to change may not generalize across samples or interventions.

Idiographic approaches to developing shortened rating scales involve utilizing knowledge about individual children, either from anecdotal evidence or prior assessments, to create customized rating scales specific to each child. Two idiographic approaches to developing brief behavior rating scales have been employed. The first method involves allowing informants to select items from several relevant rating scales and to rate those items on a common Likert scale each day (see Hyman et al., 1998). Hyman et al. noted that such a “menu-based” approach has the potential benefit of increasing informant ownership in the evaluation process, which may serve to increase compliance to the assessment protocol. An alternative idiographic approach involves initially administering a full-length rating scale at baseline, and then selecting a small sample of the highest rated (most problematic) items for progress monitoring (Volpe, McConaughy, & Hintze, 2009). Idiographic approaches have the advantage of being tailored for each student, which should increase the social validity of the assessment process. However, because scales are potentially different for each child, evaluating the psychometric properties of these individualized scales is challenging. In addition, when extreme scores are selected at one point in time, regression to the mean may lead to overestimates of treatment effects. In other words, when two measures are less than perfectly correlated, extreme scores on one measure tend to be less extreme on the next (see Cook & Campbell, 1979).

Recently, Volpe, McConaughy, & Hintze (2009) compared two methods of creating brief teacher rating scales in a sample of children diagnosed with attention-deficit hyperactivity disorder (ADHD) who were involved in a medication titration procedure involving a no-treatment baseline and three doses of methylphenidate. In this study, the reliability and treatment sensitivity of the two 9-item DSM-IV-based ADHD scales from the ADHD-Symptom Checklist—4 (Gadow & Sprafkin, 1997) were compared to shortened 4-item versions of the scales. The shortened versions were created by either selecting items with the highest factor loadings (factor-derived approach) or selecting items with the highest ratings in a no-treatment condition (individualized approach). Results from this preliminary study were encouraging in that there were few notable differences in psychometric indicators across the two rating scale methods. The individualized method appeared to be the most sensitive to medication effects. However, the medication evaluation procedure did not include double-blind placebo control, and doses of medication were not randomized or counterbalanced to control for order effects. Therefore, it was difficult to draw firm conclusions about the relative treatment sensitivity of the individualized method. In addition, the study focused on reliability and treatment sensitivity and did not investigate the concurrent validity of the shortened measures. Finally, the sample for the study was small and included a wide age range of participants (between 4 and 17 years of age).

The purpose of the current study was to examine further the psychometric properties of abbreviated rating scales. This study builds on our prior efforts in several ways. First, the current study was conducted using stringent experimental controls (double-blind placebo controlled, counterbalanced dose conditions) in a large sample of children with oppositional defiant disorder and both comorbid chronic multiple tic disorder and ADHD. In the current study, we evaluated the psychometric properties of scales created using both factor-derived and individualized methods and compared them to the full-length original scales for three constructs related to disruptive child behavior (e.g., inattention-overactivity [IO], aggression [AG], and peer conflict [PC]). In addition to reliability and treatment sensitivity, we also examined the concurrent validity of...
of the original and abbreviated rating scales using a battery of external validation measures, including full-length rating scales measuring the same or similar constructs and systematic direct observation. It was hypothesized that the three rating scale methods would demonstrate adequate psychometric properties, and that there would be few notable differences across methods.

This article addresses the objectives of the special series through presentation of empirical data demonstrating how targets of measurement selected from behavior ratings scales can be selected to generate reliable CPMS data.

Method

Participants

Participants consisted of 65 children (53 boys, 12 girls) between 6 and 12 years old (mean = 8.9; SD = 1.8) who were recruited from a variety of sources (clinics, schools, media advertisements, parent support groups) for participation in a larger study designed to investigate whether immediate release methylphenidate (IR-MPH) is an effective treatment for children diagnosed with oppositional defiant disorder, ADHD, and chronic multiple tic disorder or Tourette’s disorder. Each child participated in a randomized, placebo-controlled, crossover trial of IR-MPH. This study was approved by a university institutional review board. Prior to enrollment, parents were repeatedly warned of the possible risk of irreversible tic exacerbation consequent to IR-MPH treatment. Parents and children (≥11 years) provided written informed consent and assent, respectively.

