Psychological Assessment

Integration of Symptom Ratings From Multiple Informants in ADHD Diagnosis: A Psychometric Model With Clinical Utility

Michelle M. Martel, Ulrich Schimmack, Molly Nikolas, and Joel T. Nigg

Online First Publication, March 2, 2015. http://dx.doi.org/10.1037/pas0000088

Psychological Assessment

1040-3590/15/$12.00 http://dx.doi.org/10.1037/pas0000088

Integration of Symptom Ratings From Multiple Informants in ADHD Diagnosis: A Psychometric Model With Clinical Utility

Michelle M. Martel
University of Kentucky

Ulrich Schimmack
University of Toronto

Molly Nikolas
University of Iowa

Joel T. Nigg
Oregon Health and Sciences University

The *Diagnostic and Statistical Manual of Mental Disorder—Fifth Edition* explicitly requires that attention-deficit/hyperactivity disorder (ADHD) symptoms should be apparent across settings, taking into account reports from multiple informants. Yet, it provides no guidelines how information from different raters should be combined in ADHD diagnosis. We examined the validity of different approaches using structural equation modeling (SEM) for multiple-informant data. Participants were 725 children, 6 to 17 years old, and their primary caregivers and teachers, recruited from the community and completing a thorough research-based diagnostic assessment, including a clinician-administered diagnostic interview, parent and teacher standardized rating scales, and cognitive testing. A best-estimate ADHD diagnosis was generated by a diagnostic team. An SEM model demonstrated convergent validity among raters. We found relatively weak symptom-specific agreement among raters, suggesting that a general average scoring algorithm is preferable to symptom-specific scoring algorithms such as the "or" and "and" algorithms. Finally, to illustrate the validity of this approach, we show that averaging makes it possible to reduce the number of items from 18 items to 8 items without a significant decrease in validity. In conclusion, information from multiple raters increases the validity of ADHD diagnosis, and averaging appears to be the optimal way to integrate information from multiple raters.

Keywords: ADHD, diagnosis, assessment, structural equation modeling

The *Diagnostic and Statistical Manual of Mental Disorders—Fifth Edition* (DSM-5), similar to the prior edition (*DSM–IV–TR*), defines attention-deficit/hyperactivity disorder (ADHD) with a list of nine inattentive and nine hyperactive-impulsive behavioral symptoms, with slightly expanded content to capture age-related variability in symptoms, of which children must manifest at least six in one of the two symptom domains, as well as substantial interference in functioning occurring in two or more settings (e.g., at home and at school; American Psychiatric Association, 2013). The DSM-5 text strongly encourages clinicians to obtain information from more than one informant who see the individual in more than one setting, such as parents and teachers. The DSM-5 text explains that "confirmation of substantial symptoms across setting typically cannot be done accurately without consulting informants who have seen the individual in those settings" (American Psychiatric Association, 2013, pp. 37–38). Yet, there remains no standardized approach to integration of these multiple sources of information in making an ADHD diagnosis for research or—perhaps more importantly—clinical purposes. The DSM-5 did not provide one due to insufficient empirical data on the best approach. This paper attempts to fill that gap.

Prior work indicates that young children are less than adequate informants about their externalizing behavior problems, particularly in the domains of inattention and hyperactivity-impulsivity, based on low agreement of children's ratings with parent and teacher ratings (Bird, Gould, & Staghezza, 1992; Loeber, Green, Lahey, & Stouthamer-Loeber, 1989; reviewed by De Los Reyes & Kazdin, 2005). Further, research on clinical assessment procedures suggests that clinician judgment, by itself, can be less than ideal due to the presence of rating biases (Dawes, Faust, & Meehl, 1989; Holmbeck et al., 2008; Voigt et al., 2007), and in the case of ADHD, it is not clear that it is cost-effective over and above parent and teacher report on symptom checklists (Pelham, Fabiano, & Massetti, 2005). Whereas parents and teachers are the key sources of information in ADHD evaluation, individual parent and teacher ratings are also subject to rater bias effects. Their appropriate

Michelle M. Martel, Psychology Department, University of Kentucky; Ulrich Schimmack, Psychology Department, University of Toronto; Molly Nikolas, Psychology Department, University of Iowa; Joel T. Nigg, Psychiatry Department, Oregon Health and Sciences University.

