Behaviorally anchored rating scale aba

What is a 360 assessment? What is graphic rating scale? What is psychological appraisal? What is MBO in performance appraisal? What is a forced distribution performance appraisal? What is performance appraisal system? What is behavior observation scale? What is the bars rating scale used for ABA? What are behavioral anchors? What is mixed standard scale? What is forced choice method of performance appraisal? When should an informal performance appraisal be conducted? What is forced choice method? In order to continue enjoying our site, we ask that you confirm your identity as a human. Thank you very much for your cooperation. Behaviourally Anchored Rating Scales (BARS) definition Behaviourally Anchored Rating Scales (BARS) are designed to bring the benefits of both qualitative and quantitative data to the employee appraisal process. BARS compare an individual's performance against specific examples of behaviour that are anchored to numerical ratings. For example, a level four rating for a nurse may require them to show sympathy to patients while a level six rating may require them to show higher levels of empathy and ensure this comes across in all dealings with the patient. The behavioural examples used as anchor points are often collected using Critical Incident Techniques (CIT), which are procedures used for documenting human behaviour that have significance in a particular area. BARS was originally developed to counteract the perceived subjectivity in using basic ratings scales to judge performance, although BARS is still subject to criticism. It is often accused of being subject to unreliability and leniency error. Although the efficiency with which a wide range of behavioral data can be obtained makes behavior rating scales particularly attractive tools for the purposes of screening and evaluation, feasibility concerns arise in the context of formative assessment. Specifically, informant load, or the amount of time informants are asked to contribute to the assessment process, likely has a negative impact on the quality of data over time and the informant's willingness to participate. Two important determinants of informant load in progress monitoring are the length of the rating scale (i.e., the number of items) and how frequently informants are asked to provide ratings (i.e., the number of occasions). The purpose of the current study was to investigate the dependability of the IOWA Conners Teacher Rating Scale (Loney & Milich, 1982), which is used to differentiate inattentive-overactive from oppositional-defiant behaviors. Specifically, the facets of items and occasions were examined to identify combinations of these sources of error necessary to reach an acceptable level of dependability for both absolute and relative decisions. Results from D studies elucidated a variety of possible item–occasion combinations reaching the criteria for adequate dependability. Recommendations for research and practice are discussed.Behavior rating scales are among the most common assessment methods used by school psychologists, with over 75% of school psychologists reporting inclusion of either parent or teacher scales in the majority of recent referral cases (Shapiro & Heick, 2004). These measures generally assess a broad spectrum of constructs relating to social behaviors (e.g., aggressive behavior and anxiety symptoms) and typically demonstrate sound psychometric properties. Because they can be used to sample behavior over a long period of time, they afford the advantage of measuring low-frequency behaviors that might not be captured by other assessment methods such as systematic direct observation (Merrell, 2008). In addition, because rating scales easily can be completed by several informants, they offer a highly efficient means to obtain information about child behavior from multiple settings and perspectives (see Achenbach, McConaughy, & Howell, 1987). Furthermore, in comparison to other behavior assessment tools such as systematic direct observation, rating scales are fairly cost effective (e.g., they require only modest training for informants and take relatively little time to complete).Although the overwhelming majority of rating scales were developed primarily as screening and diagnostic measures, they have long been used to evaluate the summative effects of interventions targeting emotional and behavior problems (DuPaul & Stoner, 2003). However, when applied to progress monitoring, feasibility concerns emerge (Briesch & Volpe, 2007; Volpe & Gadow, in press; Volpe, Gadow, Blom-Hoffman, & Feinberg, 2009). Specifically, due to the fact that typical broad-band rating scales often contain over 100 items, they are not typically considered for use as progress monitoring tools. This pattern of use is due largely to informant load (i.e., the amount of time informants are asked to contribute to the assessment process), which likely has a negative impact on the informant's acceptability of assessment procedures and their willingness to participate in the assessment process (Elliott, Witt, Galvin, & Peterson, 1984). Alternatively, several rating scales, including the ADHD Symptom Checklist (Gadow & Sprafkin, 2008) and the BASC Monitor for ADHD (Reynolds & Kamphaus, 2004), have been developed specifically to measure behavior in a formative fashion. Although these shorter rating scales offer the advantage of fewer overall items, their utility for frequent assessment is still somewhat limited given the fact that they often contain over 40 items. One exception is the IOWA Conners Teacher Rating Scale (Loney & Milich, 1982), which was evaluated in the present study. The IOWA Conners was developed to discriminate inattentive-overactive and oppositional-defiant symptoms and is composed of only 10 items. Although it was designed as an efficient measure for diagnostic purposes, because of its brevity, technical characteristics, and relevance in assessing the associated symptoms of inattention-overactivity and oppositional-defiant behavior, it has been used widely in evaluating the effects of intervention for students with attention-deficit/hyperactivity disorder (ADHD) and related disorders. Several studies have documented that the IOWA Conners is sensitive to the effects of both stimulant medication and psychosocial interventions (see Pelham, Fabiano, & Massetti, 2005). Moreover, the IOWA Conners frequently has been used to aid in the titration of stimulant medication and to monitor treatment effects. Although weekly teacher ratings are common in published medication titration procedures, the IOWA Conners has been administered daily in school analog settings (e.g., Fabiano et al., 2007).A problem-solving approach to treating child emotional and behavior problems involves frequent monitoring of student functioning in response to intervention in order to assess progress toward meaningful and socially valid goals. In the realm of academic assessment, general outcome measures are available to assess growth in the basic academic skill areas of writing, reading, and mathematics (e.g., Shinn, 2008). However, no such general outcome measure has emerged for assessing student social-emotional functioning (see Chafouleas, Volpe, Gresham, & Cook, in press).Evaluating the effects of psychosocial interventions to address student emotional and behavioral functioning is not new to school-based professionals. School-wide positive behavior support (SWPBS; Sugai & Horner, 2002) is one example wherein emphasis has been placed upon assessment of both individual- and system-level change in response to tiers of intervention. Within SWPBS systems, office discipline referrals (ODRs) are typically used as an indicator of individual student risk, with students receiving two or more ODRs identified as being in need of additional supports (Walker, Cheney, Stage, & Blum, 2005). However, these outcome measures are also often used to assess changes in overall school climate as a function of SWPBS implementation. Unfortunately, little attention has been directed toward the validation of these assessment measures (McIntosh, Spaulding, & Frank, in press), purportedly because the instruments have a high degree of face validity. The use of rating scales in SWPBS has typically been limited to multiple-gated screening processes, wherein ratings are completed for those students ranked highly for emotional or behavior problems (e.g., Walker & Severson, 1990), though arguably ODRs are more frequently used to identify students in need of intervention.Another example of evaluation of behavioral functioning in school settings can be found in the literature pertaining to school-based procedures for the titration of stimulant medication in the treatment of ADHD wherein evaluators conduct brief controlled trials to compare student response to several doses of medication and a no treatment baseline or placebo condition. Typically, such procedures have involved assessing several domains of functioning including psychiatric symptoms, academic and social functioning, and stimulant side effects (DuPaul & Stoner, 2003); however, there has been little consistency in assessment methods across studies. Much of the published research concerning school-based medication evaluation has been conducted in the context of federally-funded research studies utilizing resources from outside of the school setting to orchestrate the assessment procedures (e.g., Gadow, Nolan, Paolicelli, & Sprafkin, 1991) and has often involved features such as intensive data-collection procedures (e.g., systematic direct observation and long rating scales) and experimental controls (e.g., double-blinding and use of placebos) that often are not feasible in typical school settings in the absence of external funds and other