Inclusion criteria. To participate in the study each child had to meet Diagnostic and Statistical Manual, third edition, revised (American Psychiatric Association, 1987) or DSM-IV (American Psychiatric Association, 1994) diagnostic criteria for ADHD and either chronic motor tic disorder or Tourette’s disorder. Moreover, most children were above the clinical cutoff on two out of three parent- or teacher-completed ADHD behavior rating scales. Each child met ADHD clinical criteria in both school and home. Almost all children met research diagnostic criteria (Kurlan, 1989) for Tourette’s disorder, either definite or by history.

Comorbidities. Comorbid disorders were assessed with the parent interview version of the Diagnostic Interview for Children and Adolescents (Reich, 2000), which is a semistructured clinical interview. Approximately one half met criteria on the Diagnostic Interview for Children and Adolescents for oppositional defiant disorder, and one third met criteria for an anxiety disorder, sometimes in conjunction with a depressive disorder.

Exclusion criteria. Children who exhibited one or more of the following were excluded from the study: (a) their tics were the major clinical management concern; (b) they were too severely ill (dangerous to self or others), psychotic, or mentally retarded (IQ < 70); (c) there were concerns about danger to self or other; or (d) there was evidence of a seizure disorder, major organic brain dysfunction, major medical illness, medical or other contraindication to medication (other than tics), or pervasive developmental disorder.

Procedure

Participants received placebo and three doses of methylphenidate (0.1 mg/kg, 0.3 mg/kg, and 0.5 mg/kg) for 2 weeks each under double-blind conditions. The upper limit for the 0.5 mg/kg dose was 20 mg. All children received Novartis Brand Ritalin-IR (i.e., generic substitutions were not permitted). Dose schedules were counterbalanced and assigned on a random basis. Medication was administered twice daily, approximately 3.5 hr apart, 7 days a week, and dispensed in dated, sealed envelopes at 2-week intervals. Detailed descriptions of the procedure and all measures have been previously published (Gadow et al., 1995; Gadow, Sverd, Nolan, Sprafkin, & Schneider, 2007).

Once the medication evaluation began, each child and at least one parent (typically the mother) were required to come to the clinic at
2-week intervals for interviews with the prescribing physician that included observations of tics and assessment of clinical status and cardiovascular/weight measurement. Approximately 1.25 hr after drug ingestion, all children were observed in a simulated classroom. Parents and teachers completed a battery of rating scales for each child’s behavior, 2 days per week for the duration of the drug evaluation. Teachers rated child behavior during periods of maximum drug efficacy (specific time duration for the morning) on Tuesdays and Thursdays.

Measures

The Inattention/Overactivity With Aggression (IOWA) Conners Teacher’s Rating Scale (Loney & Milich, 1982) contains two 5-item subscales, Inattention-Overactivity (IO) and Aggression (AG). Individual items are rated 0 = not at all, 1 = just a little, 2 = pretty much, 3 = very much. The IOWA scales demonstrate adequate internal consistency (.80–.92) and 1-week test–retest reliability (r values = .86–.89). A public school direct observation study indicated that IO scores correlated with off-task behavior (r = .46), and the AG scores correlated with non-compliance (r = .60) and interference (r = .43; Nolan & Gadow, 1994).

The Peer Conflict Scale (Gadow, 1986) contains 10 items that assess interpersonal peer aggression, are based on the physical and nonphysical aggression categories of the ADHD School Observation Code (Gadow, Sprafkin, & Nolan, 1996), and correspond to 6 of the 15 DSM-IV-defined symptoms of conduct disorder (Gadow & Sprafkin, 2008). Individual items are rated 0 = never, 1 = sometimes, 2 = often, 3 = very often. The findings of numerous studies of children between 3 and 12 years have indicated that the Peer Conflict Scale demonstrates excellent psychometric properties (Sprafkin, Gadow, & Nolan, 2001; Gadow & Nolan, 2002).

The Child Symptom Inventory—4 (CSI-4; Gadow & Sprafkin, 2002) is a teacher-completed rating scale containing 77 items that screen for the behavioral symptoms of most childhood disorders described in the DSM-IV. Individual items bear one-to-one correspondence with DSM-IV symptoms. Symptom severity is assessed as follows: 0 = never, 1 = sometimes, 2 = often, 3 = very often. In this study, we used severity scores for ADHD—Inattentive (ADHD: 9 items), ADHD—Hyperactive-Impulsive (ADHD, 9 items), and oppositional defiant disorder (8 items) symptom categories. Numerous studies indicate that the CSI-4 demonstrates satisfactory internal consistency (Cronbach’s alpha), reliability, and validity in community-based normative, clinic-referred, and ADHD samples (Gadow & Sprafkin, 2008; Gadow, Sprafkin, Salisbury, Schneider, & Loney, 2004).