The project was supported by Award Number R01-MH070004-01A2 from the National Institute of Mental Health. Michelle M. Martel was supported by K12 DA 035150. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health. The authors also thank all participating children and their families for making this work possible.

Correspondence concerning this article should be addressed to Michelle M. Martel, Psychology Department, University of Kentucky, Lexington, KY 40506. E-mail: michelle.martel@uky.edu

combination, however, is likely to enhance diagnostic validity. Parents and teachers both appear to provide reliable, discriminant, and dimensional ratings of childhood ADHD symptoms, and these ratings have demonstrated convergent validity (Gomez, 2007, 2008; Gomez, Vance, & Gomez, 2011).

Crucially, parent and teacher ratings of child ADHD symptoms are only moderately correlated ($r \sim .6$). These moderate correlations may reflect systematic measurement error due to response styles, rating biases, unique perspectives of raters, or cross-situational variability in child symptom expression (Achenbach, 2011; De Los Reyes, 2013; De Los Reyes et al., 2011; Dirks, De Los Reyes, Briggs-Gowan, Cella, & Wakschlag, 2012; Piacentini, Cohen, & Cohen, 1992). In particular, data on youth psychopathology, though not specific to ADHD, suggest that some of the unique variance reflects situation-specific behaviors and that this information can enhance the prediction of mood problems and functioning (De Los Reyes, Henry, Tolan, & Wakschlag, 2009; Dirks, Boyle, & Georgiades, 2011; Dirks et al., 2012). Despite these insights, there remains no empirically based consensus on the best way to integrate parent and teacher reports of ADHD symptoms to diagnose ADHD, and the sources of disagreement among raters of ADHD have not been fully explored. Currently, researchers and clinicians use a variety of ad hoc approaches to integrate across parent and teacher ratings.

One such approach, commonly utilized in research studies because it was relied on in the DSM–IV field trials (Lahey et al., 1994), is often referred to as the or algorithm. This approach specifies that a symptom is present if either the parent or the teacher endorses a specific symptom. Each symptom is only counted once (even if both raters endorse it), so that the number of symptoms present when counted still yields a score from 0 to 9 for the 9 items of the inattentive and hyperactive-impulsive scales, respectively. This approach to integration of symptom ratings across multiple informants is simple, easily standardized, and reasonably easy to utilize in clinical settings. Yet, this approach has several problems that may lower the validity of ADHD diagnoses (Solanto & Alvir, 2009; Valo & Tannock, 2010). For example, if one rater has an acquiescence bias and agrees to all symptoms, ratings of the other rater are essentially ignored because it is sufficient for one rater to endorse a symptom. Additionally, random measurement error will not cancel out symmetrically. That is, if random error leads to the false rating that a symptom is present, a symptom is coded as present. If, however, random error leads to a false rating that a symptom is absent, it may have no effect as long as the other rater correctly rates the symptom as present. Additionally, if each rater rates three different symptoms, a child can meet criteria for ADHD despite exhibiting few problems in either setting. Such a diagnosis of ADHD would look different, but be similarly labeled as a child who exhibits extensive symptoms in both settings. Further, the or algorithm becomes increasingly problematic as the number of raters increases because, as the number of reporters increase, symptom levels necessarily become inflated to the point that, with enough reporters, a child with few symptoms in any one setting may still be described as exhibiting every ADHD symptom. These undesirable consequences of the or algorithm undermine the benefit of multitrait multimethod assessments, wherein adding more data sources should increase validity (Campbell & Fiske, 1959; Connelly & Ones, 2010; Schimmack, 2010). It is difficult to escape the impression that, despite its convenience, the or algorithm is a suboptimal scoring method to integrate ratings across multiple individuals.

A second approach, which we refer to as the and algorithm, counts symptoms as present only if parent and teacher agree that a symptom is present. This scoring algorithm suffers from the same psychometric problems as the or algorithm, but with the opposite bias such that children with ADHD problems may remain undiagnosed and untreated. The reason is that a single rater with a conservative bias can dominate the diagnosis. In conclusion, to improve the diagnosis of ADHD, it is necessary to develop scoring methods that reduce systematic measurement error that is unique to individual raters, to evaluate these two approaches directly against one another for validity, and to consider additional alternative approaches as well.