resources.Several authors, however, have proposed methods and procedures that reduce the demands on informants and evaluators. For example, Volpe, Heick, and Guerasko-Moore (2005) have placed school-based medication evaluation into a flexible behavioral consultation framework wherein assessment procedures are adapted based on the resources available in each assessment situation. Briefly, the process involves the use of rating scales that describe various assessment components and the surveying of key stakeholders to assess the acceptability and feasibility of each assessment component. These data are then used to develop an assessment procedure that is embedded in a behavioral consultation framework which takes into consideration these assessment preferences. Pelham (1993) has promoted the use of flexible and user-friendly daily report cards to obtain teacher ratings of student progress on a small group of target behaviors. These instruments do not require trained observers and take relatively little time to complete. Although the use of these daily report cards as an intervention for classroom behavior problems shows much promise (e.g., Fabiano et al., 2007; Fabiano et al., in press), little is known about their measurement characteristics, and as such, their potential as progress-monitoring measures remains unknown.One general approach that has experienced increased adoption in recent years is the extraction of a subset of items from a larger established rating scale. Hyman and colleagues (1998) suggested selecting items from existing rating scales to create shorter progress-monitoring measures wherein only the most relevant items would be rated by informants (e.g., by highlighting selected items on a existing rating scale once teachers have selected those most relevant for a specific child). One potential disadvantage of this approach, however, is that once items are removed from existing scales little is known about their psychometric properties. An alternative method utilized by several authors involves selecting items with the highest factor loadings from existing rating scales (e.g., Reynolds & Kamphaus, 2004), which results in scales containing items that are most highly intercorrelated and those that should be most strongly associated with the construct of interest. Another suggested option is to select items based on their relative sensitivity to intervention using complex decision rules involving multiple statistical methods (e.g., Gresham et al., in press; Meier, McDougal, & Bardos, 2008).The resultant measures can be referred to as "brief rating scales" because they are shorter than the original measures. However, these represent nomothetic approaches to shortening rating scales, as data from large samples are used to inform the deletion of items. Several authors have instead advocated for individualized methods of item selection wherein teachers select from a menu of items that are most relevant for individual students (Hyman et al., 1998) or items are selected based on their rated severity (Volpe et al., 2009; Volpe & Gadow, in press). A more descriptive term for these brief rating measures would be "customized rating scales" given that the composition of each scale is tailored to each student.The psychometric properties of rating scales have been overwhelmingly examined through the application of classical test theory. The most common methods of examining reliability involve evaluating the consistency of measurement over repeated assessments by the same informant (i.e., test–retest reliability), across different informants (i.e., inter-rater reliability), and across items (i.e., internal consistency reliability). In each case, the underlying assumption is that the behavior of interest is a stable, enduring trait and therefore any variability observed is due to measurement error (Cone, 1978). Each classical test theory approach to reliability involves partitioning true variation in scores from error, wherein all error variance is considered random or unexplained. If one were to measure the phenomena of interest an infinite number of times, classical test theory suggests that random, uncorrelated errors would cancel each other out, with the obtained average measurement thus representing the true score.Although such an approach is well-suited to the measurement of traits that are expected to remain stable over time, the relevance of traditional classical test theory approaches to the assessment of state behaviors has been questioned (Silva, 1993). Generalizability (G) theory however, offers an alternative to the classical test theory approach that not only acknowledges, but accounts for, differences in measurement conditions by suggesting that each occasion sampled is exchangeable with a universe of other possible measurements (Cronbach, Gleser, Rajaratnam, & Nanda 1972). The focus in using G theory is therefore not on the consistency of measurement but on how accurately one can generalize from a specific sample of behavior to all possible samples of interest. Rather than combining multiple sources of measurement variance under a general residual error term, G theory enables the partitioning of error variance by its source (Brennan, 2001; Shavelson & Webb, 1991), thereby informing decisions regarding how various sources of error can be minimized in future assessment activities.One of the greatest strengths of a generalizability approach is that the variance components obtained in a generalizability (G) study can then be used to inform decision (D) studies. Analogous to the Spearman Brown prophecy formula, D studies allow one to estimate how reliability coefficients would improve if different aspects of measurement were altered. Given that it is possible to determine optimal measurement procedures without extensive direct testing, G theory has been used within the school psychology literature to examine both academic and behavioral assessment tools. In evaluating different forms of curriculum-based measurement, for example, D studies have been employed to assess how dependability changes as a function of the length of probes (e.g., Christ, Johnson-Gros, & Hintze, 2005) or the number of probes (e.g., Hintze, Christ, & Keller, 2002; Poncy, Skinner, & Axtell, 2005) administered. Although several studies have recently applied G theory to the examination of rating scales (Bergeron, Floyd, McCormack, & Farmer, 2008) and similar behavioral assessment methods (e.g., Chafouleas, Briesch, Riley-Tillman, Christ, Black, & Kilgus, 2010; Volpe, McConaughy, & Hintze, 2009), the predominant focus across D studies has been on manipulating the number of behavior samples (e.g., observations and ratings) needed to reach a criterion for dependability.Implementing a large-scale, problem-solving model related to emotional and behavior problems requires careful consideration of the efficiency of intervention and assessment methods. In other words, it is unlikely that inefficient models will be sustainable in typical school settings. In the case of intervention, efficiency has been operationalized by dividing each unit of change in the desired direction (e.g., increased minutes engaging in appropriate social behavior; words read correctly) by the number of instructional minutes (see Cates et al., 2003; Nist & Joseph, 2008). In this manner, an intervention requiring the fewest number of instructional minutes per unit of change is considered to be most efficient. The concept of efficiency also has been applied to academic assessment (Griffiths, VanDerHeyden, Skokut, & Lilles, 2009) and behavioral assessment (Volpe et al., 2009; Volpe & Gadow, in press), wherein the psychometric properties of more time-intensive assessment approaches have been compared to less time-intensive approaches. Here, if a less time-intensive method reaches the threshold for reliability or validity, it can be chosen over the more time-intensive method.G theory offers a useful tool for operationalizing the efficiency of assessment methods (see Parkes, 2000). Specifically, an investigator can select an acceptable criterion for dependability a priori and then conduct D studies to examine different combinations of error that can be modeled to reach the criterion threshold. In D studies two reliability-like coefficients are calculated. The generalizability coefficient (ρ2) is the relevant index for making relative decisions, in that it considers variance that has only to do with the rank ordering of persons. The dependability coefficient (φ) is used for absolute decisions in that it does not consider variance across persons, but rather only accounts for variance components due to inconsistencies in measurement conditions (Brennan, 2001). Typically, D studies are conducted one-at-a-time involving manipulation of a single facet (e.g., occasions), and a point is reported wherein the criterion is reached (e.g., 10 observations are necessary to reach a dependability coefficient ≥ .80). Cone (1978), however, proposed six sources of error that can be conceptualized in G theory as universes of generalization and can be manipulated simultaneously: (a) scorer (e.g., rater), (b) item, (c) time, (d) method, (e) setting, and (f) dimension. As a result, it is possible that several different combinations of these sources of error (e.g., more raters and less time as well as fewer raters and more time) can be used to reach the desired threshold of dependability.Generating several different assessment strategies with acceptable levels of dependability has several advantages. First, if one were to quantify the cost of each source of error (e.g., cost of having additional informants provide ratings or time needed to complete each rating scale item), then the efficiency of each method can be quantified, and models of equal dependability