For the classroom analogue observation (Roberts, Ray, & Roberts, 1984), the child sits alone at a desk in a small classroom and is instructed to complete worksheets and not to play with toys on an adjacent table. The 15-min observation sessions are divided into 180 five-second intervals, and behaviors are coded as either present or not present in each interval. Off Task is defined as not attending to the worksheet for any part of the five-second interval. Out of Seat is defined as the child not sitting or kneeling on his or her chair at any point during an interval. The simulated classroom code categories show high inter-rater reliability when used by trained observers, temporal stability over a 2-year time interval, correlation with corresponding teacher-completed behavior rating scales (Off Task), discriminant validity for nonreferred children and ADHD children with and without comorbid oppositional behavior, and sensitivity for assessing stimulant drug and dose effects (Gadow et al., 1995, 2007; Milich, 1984; Milich, Loney & Landau, 1982; Roberts, 1990; Roberts et al., 1984).

Scale Construction

Teachers were administered the full 5-item IO and AG scales of the IOWA Conners and the full 10-item Peer Conflict Scale. The original scales were examined along with two sets of 3-item scales created using two
methods (the factor-derived and individualized approaches are described in the following paragraphs). The number of items for the abbreviated scales was selected in an attempt to minimize the length of each scale, while still maintaining adequate technical characteristics.

The factor-derived method involved selecting items from the original scales with the highest factor loadings. Because of space limitations, only a brief description of the factor-analytic studies is provided here. Two principal component analyses were conducted using direct oblimin rotation, which does not assume that factors are uncorrelated. Teacher ratings for the IO and AG scales for 936 clinic-referred children between the ages of 5 and 13 years \( (M = 9.73; SD = 2.49) \) were submitted for principal component analysis. A two-factor solution explained approximately 69% of the variance in scores. Three items from the IO (“fails to finish things he starts—short attention span,” “constantly fidgeting,” “inattentive, easily distracted”) and AG (“temper outbursts, explosive and unpredictable behavior,” “quarrelsome,” “defiant”) scales with the highest factor loadings comprised the factor-derived scales. The average factor loadings of items retained for the factor-derived IO (0.82) and AG (0.90) scales were notably higher than the average factor loadings for the remaining items (0.54 and 0.81, respectively).

Teacher ratings on the Peer Conflict Scale were obtained from a normative data study of the ADHD-Symptom Checklist—4. Teacher ratings were completed for students in regular education \( (N = 493) \) who were between 5 and 12 years of age \( (mean = 8.1; SD = 1.65) \). The 10 items of the Peer Conflict Scale were submitted to principal component analysis using direct oblimin rotation. All items loaded on a single factor, which accounted for approximately 63% of the variance in scores. The average factor loadings for the 3 items selected for factor-derived PC scale (“gives dirty looks or makes threatening gestures to other children,” “hits, pushes, or trips other children,” “threatens to hurt other children”) were only slightly higher (.83) than loadings for the remaining items (.78).

Next, we created individualized scales for each participant by selecting 3 items with the highest ratings in the placebo condition. In most cases there were 3 or more items endorsed with the highest possible item rating (item score = 3). For cases in which there were more than 3 items rated as occurring very often or very much, an online random number generator (Random.org) was used to select items. In some cases, fewer than 3 items received ratings of “3.” In these cases, the same procedure was employed to select the balance of highest rated items.

Results

Reliability Analyses

The internal consistency and temporal stability of the three constructs (IO, AG, PC) across three methods (factor-derived, individualized, and full original scale) was examined across both placebo and low-dose conditions. Internal consistency was calculated using data from the first assessment occasion in the placebo and low-dose conditions, respectively. There were no data missing for this analysis. We selected the criterion for acceptable alpha coefficients at .70 based on a determination that these measures were moderate stakes tests (see Schmitt, 1996). Internal consistency was acceptable across constructs, methods, and conditions (Table 1). The alpha coefficients for the IO and AG scales ranged between .84 and .93, with small differences across constructs, methods, and conditions. However, the coefficients for the PC were notably lower for the factor-derived method \( (\alpha \text{ placebo} = .74, \alpha \text{ low dose} = .71) \) than for other methods \( (\text{range } = .80-.93) \). In addition, the internal consistency for PC was somewhat lower in the low-dose \( (\text{range } = .71-.85) \) compared to the placebo condition \( (\text{range } = .74-.93) \).