We propose an alternative method that does not rely on dichotomous scoring of symptoms as either present or absent. Rather, we consider symptom ratings as a probabilistic function of symptom severity. This scoring method can take severity of symptom ratings into account (i.e., if each symptom is rated on a 0–3 scale as is commonly done in ADHD evaluations, that information is retained). Further, symptom ratings can be averaged even if raters disagree in the rating of a specific symptom. Thus, this approach allows for severity and more continuous aspects of symptomatology to be taken into consideration and is consistent with taxometric analyses of ADHD (Marcus & Barry, 2011). This scoring method is as simple as the *or* or the *and* algorithm, but it has the potential additional advantage that rater biases are reduced by averaging across raters with different rating biases. We evaluate this hypothesis here.

The current paper provides, to our knowledge, the first empirical comparison of the validity of or, and, and average rating algorithms for combining informant data for ADHD. It does so by pitting each method against a "gold standard" of a best-estimate diagnosis provided by a clinical diagnostic team comprising a licensed child clinical psychologist and a board-certified child psychiatrist, who utilized a structured diagnostic interview, clinician observations, as well as additional parent and teacher ratings and comments to arrive independently at their best estimate opinion. Their consensus opinion becomes the gold standard in each case.

In addition, using a multitrait multimethod structural equation modeling approach, we provide a critical evaluation of how much additional raters (i.e., mother, father, teacher) can increase validity of ADHD diagnosis and whether specific raters (e.g., teachers, mothers) outperform other raters (e.g., parents, fathers). As a secondary check on the effectiveness of the winning method, we consider the issue of the potential redundant information in the ADHD symptom list, one of the longest symptom lists of any disorder in the DSM. We include herein what is, to our knowledge, the first test using these methods of whether the nine symptoms per symptom domain can be cut in half with comparable validity (Kessler et al., 2010).

Method

Participants

Participants were 725 children (55.3% male), 6 to 17 years old ($M = 10.82$; $SD = 2.33$), and their primary caregivers ($N = 675$

for mothers and $N = 476$ for fathers) and teachers ($N = 629$), recruited from the community and deliberately oversampled for ADHD. After our evaluation, 330 children (45.5% of the sample) exhibited clinically significant (or diagnostic) levels of ADHD symptoms. The most common child clinical problems were oppositional defiant disorder ($n = 139$ [19.2%] in full sample; $n = 97$ [29.4%] in those with ADHD), learning disorders ($n = 98$ [13.5%] in full sample; $n = 60$ [18.2%] in those with ADHD), generalized anxiety disorder ($n = 55$ [7.6%] in full sample; $n = 31$[9%] in those with ADHD), and major depressive disorder ($n = 33$ [4.6%] in full sample; $n = 26$ [7.9%] in those with ADHD). The percentage of the ethnic minority or Hispanic were similar to the local community (28.5%) and family income ranged from $0 to $600,000 per year ($M = $67,550, $SD = 47,323$), as reported by parents. Children came from 426 families; 299 families had two children in the study, and the nonindependence of sibling data was handled statistically, as explained later. Approximately 265 families (60%) lived together (vs. being divorced or separated). All families completed parent written informed consent and child written informed assent, and study procedures were approved by the university Institutional Review Board, as well as were consistent with the guidelines of the American Psychological Association and National Institutes of Health.

Recruitment and Identification

To avoid the well-known inferential biases that attend upon clinic-based enrollment, a community-based recruitment strategy was used, with mass mailings to parents in local school districts, public advertisements, as well as flyers at local clinics. Families who volunteered then passed through a standard multigate screening process to identify cases and noncases eligible for the study based on standard DSM–IV criteria (in use at the time of recruitment). At Stage I, all families were screened by phone to rule out youth prescribed long-acting psychotropic medication (e.g., antidepressants), neurological impairments, seizure history, head injury with loss of consciousness, other major medical conditions, or a prior diagnosis of mental retardation or autistic disorder, as reported by parent. Approximately 20% of families were screened out at this point.