can be ordered hierarchically in terms of cost of administration. Second, informants can be offered a menu of assessment options (e.g., X number of items and Y number of times) rather than a single mandate and can select the combination that is most acceptable to them given their unique preference (cf. Volpe et al., 2005). Finally, both the purpose of assessment and unique assessment situation can help to inform selection. For example, it is likely that in many cases the number of items in a scale is inversely related to the number of times the scale must be administered in order to obtain an acceptable level of reliability. When decisions must be made rapidly, a longer scale can be selected.The purpose of the current study was to explore how G theory may be used to maximize data collection procedures by examining rating scale data from the IOWA Conners (Loney & Milich, 1982), which is used to differentiate inattention-overactivity from oppositional-defiant behavior. The facets of item and time (i.e. occasions) were manipulated to identify combinations of these sources of error necessary to reach an acceptable level of dependability for both absolute and relative decisions. G theory studies involving the assessment of child classroom behavior have examined non-referred children under typical classroom conditions with few exceptions (Volpe et al., 2009). This pattern in the research may be problematic because school psychologists typically observe children who are referred for emotional and behavior problems, and the behavior of referred children tends to be more variable than typically developing children. In addition, the introduction of an intervention, in addition to having an effect on the level and trend of data, often impacts the variability in scores. All of these factors can lead to differences in reliability across phases (e.g., baseline versus intervention). In order to maximize the degree to which psychometric evidence can be generalized to the authentic applications of progress monitoring measures, it is important to examine the psychometric properties of these measures in referred populations and under both baseline and treatment conditions. Children and adolescents in the current study were clinic referrals participating in a randomized, placebo-controlled, crossover trial of immediate release methylphenidate (IR-MPH). G and D studies were conducted both for a placebo and IR-MPH condition.Participants consisted of 71 children (57 boys, 14 girls) between 6 and 13 years old (M = 8.9; SD =1.9) who were recruited for participation in a larger study (Gadow, Nolan, Sverd, Sprafkin, & Schneider, 2008; Gadow, Sverd, Nolan, Sprafkin, & Schneider, 2007). Students were recruited from a variety of sources (i.e., clinics, schools, media advertisements, and parent support groups) for participation 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 notified on several occasions of the possible risk of irreversible tic exacerbation caused by IR-MPH treatment (although the results of the controlled trial indicated that this was not the case for the sample as a whole). Parents and children (≥ 11 years) provided written informed consent and assent, respectively.A total of four students were excluded from the current study because the overwhelming majority of their data on the dependent variables were missing. The remaining sample consisted of 67 children (54 boys, 13 girls) between 5 and 13 years-old (M = 8.9; SD = 1.8). Participants consisted of two cohorts, with all data being collected from 1989 to 2004. The majority of students were Caucasian (91%), with the remainder being either Black (4.5%) or Hispanic (4.5%).Inclusion criteria To participate in the study each child had to meet Diagnostic and Statistical Manual of Mental Disorders, Third Edition, Revised (DSM-III-R; American Psychiatric Association, 1987) or Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV; American Psychiatric Association, 2004) diagnostic criteria for ADHD and either chronic motor tic disorder or Tourette's disorder. Each child met ADHD clinical criteria in both school and home. Almost all children met research diagnostic criteria (Kurlan, 1989) for Tourette syndrome, either definite or by history.Co-morbidities Co-morbid disorders were assessed with the parent interview version of the Diagnostic Interview for Children and Adolescents (DICA; Reich, 2000). Approximately one-half of participants met criteria on the DICA for oppositional defiant disorder, and one quarter met criteria for an anxiety disorder, sometimes in conjunction with a depressive disorder.Exclusion criteria Children who met one or more of the following criteria were excluded from the study: (a) their tics were the major clinical management concern; (b) they were too severely ill (i.e., dangerous to self or others), psychotic, or mentally retarded (IQ < 70); or (c) they had a seizure disorder, major organic brain dysfunction, major medical illness, medical or other contraindication to medication (other than tics), or pervasive developmental disorder.The IOWA Conners Teacher's Rating Scale (Loney & Milich, 1982) contains two, five-item scales: Inattention-Overactivity (IO) and Oppositional-Defiant (OD). The five-item IO scale contains items relating to both the inattentive and hyperactive-impulsive symptoms of ADHD, whereas the OD scale contains items related to oppositional defiant disorder. The 10 items that compose the IOWA Conners were originally drawn from the 39-item version of the Conners Rating Scale that was incorporated into the NIMH Early Clinical Drug Evaluation Unit (see Loney & Milich, 1982). Individual items are rated on a four-point scale (i.e., 0 = not at all, 1 = just a little, 2 = pretty much, 3 = very much). The IO and OD scales were shown to be moderately correlated with corresponding chart ratings of hyperactivity and aggression, demonstrate adequate internal consistency (r = .80 to .92) and 1-week test–retest reliability (rs = .86 to .89), and differentiate between non-referred children and children with ADHD with and without co-morbid oppositional behavior (e.g., Atkins, Pelham, & Licht, 1989; Loney & Milich, 1982; Milich & Landau, 1988; Pelham, Milich, Murphy, & Murphy, 1989; Waschbusch & Willoughby, 2008). A direct observation study conducted in a public school setting indicated that IO scores correlated with off-task behavior (r = .46), and the OD scores correlated with noncompliance (r = .60) and interference (r = .43; Nolan & Gadow, 1994).Participants received placebo and three doses of IR-MPH (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. In this study we examine data only from the placebo and 0.3 mg/kg conditions. All children received Novartis Brand Ritalin-IR. Dose schedules were counter-balanced and assigned on a random basis. Medication was administered twice daily, approximately 3.5 hours apart, 7 days a week, and dispensed in dated, sealed envelopes at 2-week intervals. Detailed descriptions of the procedure and all measures were previously published (Gadow et al., 1995, 2007).Teachers completed a battery of rating scales for each child's behavior, two days per week for the duration of the drug evaluation. Hence, a total of four ratings (two ratings each week for two weeks) were available for each student in each dose condition. Teachers knew all students for at least two months before completing these ratings. Approximately 28 percent (19 of 67) of students were receiving full-time special education services. For these students, ratings were completed by a special education teacher. Another 33 percent (22 of 67) of students received some kind of special education service, but for these students ratings were completed by regular education teachers. Each teacher completed ratings for only one student. Data were collected on a total of four occasions per medication condition (two ratings each week for two consecutive weeks). For the majority of participants, data were available for all four assessment occasions within each condition (placebo and IR-MPH). In cases where data were missing, available data were entered for the first three assessment occasions, with the fourth assessment occasion being left blank. This method resulted in a disproportionate amount of missing data for the fourth assessment occasion in each condition. Child behavior was rated during periods of maximum drug efficacy.Data Analysis Variance component analyses were conducted using GENOVA, an ANSI C computer program capable of computing variance components for a variety of complete, balanced ANOVA designs (Crick & Brennan, 1983). Within each scale (i.e., IO and OD), four sources of variance (i.e., facets) were identified and both the facets and all possible interactions were investigated within the model. Person (p) served as the target of measurement in the current study and refers to the individual student. Multiple items (i) were used to assess each student's behavior across time. Given that four rating occasions occurred within each medication condition, the facet of occasion (o) was treated as nested within medication condition (c; placebo/IR-MPH). This treatment resulted in a partially-nested p x i x (o:c) design. Because the goal of the study was to generalize beyond the specific students and rating occasions sampled, the facets of persons and occasions were treated as random.Of a total of 2,680 possible data points (5 items × 4 occasions × 2 conditions × 67 students), 886 (33%) were identified as missing. Because the reasons for missing data were limited to teacher or student absences, data were considered to be missing at random (Enders, 2001) and, therefore, use of multiple imputation procedures was considered appropriate. Through use of multiple imputation, observed values are used to predict several (m > 1) plausible missing values. Identical analyses are then conducted on each (m > 1) of the generated datasets, and results are combined to produce the final estimates (Schafer & Olsen, 1998). All variables and interactions targeted in generalizability analyses were included in the imputation model, and a total of five datasets were generated for the purpose of the current study. The mean rating obtained on the IO scale was 1.36 (SD = 1.05) in the placebo condition and .96 (SD = .93) in the IR-MPH condition. The mean rating obtained on the OD scale was .65 (SD = .93) in the placebo condition and .35 (SD = .70) in the IR-MPH condition. Mean ratings and standard deviations are presented for each item in Table 1.Mean Item Ratings and Standard Deviations across Scales and Conditions (N = 67)ConditionPlaceboIR-MPHMSDMSDInattention-Overactivity Item 11.311.120.830.93 Item 21.361.051.101.03 Item 31.601.021.100.98 Item 41.701.021.211.01 Item 50.831.020.580.72Oppositional-Defiant Item 10.580.940.340.71 Item 20.670.930.380.75 Item 30.550.860.230.57 Item 40.610.900.350.72 Item 50.841.000.440.77Mean internal consistency coefficients across occasions were calculated for the IO and OD scales separately under each medication condition (see Table 2). The mean reliability (i.e., alpha coefficient) of the IO scale was found to be identical under both conditions (M =.89) and similar to the mean alpha coefficients identified for the OD scale under placebo (M =.90) and IR-MPH conditions (M = .91). Test–retest reliability and the standard error of measurement were derived using intraclass correlation coefficients (ICC) in order to handle multiple rating occasions and account for systematic error (McGraw & Wong, 1996; see Table 2). The ICCs for the AG scale (Placebo =.85, IR-MPH =.89) were slightly higher than those identified for the IO scale (Placebo =.86, IR-MPH =.87). All coefficients were found to exceed .80, thereby indicating high levels of agreement across time (Landis & Koch, 1977).Coefficient Alpha, 1-week Test-Retest Reliability Coefficients, and Standard Errors of Measurement by Scale and Condition (N = 67)PlaceboIR-MPHAlphaICC (SEM)AlphaICC (SEM)Inattention-Overactivity.890.86 (2.41).890.87 (3.01)Oppositional-Defiant.900.85 (2.81).910.89 (2.16)In the full-model G study, the proportion of variance in item ratings was calculated independently for each of the two IOWA Conners scales (i.e., IO and OD). The model involved all students (p) being rated on five items (i) across four occasions (o) within two medication conditions (c), resulting in a partially-nested p x i x (o:c) design. As evident in Table 3, the largest percentage of variance was attributable to the facet of person (19% IO, 29% OD), indicating that substantial differences among students were identified for both behaviors (across items and rating occasions). In contrast, a relatively small percentage of variance (7% IO, 6% OD) was attributable to the facet of condition, suggesting that there were only minor differences in overall student behavior between medication conditions. The moderate size of the interactions between person and occasion nested within condition (19% IO, 22% OD) and person and condition (14% IO, 12% OD), however, together suggest that the relative standing of individual students did not remain constant across medication conditions. That is, the students who received the highest ratings under placebo conditions were not necessarily the same students who received the highest ratings under the IR-MPH condition. Not surprisingly, the interaction between persons and items also helped to explain a small amount of rating variance (7% IO, 6% OD), suggesting that the rank ordering of students varied depending on the item rated. A minimal proportion (i.e., < 5%) of rating variance was explained by the remaining facets (i.e., occasion within condition) and interactions (i.e. condition by item, item by occasion within condition, person by condition by item). Finally, roughly one-fifth (20% IO, 22% OD) of the overall rating variance was not explained by the facets specified in the full model.Full Model Variance Component Results across Scales (N = 67)IOWA ConnersTeacher Rating ScaleFacetIOODPerson.23 (19%).21 (29%)Condition.09 (7%).04 (6%)Occasion: Condition.00 (0%).00 (0%)Item.09 (8%).01 (1%)Person × Condition.17 (14%).09 (12%)Person × Occasion: Condition.23 (19%).16 (22%)Person × Item.09 (7%).04 (6%)Condition × Item.00 (0%).00 (0%)Item × Occasion: Condition.00 (0%).00 (0%)Person × Condition × Item.07 (6%).02 (3%)Person × Occasion: Condition × Item + Residual.24 (20%).16 (22%)Although G study results based on the full model provide important information regarding the sensitivity of the measures to treatment conditions (i.e., the person by condition interaction), conducting D studies based on the full model was deemed to provide information with limited utility. That is, given the finding that one-third of the rating variance observed was attributable to differences across conditions (i.e., person by condition and person by occasion within condition facets), it was determined that recommendations for data collection needed to be made within condition. Reduced models were therefore investigated within treatment condition (i.e., placebo and IR-MPH) in which every item was rated on every occasion for all children (fully crossed p x i x o design). All G and D studies were performed for each scale (i.e., IO and OD) independently.Separate G studies were conducted within treatment condition to examine the independent influence of the facets (i.e., factors) of person, item, and occasion, as well as all interactions between these facets. Results for both the placebo and IR-MPH conditions are presented in Table 4, given the model in which five items were rated on four separate occasions. Across scales and conditions, the largest proportion of rating variance (34%-48%) was attributable to the facet of persons. This finding demonstrates that there were notable individual differences in both oppositional-defiant and inattention-overactive behaviors across items and rating occasions. These differences among students were found to be more pronounced when rating OD (40-48%) than IO (34-37%). Variance component results unique to each scale are presented next.Reduced Model Variance Component Results across Scales and Conditions (N = 67)IOWA ConnersTeacher Rating ScaleIOODFacetPlaceboIR-MPHPlaceboIR-MPHPerson.42 (34%).37 (37%).35 (40%).25 (48%)Occasion.00 (0%).00 (0%).00 (0%).00 (0%)Item.13 (10%).06 (6%).01 (1%).00 (1%)Person × Occasion.25 (21%).21 (21%).22 (25%).11 (21%)Person × Item.16 (13%).15 (15%).07 (8%).06 (12%)Occasion × Item.01 (0%).00 (0%).00 (0%).00 (0%)Person × Occasion × Item + Residual.26 (21%).21 (21%).23 (26%).09 (18%)ρ2.79.80.82.83Φ.76.78.81.83For the IO scale, the interaction between persons and items accounted for a moderate proportion of variance (13% placebo, 15% IR-MPH), whereas variance attributable to items was somewhat lower (10% placebo, 6% IR-MPH). That is, a greater proportion of rating variance was explained by changes in the rank order of students across items (i.e., different items were rated as most problematic for different students) than by differences in behavior (summed across students) across items. The facet of occasion (i.e., changes in overall student behavior across occasions) and the interaction between items and occasions (i.e., changes in the relative severity with which particular items were rated across occasions) contributed negligible variance to the model (i.e., < 5%) across both conditions. Finally, the percentage of rating variance explained by the residual error term (i.e., variance attributable to the three-way interaction plus unexplained variance) and the interaction between persons and occasions (i.e., changes in rank ordering of students across occasions) were found to be identical (21% placebo/IR-MPH) across both conditions. Given the enacted model (representing actual data collection procedures), in which teachers completed the IO scale on four separate occasions, reliability-like coefficients were found to be slightly higher for the purposes of relative decision making (i.e., ρ2: placebo: = .79; IR-MPH = .80) than absolute decision making (i.e., Φ: placebo = .76; IR-MPH = .78).For the OD scale, the interactions between persons and occasions (25% placebo, 21% IR-MPH) and persons and items (8% placebo, 12% IR-MPH) together helped to explain roughly one-third of the observed rating variance. In contrast, the variance components for item, occasion, and the interaction between items and occasions fell at or below 1%. The size of the residual error term was found to vary somewhat across conditions, with a greater proportion of variance within the placebo condition considered unexplained (26%) than in the IR-MPH condition (18%). Given the enacted model, wherein teachers completed the OD scale on four separate occasions, reliability-like coefficients were found to be comparable within each treatment condition for the purposes of relative (ρ2: placebo = .82; IR-MPH = .83) and absolute (Φ: placebo = .81; IR-MPH = .83) decision making.Utilizing the results of G studies based on the enacted model, it is possible to mathematically determine the optimal conditions of measurement through a series of D studies. In the current study, the facets of items and occasions were systematically manipulated in order to determine the most cost-effective measurement situation for which adequate generalizability (ρ2) and dependability (Φ) coefficients were achieved for the IO and OD scales (see Figures 1 & 2). A benchmark of .80 was employed, which is a commonly accepted criterion for reliability coefficients when making screening and progress monitoring decisions (Salvia, Ysseldyke, & Bolt, 