Temporal stability was calculated using data from the first two occasions (2 days apart) in the placebo and low-dose conditions, respectively. Less than 1% of data were missing in the placebo condition, and approximately 1% of data were missing in the low-dose condition. Missing data were replaced with the mean item score for that participant. Means,
standard deviations, and Pearson correlations are summarized in Table 2. To summarize estimates across methods and conditions, correlation coefficients were transformed into Fisher’s Z values. Next, Fisher’s Z values were averaged by method and by condition, and these averaged values were then transformed back to Pearson r values. The reliability of the IO scale (mean r = .67) was somewhat lower than both AG (mean r = .71) and PC (mean r = .78) across methods and conditions. Mean correlations differed little across factor-derived, individualized, and full methods (.72, .72, and .71, respectively). However, a slightly larger difference was found between the placebo (mean r = .75) and low-dose (mean r = .69) conditions.

### Concurrent Validity

Concurrent validity analyses were performed using data from the initial placebo assessment for the predictors (IO, AG, and PC) and clinic observation criterion variables. The remaining criterion variables (from the teacher-rated CSI-4) were collected at the intake assessment. Although there were no missing data for the predictor variables, approximately 6% of data were missing for criterion variables. These data were replaced with sample means by variable. A partial summary of correlations between predictor and criterion variables is provided in Table 3. To control the size of the correlation matrix, correlations not of central relevance to the current study were excluded. Using the Fisher Z transformation method described earlier to average correlation coefficients, only small differences were noted in validity coefficients across methods. Specifically, average correlations between IO and criterion variables (CSI-4 ADHD scales, and the three categories from the classroom analogue) were comparable across factor-derived (.31), individualized (.33), and full (.34) methods. The oppositional defiant disorder and conduct disorder scales of the CSI-4 served as criterion variables for both AG and PC. Similar to IO, average correlations between versions of AG and PC and criterion measures were comparable across factor-derived (.18 and .24, respectively), individualized (.18 and .27, respectively) and full (.15 and .26) methods.

IO demonstrated good convergent and divergent validity across methods, with validity coefficients (teacher ratings and observations of ADHD behaviors) reaching the level of statistical significance (p < .05) in the low to moderate range (r values between .24 and .41) and much lower correlations (range = .10–.18; ns), with teacher ratings of oppositional defiant disorder and conduct disorder. The AG and PC demonstrated low to moderate correlations with teacher ratings of oppositional defiant disorder and conduct disorder (range = .10–.18; ns). However, divergent validity was demonstrated only with teacher ratings of the inattentive symptoms of ADHD (range = .12–.17; ns) and clinic observations (range = .00–.18; ns). Correlations were notably higher for teacher ratings including the hyperactive-impulsive symptoms of ADHD (range = .27–.44; p < .05). Indeed, these correlations were comparable to those with oppositional defiant disorder. Teacher ratings of conduct disorder behaviors demonstrated low correlations across the IO, AG, and PC (range = .00–.15; ns).