At Stage II, parents and teachers of remaining eligible youth completed several standardized rating scales, including the ADHD Rating Scale (ADHD-RS; DuPaul, Power, Anastopolous, & Reid, 1998). In addition, one parent completed a structured clinical interview, Kiddie Schedule for Affective Disorders and Schizophrenia (KSADS-E; Puig-Antich & Ryan, 1986), modified for DSM–IV, to ascertain symptom presence, onset, duration, and impairment (in the case of siblings, often, but not always, the same parent for both siblings). Exactly 134 children in the current sample (40% of the ADHD youth) had ever been prescribed psychostimulant medication, similar to population estimates in this age range (Froehlich et al., 2007). Parents and teachers were instructed to rate children's behavior when not taking psychostimulant medication, although it is recognized that parents may be more able to do this than teachers.

The data from the interview and parent and teacher rating scales were then presented to a clinical diagnostic team consisting of a board-certified child psychiatrist and licensed clinical child psychologist to implement a best-estimate diagnostic procedure. Their agreement rates were acceptable for ADHD diagnosis ($\kappa \geq .89$). This best-estimate diagnostic procedure with attendant symptom counts was utilized as a "gold-standard" criterion to examine predictive validity.

Measures

ADHD symptoms. For primary analysis, maternal, paternal, and teacher report on ADHD symptoms was obtained on the ADHD-RS (DuPaul et al., 1998), a common method used by researchers and clinicians. Here, each ADHD symptom is rated using a 4-point scale ranging from 0 (*never, or rarely*) to 3 (*very often*). Each informant provided ratings for all of the 18 DSM–IV symptoms (i.e., nine inattentive, nine hyperactive-impulsive symptoms).

Executive function. All children completed a neuropsychological testing battery after a minimum washout period of 24 hr for short-acting preparations and 48 hr for long-acting preparations (washout range 24–152 hr, M = 58 hr). The testing battery included tasks chosen to assess a variety of neuropsychological domains deemed especially relevant to ADHD. The neuropsychological battery and its factor analysis are described in Nikolas and Nigg (2013). They were administered in a fixed order as follows:

(1) *Working memory: Stars task.* We developed a computerized task modeled on Engle (2002). This was a dual task working memory assessment; on each trial of the task, participants either counted objects or remembered their location, while remembering a changing rule and keeping track of the sum.

(2) *Memory span and working memory: Spatial span.* Children completed a computerized version of the spatial working memory task (see Martinussen & Tannock, 2006) to examine visuospatial span and working memory capabilities, which required them to recall and reproduce the location of up to 9 objects in forward or backward order

(3) *Memory span and working memory: Digit span.* Youth completed the WISC-IV Digit Span task to assess verbal span (forward) and working memory (backward).

(4) *Interference control: DKEFS color-word interference.* This subtest from the Delis-Kaplan Executive Function System (DKEFS; Delis, Kaplan, & Kramer, 2001) was administered to assess interference control; it is similar to the classic Stroop task except with four conditions: color naming of nonword patches, reading color names in black ink, an inhibition condition (incongruent color word and ink color, name the color), and a switch condition in which some items must be read rather than named.

(5) *Response suppression/inhibition: Stop task.* The stop task (Logan, 1994) was administered to assess response inhibition; it requires the suppression of a prepotent motor response.

(6) *Reaction time variability.* The within-child variability of the reaction time (RT) on the Go Response trials from the stop task was retained as a measure of response variability.

(7) *Signal detection (putative indicator of arousal): Continuous performance task.* A version of the identical pairs continuous performance task (Cornblatt, Risch, Faris, Friedman, & Erlenmeyer-Kimling, 1988) similar to that used by Halperin, Sharma, Greenblatt, and Schwartz (1991) was used to examine vigilance and sustained attention.

(8) *Temporal information processing: Tapping task.* A computerized tapping task was administered to assess temporal information processing abilities (Toplak, Dockstader, & Tannock, 2006).

(9) *Processing speed and set shifting: DKEFS trailmaking task.* The DKEFS trailmaking task (Delis et al., 2001) was administered to assess cognitive-control and set-shifting abilities.

As described in Nikolas and Nigg (2013), a seven-factor neuropsychological model (factors labeled as inhibition, working memory, processing speed, memory span, response variability, arousal, and temporal information processing) exhibited the best fit, compared with competing models. Three factor scores (working memory, speed, inhibition) deemed most relevant to ADHD based on prior literature (Barkley, 1997; Nigg, 2006) were retained for the present study analyses to serve as further cross-validation evidence. They were standardized and summed to form a composite of executive function problems often associated with ADHD.