2010), and is represented by a dashed line in the figures.For the IO scale, differences were noted between the coefficients obtained for the purposes of relative decision making (i.e., generalizability coefficients; see top row of Figure 1) and absolute decision making (i.e., dependability coefficients; see bottom row of Figure 1), suggesting that guidelines for data collection may vary depending on the intended purpose of assessment. For both placebo and IR-MPH conditions, adequate (i.e., .80) levels of reliability for either decision making purpose (i.e., relative or absolute) were not reasonably achieved (i.e., with less than 20 occasions) given the use of either a single-or 2-item scale. Within the placebo condition, the .80 criterion was met for the purposes of relative decision making if the number of administrations was as follows: the 3-item scale was administered across 7 occasions, the 4-item scale across 6 occasions, or the full 5-item scale across 5 occasions. Slightly fewer administrations were necessary to reach this criterion within the IR-MPH condition: the 3-item scale across 7 occasions, the 4-item scale across 5 occasions, or the full 5-item scale across 4 occasions. When examining reliability for the purposes of absolute decision making (see bottom row of Figure 1), however, the required number of administrations increased substantially. Additionally, the number of rating occasions necessary to achieve adequate levels of reliability under the placebo condition was found to be higher than under the IR-MPH condition. The 3-item scale would need to be administered across 12 (IR-MPH) to more than 20 (placebo) occasions, whereas the 4-item scale would require 7 (IR-MPH) to 10 (placebo) administrations. The full 5-item scale would need to be administered across 5 (IR-MPH) to 7 (placebo) occasions in order to achieve adequate reliability.For the OD scale, identified coefficients were roughly equivalent for the purposes of relative (see top row of Figure 2) and absolute decision making (see bottom row of Figure 2) across both placebo and IR-MPH conditions. Using a single-item scale, adequate reliability was not achieved given 20 measurement occasions under each condition. However, the .80 criterion was met for the purposes of relative decision making given the administration of a 2-item scale across 6 (placebo) to 7 (IR-MPH) occasions, a 3-item scale across 5 occasions (both conditions), or a 4-item scale across 4 occasions (both conditions). Across both placebo and IR-MPH conditions, the full 5-item rating scale would need to be completed on 4 occasions in order to achieve an adequate level of reliability for the purposes of relative decision making. Obtained coefficients for the purposes of absolute decision making were found to be similar in magnitude. Results indicated that adequate levels of reliability could be achieved for a 3-item scale across 5 occasions and a 4-item scale across 4 occasions in both conditions. The full 5-item rating scale would need to be administered across 4 occasions in both conditions in order to achieve sufficient levels of dependability.The purpose of the current study was to explore how G theory may be used to maximize data collection procedures by examining rating scale data from the IOWA Conners Teacher Rating Scale (Loney & Milich, 1982). First, two classical test theory methods were employed to examine the reliability of the two scales of the IOWA Conners across placebo control and IR-MPH conditions. Internal consistency coefficients were high for both scales (i.e., r =.87-.95), with average coefficients either identical (IO) or comparable (OD) across the medication conditions. Likewise, ICC involving repeated measures within condition indicated very strong agreement (i.e., 0.85-0.87) for both the IO and OD scales across time. This finding is consistent with other studies examining child classroom behavior (Bergeron et al., 2008; Chafouleas, Christ, Riley-Tillman, Briesch, & Chanese, 2007; Hintze & Matthews, 2004) and suggests minimal systematic influence of time on behavior averaged across students. These results are consistent with reliability evidence previously established for the IOWA Conners (e.g., Atkins et al., 1989; Loney & Milich, 1982; Pelham, Milich, Murphy, & Murphy, 1989); however, the limitations of information provided through such analyses has been noted (Cronbach, 2004). That is, traditional reliability coefficients describe only the relative ranking of persons, and therefore cannot be used to inform decision making for absolute (i.e. intra-individual) decisions, as would be the focus within a progress monitoring context. Furthermore, both coefficient alpha and ICC essentially represent one-facet G studies, in which a single source of measurement error (i.e., item and time) is of interest and all other variance is considered to be random or unexplained (Shrout & Fleiss, 1979). The magnitude of the ICC values, for example, indicates that the rank ordering of students on IO and OD dimensions was highly stable across rating occasions. G theory, however, permits the partitioning of error by its source in order to facilitate simultaneous examination of multiple sources of variance (Lei, Smith, & Suen, 2007; Suen & Ary, 1989, Suen & Rzasa, 2004). Such an approach has greater utility in the identification and design of efficient assessment approaches.The G studies employed in the current study were examined to inform assessment design by identifying and quantifying multiple sources of error and their interactions. Specifically, the facets of items and occasions were manipulated under placebo and stimulant medication conditions to identify combinations of these sources of error necessary to reach an acceptable level of dependability for both inter- and intra-individual decisions. Results of the full model indicated that ratings assigned to students differed substantially depending on whether students were evaluated during placebo or the stimulant medication condition. That is, roughly one-third of the observed rating variance could be explained by changes in the relative standing of students across the conditions (i.e., person by condition and person by occasion within condition facets). One possible explanation of this finding was that students demonstrated differential response to medication, which is a common finding in studies evaluating the effects of stimulant medication (e.g., see Barkley, 2006). Also, it is not surprising that variability in teacher perceptions of inattentive-overactive and oppositional-defiant behaviors differed depending on whether or not the student was receiving medication, as the behavior of children with behavior problems tends to be less variable in medicated compared to non-medicated conditions (e.g., Gadow et al., 1995). Findings that variability in inattentive-overactive and oppositional-defiant behavior differed across treatment conditions do, however, highlight the need to consider behavioral variability when recommending assessment procedures.Due to the large proportion of variance associated with condition in the full model, reduced models involving the facets of items and occasions within each medication condition were examined. Such models would be most informative for practice because, in data-based decision-making, the goal is typically to generalize estimates within a specific condition and then compare these estimates across conditions. Across scales and treatment conditions examined within a reduced model, the facet of occasions contributed no variance to models, suggesting that there were no overall changes in behavior (summed across students) over time. Similarly, the interaction between items and occasions did not contribute variance to any model, suggesting that there were no universal differences in how individual items were rated over time. Interesting differences were identified, however, between the IO and OD scales for the variance attributable to the facet of item. For the IO scale, between 6% and 10% of rating variance was explained by differences between items; however, negligible variance was attributable to items of the OD scale. These findings indicate a greater degree of variability in how IO items were rated. One explanation for these findings is that the OD scale arguably measures a single construct of defiant behavior, whereas IO scale items assess both inattention and hyperactivity-impulsivity (see Waschbusch & Willoughby, 2008). The defining symptoms of ADHD in the various iterations of the Diagnostic and Statistical Manual of Mental Disorders (DSM) have changed over the last 20 years (see Barkley, 2006), but recent research supports the view that inattention should be considered distinct from hyperactivity-impulsivity (see Dumenci, McConaughy, & Achenbach, 2004; Milich, Balentine, & Lynam, 2001; Pillow, Pelham, Hoza, Molina, & Stultz, 1998). Moreover, this distinction is reflected in the current DSM-IV (American Psychological Association, 2000).The interaction between persons and occasions consistently contributed a substantial proportion of variance (roughly 21%) across both the IO and OD scales. This finding is consistent with prior G studies relating to student classroom behavior (Chafouleas, Briesch et al., 2010; Hintze & Matthews, 2004; Volpe et al., 2009) and indicates that both scales were sensitive to relative (i.e., inter-individual) changes in student behavior across time. Current results suggest that behavioral change over time was demonstrated to an equivalent degree across the two 5-item scales.The interaction between persons and items was somewhat more pronounced for the IO scale (between 13% and 15%) than the OD scale (between 8% and 12%). These differences indicate that the relative standing of students differed from one