---

**Table 1**

**Summary of Alpha Coefficients for Three Rating Scale Methods (N = 65)**

<table>
<thead>
<tr>
<th>Method</th>
<th>Dose Condition</th>
<th>Placebo</th>
<th>Low Dose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor derived</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inattention-Overactivity</td>
<td></td>
<td>.91</td>
<td>.87</td>
</tr>
<tr>
<td>Aggression</td>
<td></td>
<td>.88</td>
<td>.90</td>
</tr>
<tr>
<td>Peer Conflict</td>
<td></td>
<td>.74</td>
<td>.71</td>
</tr>
<tr>
<td>Individualized</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inattention-Overactivity</td>
<td></td>
<td>.88</td>
<td>.84</td>
</tr>
<tr>
<td>Aggression</td>
<td></td>
<td>.92</td>
<td>.89</td>
</tr>
<tr>
<td>Peer Conflict</td>
<td></td>
<td>.93</td>
<td>.80</td>
</tr>
<tr>
<td>Full</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inattention-Overactivity</td>
<td></td>
<td>.89</td>
<td>.87</td>
</tr>
<tr>
<td>Aggression</td>
<td></td>
<td>.92</td>
<td>.93</td>
</tr>
<tr>
<td>Peer Conflict</td>
<td></td>
<td>.89</td>
<td>.85</td>
</tr>
</tbody>
</table>
A 2 (Dose) × 3 (Method) within-subjects multivariate analysis of variance was performed to examine the treatment sensitivity of the three rating scale methods across dose conditions for the three constructs of interest (IO, AG, PC). Within dose condition, item scores were averaged for each participant across all assessment occasions. This involved averaging data for all four time points in the placebo and low-dose condition, respectively, for approximately 60% of participants. Data were available for at least three of four assessment occasions for 85% of the participants. For only one participant were data missing for three out of the four assessment occasions, and only for the low-dose condition. An alpha level of $p < .05$ was established for the omnibus analysis. Using Wilks’s lambda criterion for the omnibus test of within-subjects effects, main effects for dose, $F(62, 3) = 10.99, p < .001$, $\eta^2 = .35$, method, $F(59, 6) = 22.97, p < .001$, $\eta^2 = .70$, and the Dose × Method interaction, $F(59, 6) = 9.81, p < .001$, $\eta^2 = .50$, were statistically significant. Subsequently, a series 2 (Dose) × 2 (Method) analyses of variance were performed to examine pair-wise Dose × Method interactions for each construct. The number of contrasts was limited to minimize the chance of Type I error. As an additional precaution, a Bonferroni adjusted alpha of .006 was utilized as a criterion of significance for all post hoc analyses.

For IO, the Dose × Method interaction was significant only for the comparison of individualized and full methods, with the former demonstrating superior sensitivity, $F(64, 1) = 21.97, p < .001$, $\eta^2 = .26$. The analyses of variance comparing the factor-derived method to the individualized method, $F(64, 1) = 6.84$, $\eta^2 = .10$, and the full method, $F(64, 1) = 4.31$, $\eta^2 = .06$, were not statistically significant. For AG, there was a significant Dose × Method interaction comparing the individualized and factor-derived method, again favoring the individualized method $F(64, 1) = 12.24, p < .002$, $\eta^2 = .16$, but no significant interaction was found com-

### Table 2

Means (SD) and Test–Retest Coefficients for Three Rating Scale Methods ($N = 65$)

<table>
<thead>
<tr>
<th>Method</th>
<th>Placebo Time 1</th>
<th>Placebo Time 2</th>
<th>Low Dose Time 1</th>
<th>Low Dose Time 2</th>
<th>r Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor derived</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inattention-Overactivity</td>
<td>1.65</td>
<td>.97</td>
<td>1.75</td>
<td>.97</td>
<td>.60</td>
<td>.88</td>
<td>1.24</td>
<td>.88</td>
<td>1.10</td>
<td>.87</td>
<td>.68</td>
</tr>
<tr>
<td>Aggression</td>
<td>.55</td>
<td>.83</td>
<td>.64</td>
<td>.81</td>
<td>.78</td>
<td>.65</td>
<td>.36</td>
<td>.65</td>
<td>.35</td>
<td>.64</td>
<td>.68</td>
</tr>
<tr>
<td>Peer Conflict</td>
<td>.47</td>
<td>.72</td>
<td>.54</td>
<td>.72</td>
<td>.81</td>
<td>.49</td>
<td>.24</td>
<td>.49</td>
<td>.19</td>
<td>.42</td>
<td>.71</td>
</tr>
<tr>
<td>Individualized</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inattention-Overactivity</td>
<td>1.71</td>
<td>.96</td>
<td>1.70</td>
<td>.98</td>
<td>.65</td>
<td>.85</td>
<td>1.16</td>
<td>.85</td>
<td>1.03</td>
<td>.86</td>
<td>.69</td>
</tr>
<tr>
<td>Aggression</td>
<td>.77</td>
<td>.91</td>
<td>.71</td>
<td>.86</td>
<td>.72</td>
<td>.67</td>
<td>.43</td>
<td>.67</td>
<td>.35</td>
<td>.61</td>
<td>.66</td>
</tr>
<tr>
<td>Peer Conflict</td>
<td>.57</td>
<td>.83</td>
<td>.58</td>
<td>.81</td>
<td>.85</td>
<td>.49</td>
<td>.24</td>
<td>.49</td>
<td>.18</td>
<td>.43</td>
<td>.73</td>
</tr>
<tr>
<td>Full</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inattention-Overactivity</td>
<td>1.40</td>
<td>.88</td>
<td>1.49</td>
<td>.86</td>
<td>.66</td>
<td>.76</td>
<td>1.02</td>
<td>.76</td>
<td>.93</td>
<td>.79</td>
<td>.71</td>
</tr>
<tr>
<td>Aggression</td>
<td>.61</td>
<td>.81</td>
<td>.67</td>
<td>.80</td>
<td>.76</td>
<td>.64</td>
<td>.37</td>
<td>.64</td>
<td>.31</td>
<td>.58</td>
<td>.67</td>
</tr>
<tr>
<td>Peer Conflict</td>
<td>.29</td>
<td>.45</td>
<td>.38</td>
<td>.52</td>
<td>.81</td>
<td>.29</td>
<td>.16</td>
<td>.29</td>
<td>.11</td>
<td>.24</td>
<td>.64</td>
</tr>
</tbody>
</table>