Data Analysis

Structural equation modeling (SEM) was conducted in Mplus (version 7.11) using theta parameterization and weighted least squares means and variance adjusted (WLSMV) estimation, as recommended for ordinal items (i.e., symptoms; Muthen & Muthen, 2013). The dependency among data from siblings was accounted for using the clustering feature of Mplus. This function adjusts standard error estimates based on the intraclass correlation of the data. Full information maximum likelihood was used to address missingness (7% for maternal report, 13% for teacher report; 35% for father report). This data was missing due to some informants being unavailable or choosing not to complete questionnaires. Some teacher ratings were missing due to data collection continuing in summers.

Our model of ADHD symptom ratings is illustrated in Figure 1. To reflect the presumptive model in DSM-5, we fitted a model with two distinct but correlated ADHD factors. The magnitude of this correlation provides information about the discriminant validity of the two factors. A correlation of 1 would indicate that the two factors are redundant, whereas correlations less than 1 indicate that children can have distinct profiles with higher scores on one factor than the other. Each ADHD factor (i.e., inattention and hyperactivity-impulsivity) was based on the three rater-specific factors for ratings by the mother, father, and teacher.

One major challenge in modeling of multiple-method (or rater) data is model identification (Schimmack, 2010). A minimum of three raters are needed to estimate validity coefficients for each rater, but this model makes the restrictive assumption that rater biases are independent. This assumption is typically violated for ratings by mothers and fathers, who often show a common rating bias (Zou, Schimmack, & Gere, 2013). To examine multimethod data with partially correlated method variances, a minimum of four methods are needed. To obtain a fourth method, we capitalized on the well-established finding that ADHD symptoms are correlated with performance on executive function tasks (Barkley, 1997; Nikolas & Nigg, 2013; Pennington & Ozonoff, 1996; Willcutt, Doyle, Nigg, Faraone, & Pennington, 2005). The advantage of executive function tasks is that these laboratory tasks do not share method variance with symptom ratings while they still pick up aspects of the hypothesized latent ADHD variable. Therefore, to facilitate model identification, the performance score on the executive function tasks was included in the model.

A potential concern with use of these tasks is that they are not a diagnostic criterion for ADHD per se. However, for our purposes of validating ADHD symptom ratings, it is not necessary that executive functioning is used for ADHD diagnosis. It is merely sufficient that performance on these tasks is correlated with ADHD symptoms and that this correlation is not spurious due to rating biases. Similarly, an objective measure of height could be used to validate self-reports of weight. With a true correlation of $r = .50$ between height and weight, self-reports of weight are more valid measures of weight if they correlate more strongly with height. Following this logic, utilization of child performance on executive function tasks in the SEM allowed us to compare the validity of ADHD symptom ratings by different raters, especially teachers versus parents. Because the model was identified by means of having four indicators, we were able to allow the residual variances in the inattention and hyperactivity-impulsivity factors of mothers and fathers to be freely correlated, rather than requiring them to be uncorrelated, and thus address their nonindependence.

Following common procedures for multimethod data, our model also allowed for correlations between the residual variances of the inattentive and hyperactive-impulsive factors by the same rater (e.g., teacher inattention and teacher hyperactivity-impulsivity; Kenny & Kashy, 1992). The rater-specific factors were defined by the nine symptom ratings for each dimension. We also allowed for symptom-specific correlations among the three raters. These correlations measure agreement between raters on a specific item after controlling for general agreement. Finally, given the large age range of the sample and the possibility that age influenced model results, we modeled the effect of child age on executive function and ADHD symptoms.