item to another and that these changes in rank occurred more often for the IO scale. Again, the relative heterogeneity of the IO over the OD scale may account for these differences. Moreover, the relevance of any one item will vary from one student to the next (see Volpe, Gadow et al., 2009). Finally, between 18% and 26% of variance was not explained by elements of each G study, which is comparable to other G studies investigating classroom behavior (e.g., Bergeron et al., 2008; Hintze & Matthews, 2004). This finding supports the utility of the models examined in this study to account for the observed variance in teacher ratings, yet additional facets might further account for sources of error.Generalizability and dependability coefficients were comparable for the OD scale, suggesting that the same methods can be used to make relative and absolute decisions (i.e., criterion-referenced or within person). However the fact that generalizability coefficients were somewhat stronger than dependability coefficients for the IO scale suggests that the scale may be slightly more efficient for relative decision making (where the goal is to rank order students) than for making absolute decisions.Results from D studies of IO and OD scales indicated several combinations of items and occasions meeting our criterion for dependability (i.e. .80; Salvia et al., 2010). Single-item scales almost always required more than 20 rating occasions, with the exception of the OD scale under placebo conditions, in which 20 occasions were needed in order to adequately inform relative decision making. In the case of the IO scale, a 2-item scale functioned minimally better, with the number of necessary rating occasions continuing to prove impractical (i.e., more than 20). In fact, sufficient reliability was only achieved within a reasonable (i.e., 4-7) number of rating occasions in the case of the full 5-item scale.In contrast, notable improvements in dependability were found with the addition of items to the OD scale, such that the 2-item OD scale functioned similarly to the 5-item IO scale with regard to suggested administration frequency (between six and eight occasions were required to meet the .80 criterion). Adding a third item to the OD scale decreased the number of rating occasions needed to five. Compared to the gains noted with the addition of one and two items to single-item scales, however, relatively minor improvements were noted with the addition of a fourth and fifth item. Similarly, notable gains in dependability occurred with the addition of the first few rating occasions; however, thereafter gains in dependability flattened. These findings suggest that the traditional Spearman–Brown-influenced idea that more is better may not always be true. That is, our results suggest that a point of diminishing returns may be reached with a relatively short list of items.The fact that fewer items and occasions were needed to achieve sufficient levels of reliability with the OD scale is not surprising, given that the facet of item contributed minimal variance to the model for this scale (in comparison to 10% for IO). This finding does, however, highlight the fact that recommendations regarding data collection will likely vary depending on the homogeneity of scale items, thus suggesting important implications for the development of brief rating scales. That is, if items are selected from a larger existing rating scale based on either factor loading criteria (e.g., Reynolds & Kamphaus, 2004; Volpe et al., 2009) or sensitivity to change criteria (e.g., Gresham et al,, in press; Meier et al., 1997, 2008), recommendations for data collection (i.e., how many data points needed) may vary depending on which items are selected.The finding that existing rating scales can be shortened while still maintaining acceptable reliability-like coefficients is interesting and important, but does not address the degree to which these brief scales demonstrate adequate criterion-related validity and treatment sensitivity. Volpe & Gadow (in press) used the current dataset to compare the criterion-related validity and treatment sensitivity of the full-length (5-item) IO and OD scales of the IOWA Conners to 3-item scales derived via factor-derived and individualized methods. All three versions demonstrated treatment sensitivity and adequate concurrent validity with teacher ratings and systematic direct observation. Hence, the 3-item scales examined by Volpe and Gadow appear defensible for progress-monitoring purposes. Although scores from 1- or 2- item instruments administered repeatedly might achieve acceptable levels of dependability, it is unknown whether they would possess adequate criterion-related validity (i.e., correlate well with well established measures of IO or OD). The magnitude of association likely would depend on the items selected for the shortened instrument. Recent work examining customized rating scales that have been individualized for each child (Volpe et al., 2009; Volpe & Gadow, in press) suggests that a traditional view of criterion-related validity (i.e., examining the degree to which these brief measures correlate with criterion measures of constructs such as ADHD), may not be the appropriate question when evaluating measures for progress monitoring tasks. In contrast to screening assessments for which that question is central, the primary concern in the context of progress monitoring is whether there is improvement in a child's unique pattern of problems as a function of intervention efforts. Nevertheless, at least in regard to single-item measures, daily behavior ratings (e.g., Chafouleas, Briesch et al., 2010) likely would be a better option for progress monitoring because they afford greater variability (typically rated on a continuous or 11-point scale) and typically provide a more detailed description of the construct of interest (see Chafouleas et al., 2007).Although rating scales may not prove to be as sensitive to the effects of treatment as more specific measures of discrete target behaviors (e.g., systematic direct observation of kicking), there is growing evidence that they may serve as viable general outcome measures for the constructs they were designed to measure. Volpe and Gadow (in press) have made the argument that measures of both discrete target behaviors (e.g., short-term performance objectives) and broader measures (e.g., long-term general outcome measures) are needed in a comprehensive progress monitoring system.The finding that a variety of possible item-occasion combinations were identified that met criteria for adequate dependability suggests that recommendations regarding data collection may be more flexible than those suggested with comparable methods. For example, recent studies have provided important recommendations with regard to the number of systematic direct observation (Hintze & Matthews, 2004; Volpe et al., 2009) and direct behavior rating (DBR; Chafouleas et al., 2007; Chafouleas, Briesch et al., 2010) data points needed to achieve adequate levels of dependability. However, it is expected that these guidelines will be followed across all potential users, rather than providing room for choice or adaptation. In contrast, results of the current study suggest that a dependable estimate of student behavior may be obtained given the use of a shorter rating scale over several days or a longer rating scale over fewer days. Use of brief rating scales can therefore be tailored to individual preference. In some cases, it may be necessary to collect data quickly, in which case the full 5-item rating scale can be administered across three occasions. When time is not a limiting factor, however, two or three items could be administered across a greater number of occasions to obtain a comparable estimate of student behavior. The consumer is therefore provided with a menu of possible assessment options to be used in designing an individualized assessment approach. This flexibility could be provided through consultation with the school psychologist, via computer software, or both. Computer software could automate the selection of items and construction of rating forms that could be rated directly on the computer or printed out and completed later. In addition, the storage and summarization afforded by a computer-based system could streamline the data management and decision-making process.Although results of this study provide important preliminary information regarding use of G theory to inform the flexible use of rating scales, inferences derived from the study should be tempered by several limitations. First, rating scale data were examined across a small number of occasions (4) within each medication condition. Although the total number of data points was sufficient for conducting analyses given the sample size (N = 67), the stability of variance component estimates is known to improve as the number of specific instances (i.e., raters, ratees, and occasions) sampled increases (Smith, 1981). Future studies might collect rating scale data across a greater number of occasions to improve the reliability of variance estimation.Next, as noted previously, roughly 20% of the variance in ratings was not accounted for by the modeled facets. One facet that could not be modeled in the current study and therefore was not explored was that of rater. In this regard, Bergeron and colleagues (2008) found that up to five percent of the variance in rating scale composite scores was attributable to differences between raters and up to 12 percent of variance was attributable to the interaction between persons and raters. Discrepancies in the way informants rate behavior in a general sense (i.e., leniency/severity effect; Thorndike & Hagen, 1977), as well as how they rate particular students, may