Note: All correlations were significant, $p < .001$. 

358
paring the full method to either the individualized, $F(64, 1) = 7.24, \eta^2 = .10$, or factor-derived method, $F(64, 1) = 6.47, \eta^2 = .09$. However, for PC, the interaction was significant for comparisons between the full method and both the individualized, $F(64, 1) = 16.81, p < .001, \eta^2 = .21$, and the factor-derived method, with the full method being the least sensitive, $F(64, 1) = 9.55, p < .004, \eta^2 = .13$.

### Discussion

The purpose of this study was to expand our initial investigation of the feasibility of two different methods of creating abbreviated behavior rating scale measures for possible consideration in a CPMS by exploring these strategies in a randomized clinical trial. In addition to examining reliability and treatment sensitivity of the original and abbreviated scales, we investigated the concurrent validity of these scales using a battery of external validation measures, including full-length rating scales measuring the same or similar constructs, and systematic direct observation. Based on our prior work in this area (Volpe, Gadow et al., 2009), we hypothesized that the original scales and the abbreviated rating scale would demonstrate adequate psychometric properties, and if any differences in technical characteristics were found across methods, they would be minor.

Internal consistency was found to be acceptable across methods (factor-derived, individualized, and full) and conditions (placebo and low dose) and was comparable across methods with one exception. For PC, internal consistency of the factor-derived scale was lower (.74, .71) when compared to both the individualized (.93, .80) and full (.89, .85) methods. In our initial study, we found a similar result for the Inattentive scale of the ADHD-Symptom Checklist—4; the internal consistency of the factor-derived scale (.74) was notably lower in comparison to the individualized (.93) or full (.87) scales (Volpe, Gadow et al., 2009). Lower estimates of internal consistency for the factor-derived method are difficult to interpret because they are not repeated across constructs. One possible explanation is that the factor analysis of the PC was conducted with normative sample data and obtained results do not generalize well to

### Table 3
Summary of Correlations Between Scales of Interest and External Validation Measures ($N = 65$)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Factor-Derived Scales</th>
<th>Individualized Scales</th>
<th>Full Scales</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IO</td>
<td>AG</td>
<td>PC</td>
</tr>
<tr>
<td>CSI-4 Scales</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADHD-I</td>
<td>.29*</td>
<td>.12</td>
<td>.16</td>
</tr>
<tr>
<td>ADHD-HI</td>
<td>.27*</td>
<td>.30*</td>
<td>.42**</td>
</tr>
<tr>
<td>ADHD-C</td>
<td>.34*</td>
<td>.27*</td>
<td>.38**</td>
</tr>
<tr>
<td>ODD</td>
<td>.10</td>
<td>.27*</td>
<td>.36**</td>
</tr>
<tr>
<td>CD</td>
<td>.11</td>
<td>-.10</td>
<td>.13</td>
</tr>
<tr>
<td>Classroom Analog</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>On Task</td>
<td>-.41**</td>
<td>-.11</td>
<td>-.05</td>
</tr>
<tr>
<td>Out of Seat</td>
<td>.29*</td>
<td>.02</td>
<td>.02</td>
</tr>
</tbody>
</table>

Note. IO = Inattention-Overactivity; AG = Aggression; PC = Peer Conflict; CSI-4 = Child Symptom Inventory—4; ADHD-I = Attention Deficit Hyperactivity Disorder—Inattentive; ADHD-HI = ADHD—Hyperactive-Impulsive; ADHD-C = ADHD—Combined; ODD = oppositional defiant disorder; CD = conduct disorder.