To examine the validity of different scoring algorithms, we estimated the correlation between the latent ADHD factors in Figure 1 and average, "or," and "and" manifest scale scores by adding scale scores to the model, regressing scale scores onto the defining items, and estimating the total correlation between the latent factors and the manifest scale scores. This correlation can be interpreted as a validity coefficient, and the square of this coefficient is an estimate of the amount of valid variance in scale scores (Campbell & Fiske, 1959; Schimmack, 2010). This procedure was carried out separately for the inattention factor and the hyperactivity-impulsivity factor. We also examined

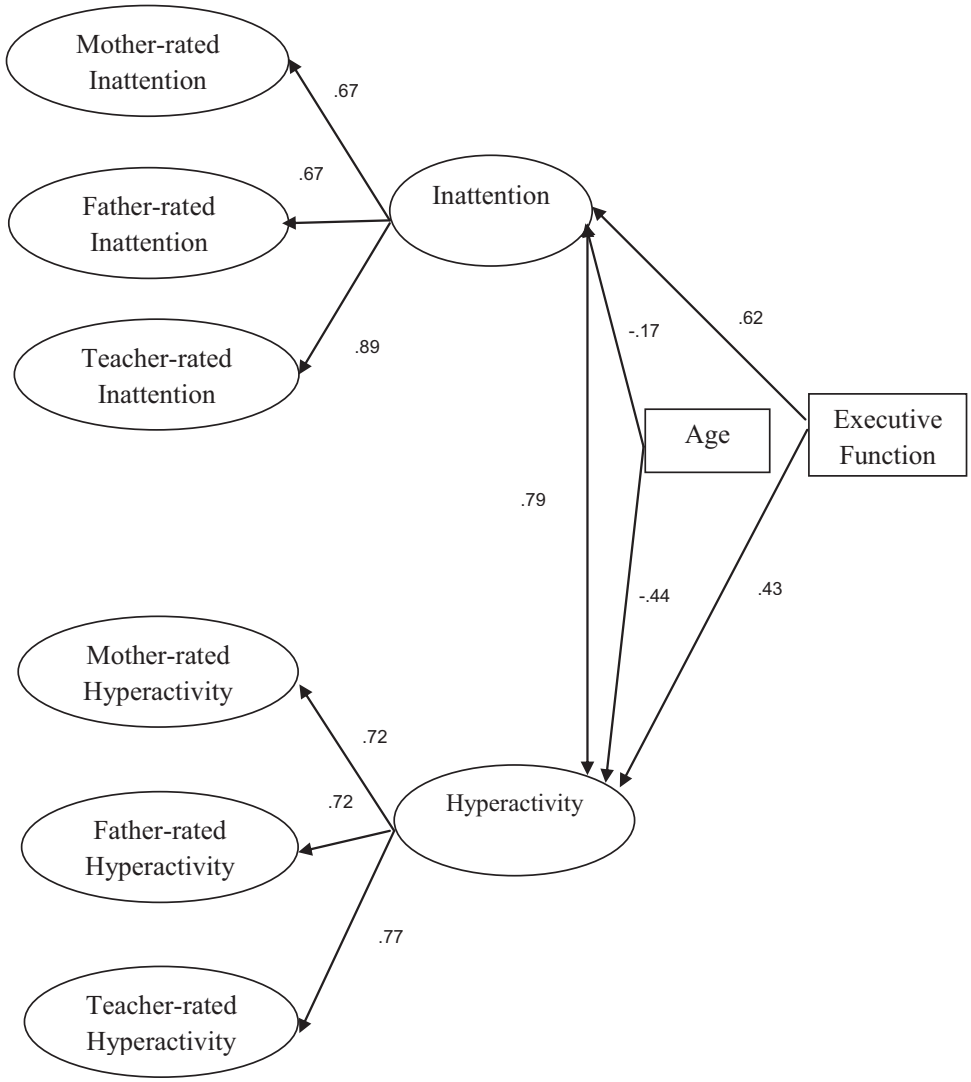

Figure 1. Simplified SEM of mother, father, and teacher ADHD symptom ratings. All loadings significant at $p < .01$.

whether the nine-item scales for each symptom dimension can be shortened without a loss of validity. Finally and critically, we compared the various scoring algorithms in terms of their ability to predict the final diagnosis of a diagnostic team.

Results

We fitted the theoretical model in Figure 1. The initial model had acceptable model fit using standard criteria of RMSEA < .06 and CFI > .95 (Hu & Bentler, 1999; McDonald & Ho, 2002). We next imposed some equality constraints on the model to create a more parsimonious model. First, we constrained item loadings to be equal (i.e., fixed the unstandardized loadings for each inattentive item and for each hyperactive-impulsive item to one). Model fit remained approximately the same, and inspection of modification indices revealed no items that consistently decreased model fit. This finding suggests that items have practically equal validity as indicators of ADHD. We also constrained the factor loadings for mothers and fathers to be equal, and model fit was unchanged, indicating that mothers and fathers are equally valid raters of ADHD symptoms. Fit of the final model was RMSEA = .042; 90% CI [.04, .044]; CFI = .97. The model parameters of the final model are shown in Figure 1. To check how the presence of missing data influenced the model, we reran it using listwise deletion with largely unchanged results (i.e., RMSEA of .041; 90% CI [.037, .045]; CFI = .97).