therefore account for an additional variability in rating that should be explored in future studies. Because each child in the current study was rated by a different teacher, the variance attributable to differences between raters becomes confounded with, and subsumed under, the facet of persons. As a result, current findings may provide an overestimate of actual person-related variance. In order to avoid this problem, and distinguish person and rater-related variance, it would be necessary to have multiple students rated by the same teacher as has been demonstrated in generalizability studies of DBR (e.g., Chafouleas et al., 2007; Chafouleas, Briesch et al., 2010) and SDO (e.g., Hintze & Matthews, 2004) conducted in general education classrooms. However, if the goal is to limit assessment procedures to students with identified behavioral problems or disabilities, it may prove difficult to identify appropriate settings (i.e., wherein multiple target students can rated by the same teacher).It should be noted that teachers rated the dependent variables along with a packet of other rating scales. Hence, it is possible that the order in which ratings were completed may have influenced teacher reports. However, order or anchoring effects are likely minor. Moreover, teachers often are asked to complete several ratings as part of a comprehensive assessment battery. Lastly, similar to use of the Spearman–Brown prophecy formula, one of the assumptions inherent in use of G theory is that the items selected are of equal value. This assumption is reasonable in making relative decisions. However, recent studies have suggested that some items may be more relevant than others when considering individual cases (Volpe & Gadow, in press) and that some items may be more sensitive to the effects of treatment than others (Gadow et al., 2008; Gresham et al., in press; Meier et al., 2008; Sprafkin, Mattison, Gadow, Schneider, & Lavigne, in press; Volpe et al., 2009). These findings suggest that dependability estimates for shortened scales may vary depending on the subset of items selected. Future research should more specifically explore this possibility, as significant implications for scale development may be indicated.Although the specific combinations of items and rating occasions identified in the current study must be considered specific to the IOWA Conners, these results have broader implications for the use of brief rating scales in formative assessment. Traditional rating scales typically demonstrate high levels of reliability; however, the administration of over 100 items is only feasible when done infrequently (e.g., when making diagnostic decisions). Results of the current study, however, illustrate the fact that increasing the length of a scale is not the only way to achieve adequate levels of reliability (Cortina, 1993). When the goal is to collect data repeatedly over time, such as when assessing student response to intervention, shorter rating scales may be administered over a greater number of occasions in order to obtain data with comparable levels of reliability. The availability of dependability data for a variety of item and occasion combinations allows the evaluator flexibility in designing assessment procedures. Informant preference (cf. Volpe et al., 2005) can be used to determine what combination of items and occasions are likely to lead to the highest levels of acceptability and adherence (e.g., Witt & Elliott, 1985). For example, for a teacher in a special education classroom it may be necessary to rate several students, thereby making even very brief ratings burdensome. Results from the current study would indicate that an OD scale as brief as two items would yield data of acceptable dependability after 6-8 days if completed each school day. In another situation, a teacher may find it more acceptable to complete a longer scale 1 to 2 days a week due to scheduling issues (e.g., on a day with an extra preparation period). In this case, the teacher might prefer to complete a full-length scale across fewer occasions. Using the current findings for the OD scale as an example, a dependable estimate of behavior could be obtained using the full 5-item scale across 4 days. Regardless of the scale examined, however, results suggest that use of a single-item scale would not be justifiable given the number of rating occasions necessary to obtain adequate levels of reliability. Across both scales, more than 20 rating occasions were needed, which would be unacceptable for most, if not all, assessment situations. Although there is at least initial support for single item scales using ten or more anchor points (e.g., Chafouleas, Briesch et al., 2010), items rated on a four-point scale seem not to be practical for progress monitoring purposes.As suggested by Parkes (2000), the utility of G theory has not been fully harnessed within the literature to date, given that desired levels of dependability need not be achieved through a single design. Rather, it is possible to identify the most efficient (i.e., maximally reliable and minimally expensive) assessment design by testing different combinations of facets. By taking the relative cost of each facet (e.g., cost of having additional informants provide ratings and time needed to complete each rating scale item) into consideration, it is possible to quantify the efficiency of a given assessment approach. Another key consideration is that of acceptability. The most efficient procedure may not necessarily be the most acceptable or sustainable. Identifying several procedures with comparable psychometric properties allows the evaluator the flexibility of taking acceptability into account without necessarily sacrificing the quality of the resultant data.Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.Achenbach TM, McConaughy SH, Howell CT. Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. Psychological Bulletin. 1987;101:213–232. [PubMed] [Google Scholar]American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 3rd. Washington, DC: Author; 1987. [Google Scholar]American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th. Washington, DC: Author; 2004. [Google Scholar]Atkins MS, Pelham WE, Licht MH. The differential validity of teacher ratings of inattention/overactivity and aggression. Journal of Abnormal Child Psychology. 1989;17:423–435. [PubMed] [Google Scholar]Barkley RA. Attention deficit hyperactivity disorder: A handbook for diagnosis and treatment. 3rd. New York: Guilford Press; 2006. [Google Scholar]Bergeron R, Floyd RG, McCormack AC, Farmer WL. The generalizability of externalizing behavior composites and subscale scores across time, rater, and instrument. School Psychology Review. 2008;37:91–108. [Google Scholar]Brennan RL. Generalizability theory. New York: Springer; 2001. [Google Scholar]Briesch AM, Volpe RJ. Selecting progress monitoring tools for evaluating social behavior. School Psychology Forum. 2007;1:59–74. [Google Scholar]Cates GL, Skinner CH, Watson ST, Meadows TJ, Weaver A, Jackson B. Instructional effectiveness and instructional efficiency as considerations for data-based decision making: An evaluation of interspersing procedures. School Psychology Review. 2003;32:601–616. [Google Scholar]Chafouleas SM, Briesch AM, Riley-Tillman TC, Christ TJ, Kilgus SJ. An investigation of the generalizability and dependability of Direct Behavior Rating Single-Item Scales (DBR-SIS) to measure academic engagement and disruptive behavior of middle school students. Journal of School Psychology. 2010;48:219–246. [PubMed] [Google Scholar]Chafouleas SM, Christ TJ, Riley-Tillman TC, Briesch AM, Chanese JAM. Generalizability and dependability of direct behavior ratings to assess social behavior of preschoolers. School Psychology Review. 2007;36:63–79. [Google Scholar]Chafouleas SM, Volpe RJ, Gresham FM, Cook CR. School-based behavioral assessment within problem-solving models: Current status and future directions. School Psychology Review in press. [Google Scholar]Christ TC, Johnson-Gros KN, Hintze JM. An examination of alternate assessment durations when assessing multiple-skill computational fluency: The generalizability and dependability of curriculum-based outcomes within the context of educational decisions. Psychology in the Schools. 2003;42:615–622. [Google Scholar]Cone JD. The behavioral assessment grid (BAG): A conceptual framework and taxonomy. Behavior Therapy. 1978;9:882–888. [Google Scholar]Cortina JM. What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology. 1993;78:98–104. [Google Scholar]Crick JE, Brennan RL. Manual for GENOVA: A generalized analysis of variance system. Iowa City, IA: The American College Testing Program; 1983. [Google Scholar]Cronbach LJ, Gleser CG, Rajaratnam N, Nanda H. The dependability of behavioral measurements. New York: Wiley; 1972. [Google Scholar]Cronbach LJ. My current thoughts on coefficient alpha and successor procedures. Educational and Psychological Measurement. 2004;64:391–418. [Google Scholar]Dumenci L, McConaughy SH, Achenbach TM. A hierarchical 3-factor model of inattention-hyperactivity-impulsivity derived from the Attention Problems Syndrome of the Teacher's Report Form. School Psychology Review. 2004;33:287–301. [Google Scholar]DuPaul GJ, Stoner G. ADHD in the schools: Assessment and intervention strategies. 2nd. New York: Guilford Press; 2003. [Google Scholar]Elliott SN, Witt JC, Galvin G, Peterson R. Acceptability of positive and reductive interventions: Factors that influence teachers' decisions. Journal of School Psychology. 1984;22:353–360. [Google Scholar]Enders CK. A primer on maximum likelihood algorithms available for use with missing data. Structural Equation Modeling. 