* $p < .05$.
** $p < .01$. 

In comparing the full method to either the individualized, $F(64, 1) = 7.24, \eta^2 = .10$, or factor-derived method, $F(64, 1) = 6.47, \eta^2 = .09$. However, for PC, the interaction was significant for comparisons between the full method and both the individualized, $F(64, 1) = 16.81, p < .001, \eta^2 = .21$, and the factor-derived method, with the full method being the least sensitive, $F(64, 1) = 9.55, p < .004, \eta^2 = .13$. The purpose of this study was to expand our initial investigation of the feasibility of two different methods of creating abbreviated behavior rating scale measures for possible consideration in a CPMS by exploring these strategies in a randomized clinical trial. In addition to examining reliability and treatment sensitivity of the original and abbreviated scales, we investigated the concurrent validity of these scales using a battery of external validation measures, including full-length rating scales measuring the same or similar constructs, and systematic direct observation. Based on our prior work in this area (Volpe, Gadow et al., 2009), we hypothesized that the original scales and the abbreviated rating scale would demonstrate adequate psychometric properties, and if any differences in technical characteristics were found across methods, they would be minor.

Internal consistency was found to be acceptable across methods (factor-derived, individualized, and full) and conditions (placebo and low dose) and was comparable across methods with one exception. For PC, internal consistency of the factor-derived scale was lower (.74, .71) when compared to both the individualized (.93, .80) and full (.89, .85) methods. In our initial study, we found a similar result for the Inattentive scale of the ADHD-Symptom Checklist—4; the internal consistency of the factor-derived scale (.74) was notably lower in comparison to the individualized (.93) or full (.87) scales (Volpe, Gadow et al., 2009). Lower estimates of internal consistency for the factor-derived method are difficult to interpret because they are not repeated across constructs. One possible explanation is that the factor analysis of the PC was conducted with normative sample data and obtained results do not generalize well to
a sample of clinic-referred children with both ADHD and either chronic motor tic disorder or Tourette’s disorder. Nevertheless, although estimates for factor-derived PC are relatively low, they still met our criteria for adequate internal consistency ($\alpha \geq .70$).

In general, the reliability of scales measuring IO was somewhat lower than those measuring both AG and PC. However, there was essentially no difference in estimates of temporal stability across methods. This is consistent with our previous study in which differences in temporal stability across dose conditions were larger than differences across methods, although this difference was not statistically significant. In the current study, differences in temporal stability between placebo and low-dose conditions were relatively small compared to our previous study.

The finding that reliability was comparable between the original and the abbreviated scales is perhaps surprising when one considers the Spearman-Brown prophecy; namely, the reliability of tests should generally increase with the number of items or scales measuring the construct of interest. However, the Spearman-Brown prophecy assumes that items on a test are parallel. It could be argued that both methods used in this study to extract items from extant rating scales involve selecting higher quality items than the ones that were rejected. This assertion might be most defensible for the individualized method, wherein only the most elevated items for individual children are selected for progress monitoring. It is important to note that the IOWA Conners and the Peer Conflict Scale were designed to be generally applicable to samples of children demonstrating disruptive behavior. Considering that rarely do two children present with the same combination of symptoms, the relevance of any given item on either measure will vary from child to child. Although a fixed number of items is necessary for diagnostic assessment, the heterogeneity of children provides an incentive to customize progress monitoring measures for individual children.

We hypothesized that the original and abbreviated measures would demonstrate acceptable concurrent validity and that there would be small if any differences in validity coefficients across rating scale methods. Evidence of concurrent validity was mixed. The IO scales demonstrated both convergent validity (high ratings with relevant measures of ADHD behaviors) and divergent validity (lower ratings with measures of constructs other than ADHD). However, although the AG and PC scales demonstrated low to moderate correlations with teacher ratings of oppositional defiant disorder (range $=.27-.40; p < .05$), divergent validity was only demonstrated with teacher ratings of the inattentive symptoms of ADHD (range $=.12-.17; \text{ns}$) and clinic observations (range $=.00-.18; \text{ns}$). There was poor differentiation between ADHD and AG and PC when ratings of ADHD included the hyperactive-impulsive symptoms of ADHD (range $=.27-.45; p < .05$). These ratings were comparable to validity coefficients examining the association between AG, PC, and oppositional defiant disorder. Of interest, teacher ratings of conduct disorder behaviors demonstrated low correlations with all ratings of AG and PC (range $.00-.15; \text{ns}$). Consistent with our hypothesis, the averaged validity coefficients across methods were comparable, suggesting that little was lost in terms of concurrent validity when the length of rating scales was reduced using the factor-derived and individualized method.