Age was a weak predictor of the inattention factor and only a slightly stronger predictor of hyperactivity-impulsivity. Weak age effects are to be expected because raters are asked to rate symptoms relative to other children of the same age. Figure 1 also shows that inattention and hyperactivity are strongly related factors, $r = .79$. However, the correlation is significantly lower than one, providing support for the hypothesis that ADHD is a multidimensional disorder with correlated factors.

The validity coefficients (i.e., loading of rater-specific factors on the general factors) show higher validity for teacher's inattention ratings than parents' inattention ratings. This could reflect the advantage of teachers to observe children in more demanding situations that require controlled attention. For hyperactivity-impulsivity, validity coefficients for parents and teachers were practically equivalent. In general, the validity coefficients are high and indicate that more than 50% of the variance in a rater-specific factor reflects variation in the general factor.

The factor loadings of individual symptom items on their respective factors were all very high (.88 for inattention and .85 for hyperactivity-impulsivity). This finding shows that raters respond to these items in similar ways and is consistent with our assumption that item responses are predominantly based on a general dimension of symptom severity and to a lesser extent on symptom specific information. The high factor loadings leave relatively little residual variance that could produce symptom specific agreement between raters (this is because some of the residual variance is simply random measurement error as well).

This impression is corroborated by an inspection of the residual correlations among ratings of the same item by the three raters (see Table 1). Rater agreement in specific symptoms would require convergent validity across all three raters. However, only three or four of nine items within ADHD symptom domain show significant convergent validity across parent and teacher ratings for inattention and hyperactivity-impulsivity. Lack of evidence for convergent validity for the other items raises concerns about scoring algorithms that focus on individual items to diagnose ADHD.

Assessing ADHD: Use of Different Scoring Algorithms

To compare relative performance of different scoring algorithms, scale composites of inattention and hyperactivity-impulsivity were generated across mother, father, and teacher ratings of ADHD symptoms using *average, or,* and *and* algorithm approaches. These composites were included in the model and regressed on all items. In addition, a latent sum score was created by regressing all items on a latent variable with fixed coefficients of 1 and setting the residual variance at zero. This variable is equivalent to a scale that simply sums all items, but because this variable is perfectly dependent on the items, it is necessary to use a latent variable. As causal arrows flow from the latent factors through the observed item ratings to the scales, it is possible to obtain the total indirect effect from the latent variables to the scale composites.

These total effects are reported in Table 2. The average inattention and hyperactivity-impulsivity composites exhibited stronger associations with the inattentive and hyperactivity-impulsivity latent factors than the composites using the or algorithm and the and algorithm. The nonoverlapping confidence intervals show that the averaging method is significantly more strongly related to the latent factors than the or and the and algorithms. Squaring these path coefficients provides an estimate of the amount of valid variance for the various scoring methods when the latent factors are assumed to be estimates of the true variance in inattention and hyperactivity-impulsivity (Schimmack, 2010). The amount of valid variance ranges from 72% for the averaging algorithm to 30% using the and algorithm for the hyperactivity-impulsivity scale and 41% using the and algorithm for the inattention scale.

Exploration of Short Version of ADHD Scale

The utility of a short average composite was also evaluated. Based on relative proportion frequencies, we chose those items that seemed to best differentiate between high and low levels of ADHD symptoms (i.e., had more nearly equal proportion of symptom endorsement across severity ratings within each rater). We were left with four items on each symptom domain scale ("close attention," "follow through," "organization," and "loses things" for inattention and "fidgets," "talks a lot," "blurts," and "interrupts, intrudes" for hyperactivity-impulsivity). As shown in Table 2, these short average composites performed comparably to the longer average composite and better than the long *or* and *and* algorithm composites in relation to their association with the latent inattention, $r = .86$, $p < .01$, and hyperactivity-impulsivity, $r = .82$, $p < .01$, factors.