2001;8:128–141. [Google Scholar]Fabiano GA, Pelham WE, Gnagy EM, Burrows-MacLean L, Coles EK, Chacko A, Robb JA. The single and combined effects of multiple intensities of behavior modification and multiple intensities of methylphenidate in a classroom setting. School Psychology Review. 2007;36:195–216. [Google Scholar]Fabiano GA, Vujnovic R, Pelham WE, Waschbusch DA, Massetti GM, Yu J, Volker M. Enhancing the effectiveness of special education programming for children with ADHD using a daily report card. School Psychology Review in press. [Google Scholar]Gadow KD, Nolan EE, Paolicelli LM, Sprafkin J. A procedure for assessing the effects of methylphenidate on hyperactive children in public school settings. Journal of Clinical Child Psychology. 1991;20:268–276. [Google Scholar]Gadow KD, Nolan EE, Sprafkin J, Sverd J. School observations of children with attention-deficit hyperactivity disorder and comorbid tic disorder: Effects of methylphenidate treatment. Journal of Developmental and Behavioral Pediatrics. 1995;16:167–176. [PubMed] [Google Scholar]Gadow KD, Nolan EE, Sverd J, Sprafkin J, Schneider J. Methylphenidate in children with oppositional defiant disorder and both co-morbid chronic multiple tic disorder and ADHD. Journal of Child Neurology. 2008;23:981–990. [PubMed] [Google Scholar]Gadow KD, Sprafkin J. ADHD Symptom Checklist -4 2008 Manual. Stony Brook, NY: Checkmate Plus; 2008. [Google Scholar]Gadow KD, Sverd J, Nolan EE, Sprafkin J, Schneider J. Immediate-release methylphenidate for ADHD in children with comorbid chronic multiple tic disorder. Journal of the American Academy of Child and Adolescent Psychiatry. 2007;46:840–848. [PubMed] [Google Scholar]Gresham F, Cook C, Collins T, Dart E, Rasetshwane K, Truelson E, Grant S. Developing a change sensitive brief behavior rating scale as a progress monitoring tool for social behavior: An example using the Social Skills Rating system-Teacher Form. School Psychology Review in press. [Google Scholar]Griffiths A, VanDerHeyden AM, Skokut M, Lilles E. Progress monitoring in oral reading fluency within the context of RTI. School Psychology Quarterly. 2009;24:13–23. [Google Scholar]Hintze JM, Christ TJ, Keller LA. The generalizability of CBM survey-level mathematics assessments: Just how many samples do we need? School Psychology Review. 2002;31:514–528. [Google Scholar]Hintze JM, Matthews WJ. The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral assessment. School Psychology Review. 2004;33:258–270. [Google Scholar]Hyman IA, Wojtowicz A, Duke-Lee K, Haffner M, Fiorello CA, Storlazzi JJ, Rosenfeld J. School-based methylphenidate placebo protocols: Methodological and practical issues. Journal of Learning Disabilities. 1998;31:581–594. [PubMed] [Google Scholar]Kurlan R. Tourette's syndrome: Current concepts. Neurology. 1989;39:1625–1630. [PubMed] [Google Scholar]Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–174. [PubMed] [Google Scholar]Lei P, Smith M, Suen HK. The use of generalizability theory to estimate data reliability in single-subject observational research. Psychology in the Schools. 2007;44:433–439. [Google Scholar]Loney J, Milich R. Hyperactivity, inattention, and aggression in clinical practice. In: Wolraich M, Routh DK, editors. Advances in developmental and behavioral pediatrics. Vol. 3. Greenwich, CT: JAI; 1982. pp. 113–147. [Google Scholar]McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological Methods. 1996;1:30–46. [Google Scholar]McIntosh K, Spaulding SA, Frank J. Establishing research-based trajectories of office discipline referrals for individual students. School Psychology Review in press. [Google Scholar]Meier ST, McDougal JL, Bardos A. Development of a change-sensitive outcome measure for children receiving counseling. Canadian Journal of School Psychology. 2008;23:148–160. [Google Scholar]Merrell KW. Behavioral, social, and emotional assessment of children & adolescents. 3rd. Mahwah, NJ: Erlbaum; 2008. [Google Scholar]Milich R, Balentine A, Lynam DR. ADHD combined type and inattentive type are distinct and unrelated disorders. Clinical Psychology: Science and Practice. 2001;8:463–488. [Google Scholar]Milich R, Landau S. Teacher ratings of inattention/overactivity and aggression: Cross-validation with classroom observations. Journal of Clinical Child Psychology. 1988;17:92–97. [Google Scholar]Nist L, Joseph LM. Effectiveness and efficiency of flashcard drill instructional methods on urban first-graders' word recognition, acquisition, maintenance, and generalization. School Psychology Review. 2008;37:294–308. [Google Scholar]Nolan EE, Gadow KD. Relation between ratings and observations of stimulant drug response in hyperactive children. Journal of Clinical Child Psychology. 1994;23:78–90. [Google Scholar]Parkes J. The relationship between the reliability and cost of performance assessments. Education Policy Analysis Archives. 2000 Mar;8(16) Retrieved March 8, 2008 from [Google Scholar]Pelham WE. Pharmacotherapy for children with attention-deficit hyperactivity disorder. School Psychology Review. 1993;21:199–223. [Google Scholar]Pelham WE, Fabiano GA, Massetti GM. Evidence-based assessment of attention deficit hyperactivity disorder in children and adolescents. Journal of Clinical Child and Adolescent Psychology. 2005;3:449–476. [PubMed] [Google Scholar]Pelham WE, Milich R, Murphy DA, Murphy HA. Normative data on the IOWA Conners teacher rating scale. Journal of Clinical Child Psychology. 1989;18:259–262. [Google Scholar]Pillow DR, Pelham WE, Hoza B, Molina BSG, Stultz CH. Confirmatory factor analyses examining attention deficit hyperactivity disorder symptoms and other childhood disruptive behaviors. Journal of Abnormal Child Psychology. 1998;26:293–309. [PubMed] [Google Scholar]Poncy BC, Skinner CH, Axtell PA. An investigation of the reliability and standard error of measurement of words read correctly per minute using curriculum-based measurement. Journal of Psychoeducational Assessment. 2005;23:326–338. [Google Scholar]Reich W. Diagnostic Interview for Children and Adolescents (DICA) Journal of the American Academy of Child and Adolescent Psychiatry. 2000;39:59–66. [PubMed] [Google Scholar]Reynolds CR, Kamphaus RW. Behavior Assessment System for Children-2 (BASC-2) Circle Pines, MN: American Guidance Service; 2004. [Google Scholar]Salvia J, Ysselydke JE, Bolt S. Assessment in special and inclusive education. 11th. Boston: Houghton Mifflin; 2010. [Google Scholar]Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate Behavioral Research. 1998;33:545–571. [PubMed] [Google Scholar]Shapiro ES, Heick P. School psychologist assessment practices in the evaluation of students referred for social/behavioral/emotional problems. Psychology in the Schools. 2004;41:551–561. [Google Scholar]Shavelson RJ, Webb NM. Generalizability theory: A primer. Newbury Park, CA: Sage; 1991. [Google Scholar]Shinn M. Best practices in curriculum-based measurement in a problem-solving model. In: Thomas A, Grimes J, editors. Best practices in school psychology V. Vol. 2. Bethesda, MD: National Association of School Psychologists; 2008. pp. 243–261. [Google Scholar]Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin. 1979;86:420–428. [PubMed] [Google Scholar]Silva F. Psychometric foundations of behavioral assessment. Newbury Park, CA: SAGE; 1993. [Google Scholar]Smith PL. Gaining accuracy in generalizability theory: Using multiple designs. Journal of Educational Measurement. 1981;18:147–154. [Google Scholar]Sprafkin J, Mattison RE, Gadow KD, Schneider J, Lavigne J. A DSM-IV-referenced teacher rating scale for monitoring behavioral improvement. Journal of Attention Disorders in press. [PubMed] [Google Scholar]Suen HK, Ary D. Analyzing quantitative behavioral observation data. Hillsdale, N.J.: Erlbaum; 1989. [Google Scholar]Suen HK, Rzasa SE. Psychometric foundations of behavioral assessment. In: Haynes SN, Heiby EM, editors. Comprehensive handbook of psychological assessment. Vol. 3. Hoboken, NJ: Wiley; 2004. pp. 37–56. [Google Scholar]Sugai G, Horner RH. The evolution of discipline practices: School-wide positive behavior supports. Child & Family Behavior Therapy. 2002;24(1/2):23–50. [Google Scholar]Thorndike R, Hagen E. Measurement and evaluation in psychology and education. 4th. New York: Wiley; 1977. [Google Scholar]Volpe RJ, Gadow KD. Creating abbreviated rating scales to monitor classroom inattention-overactivity, aggression, and peer conflict: Reliability, validity, and treatment sensitivity. School Psychology Review in press. [Google Scholar]Volpe RJ, Gadow KD, Blom-Hoffman J, Feinberg AB. Factor-analytic and individualized approaches to constructing brief measures of ADHD behavior. Journal of Emotional and Behavioral Disorders. 2009;17:118–128. [Google Scholar]Volpe RJ, Heick P, Gureasko-Moore D. An agile model for monitoring the effects of stimulant medication in schools. Psychology in the Schools. 2005;42:509–523. [Google Scholar]Volpe RJ, McConaughy SH, Hintze JM. Generalizability of classroom behavior problem and on-task scores from the Direct Observation Form. School Psychology Review. 2009;38:382–401. [Google Scholar]Walker B, Cheney D, Stage S, Blum C. Schoolwide screening and positive behavior supports: Identifying and supporting students at risk for school failure. Journal of Positive Behavior Interventions. 2005;7:194–204. [Google Scholar]Walker HM, Severson HH. Systematic Screening for Behavior Disorders: Users guide and administration manual. Longmont, CO: Sopris West; 1990. [Google Scholar]Waschbusch DA, Willoughby MT. Parent and teacher ratings on the IOWA Conners Rating Scale. Journal of Psychopathology and Behavioral Assessment. 2008;30:180–192. [Google Scholar]Witt JC, Elliott SN. Acceptability of classroom intervention strategies. In: Kratochwill TR, editor. Advances in school psychology. Hillsdale, NJ: Erlbaum; 1985. pp. 251–288. [Google Scholar]