A critical question in the present study was the degree to which the abbreviated rating scales were sensitive to the effects of treatment. In our previous study it appeared that for teacher ratings of the inattentive and hyperactive-impulsive symptoms of ADHD, the individualized method was more sensitive to treatment effects than either the factor-derived or full method. However, the individualized scales in our prior study were constructed using data from the first baseline assessment occasion; therefore, one possible explanation for this finding is that high initial scores at baseline regressed toward the mean in the next assessment, thus inflating the estimate of treatment sensitivity for the individualized method. In the current study, individualized scales were created using data from the first assess-
ment in the placebo condition, and the ordering of dose conditions was counterbalanced, therefore controlling for this potential confound. For all three constructs, the difference between scores in the placebo and low-dose condition was the greatest for the individualized method. However, significant Dose × Method interactions were found only for comparisons between the individualized and full methods for IO and PC, and between the individual and factor-derived method for AG. For PC, there also was a significant difference between the factor-derived and full methods, with the former method demonstrating superior treatment sensitivity.

Limitations and Future Directions

The present study was conducted with tight experimental controls in regard to the evaluation of the effects of stimulant medication on constructs of interest. Nevertheless, our results are limited to some extent by the fact that although we examined three rating scale methods (factor derived, individualized, and full), all teachers were administered the full versions of the IO and AG scales and the Peer Conflict Scale. Although it seems unlikely that teachers would be less accurate and consistent in rating shorter rating scales, future studies should administer the three types of rating scales separately in counterbalanced order to examine the different methods. In addition, administering both the full and various abbreviated versions would permit future investigators to evaluate the acceptability of each approach. Based on the intervention acceptability literature (e.g., Eckert & Hintze, 2000), we would expect that the informants would prefer the abbreviated over the longer scales; however, this remains a gap in the extant literature.

The findings of literally hundreds of studies have shown that IR-MPH has powerful effects on the symptoms of ADHD and childhood aggression (for a review, see Connor, 2006). Even though we compared behavior during the placebo and low-dose condition (as opposed to the higher dose condition), the sensitivity to change found for these measures may not generalize to changes resulting from other types of interventions, particularly those associated with smaller effect sizes or those targeting a narrow set of target behaviors. Although the difference in the length of the 10-item Peer Conflict Scale and PC scales was substantial, this was somewhat less the case for the IOWA Conners and the abbreviated IO and AG scales. As mentioned earlier, gains in reliability follow the law of diminishing returns, and it is not possible from the current data to determine where gains in reliability from additional items start to decelerate. To address this question, one would want to apply generalizability theory (see Brennan, 2003). Nevertheless, the results of this study suggest that 3-item scales generated from the factor-derived and individualized methods may be sufficient to obtain acceptable technical characteristics. Moreover, additional items may be superfluous for progress monitoring purposes.

Finally, we mentioned earlier that a CPMS should measure both broad domains of emotional and behavior disorders and related areas of impairment as well as specific behaviors targeted for intervention. In this study we addressed a relatively small set of broad indicators of disruptive behavior. Future studies are needed to apply this methodology to the assessment of a larger array of constructs, including internalizing problems and areas of functional impairment. The myriad of problems demonstrated by the participants of the present study (all met criteria for ADHD and either chronic motor tic disorder or Tourette’s disorder, and over half met criteria for oppositional defiant disorder) clearly indicate that children present with more than one problem. Creating short progress monitoring rating scales with acceptable psychometric properties enables the assessment of multiple constructs without overburdening the informants we rely on to provide valuable data concerning response to intervention.

References


Robert J. Volpe is Assistant Professor of Counseling and Applied Educational Psychology at Northeastern University. His primary research interests concern academic problems experienced by children with ADHD, behavioral assessment, and academic interventions. He received his doctorate in school psychology from Lehigh University in 2003.

Kenneth D. Gadow is Professor of Psychiatry at Stony Brook University School of Medicine at Stony Brook, New York. His current research interests include ADHD, Tourette syndrome, development and evaluation of assessments instruments, nosology of psychopathology, perinatal HIV infection, and candidate genes for phenotypic variation in autism spectrum disorder.