Criterion Validity

The previous conclusions about validity rested on the assumption that the shared variance across raters represents variation in ADHD factors that underlie ADHD symptoms. Another way to validate scoring algorithms is to use the various scales to predict ADHD diagnosis by a diagnostic team. This criterion is often used as a gold standard to validate less costly measures of ADHD. To

Table 1
Residual Correlations Between Individual Symptom Items Across Raters

Items: Inattention *r*	M-F	P-T	Items: Hyperactivity	*r* M-F	P-T
Close attention	.39**	.10	Fidgets	.50**	.24*
Sustained attention	.38**	.07	Leaves seat	.63**	.52**
Listens	−.15	−.19	Runs, climbs	.85**	.34*
Follow through	.52**	.39**	Plays quietly	.39*	−.05
Organization	−.04	.16	"Driven by a motor"	.45**	−.09
Sustained mental effort	.74**	.32**	Talks a lot	.54**	.09
Loses things	−.27	.02	Blurts	−.02	.04
Easily distracted	.65**	.38**	Waiting turn	.44**	.04
Forgetful	.44**	.08	Interrupts, intrudes	.31*	.26*

Note. M-F = Mother-Father; P-T = Parent-Teacher.
* $p < .05$. ** $p < .01$.

Table 2
Correlations Between Latent ADHD Factors and Manifest ADHD Scales Using Or, And, *Average, and Short Algorithms*

Latent factors: Scales	Inattention *r*	Hyperactivity-impulsivity
Inattention *or* algorithm	.69 [.61–.76]	
Hyperactivity-impulsivity *or* algorithm		.66 [.60–.71]
Inattention *and* algorithm	.64 [.57–.71]	
Hyperactivity-impulsivity *and* algorithm		.55 [.50–.59]
Inattention average	.85 [.81–.88]	
Hyperactivity-impulsivity average		.85 [.81–.88]
Inattention short average	.86 [.83–.88]	
Hyperactivity-impulsivity short average		.82 [.79–.85]

Note. All correlations significant at $p < .01$. *or* algorithms, *and* algorithms, and averages utilized mother, father, and teacher ratings [95% confidence interval].

obtain a dichotomous score of ADHD as present or absent, we used an average of one as the cutoff point on a simple average of the short-scale for inattention and hyperactivity-impulsivity. It is important to realize that this validation test is potentially biased against the averaging algorithm and the *and* algorithm, because the diagnostic team most likely tended to follow convention and to apply the or algorithm in cases of diagnostic uncertainty. Nevertheless, the averaging algorithm outperformed the or algorithm, the better of the two traditional scoring algorithms, in diagnostic sensitivity and specificity (see Table 3). Both algorithms would have correctly identified children without ADHD in most instances (91%) compared with diagnostic gold standard. However, the or algorithm was much less likely to correctly identify ADHD compared with the diagnostic team and the average algorithm; in 32% of cases, the or algorithm did not identify ADHD when the diagnostic team did, compared with the averaging algorithm that only failed to correctly identify ADHD in 17% of instances. Further, the averaging algorithm outperformed the or algorithm in both positive and negative predictive power (.92 compared with .89 for positive predictive power [PPP]; .83 compared with .74 negative predicative power [NPP]), or the proportion of positive and negative test results that are true positives and negatives, respectively.

Discussion

Valid diagnosis of ADHD is crucial for optimal treatment of ADHD, yet validity of current standardized approaches to integration of multiple informant reports of ADHD symptoms remains an unresolved issue. Like its predecessor DSM–IV–TR, the DSM-5 urges integration of information from multiple sources for diagnosis of ADHD (American Psychiatric Association, 2013). However, there is practical disagreement about the optimal way to integrate information from multiple raters. In this study, we proposed and evaluated a standardized approach grounded in psychometric theories of multitrait multimethod data (Campbell & Fiske, 1959; Schimmack, 2010) that relied on an average of parent and teacher ratings and compared this approach to alternative scoring methods such as the *or* and *and* algorithms validated using gold standard best estimate diagnosis by a clinical team.

We fitted covariances among child ADHD symptom ratings by mothers, fathers, and one teacher in a structural equation model that postulated two latent ADHD symptom domains: inattention and hyperactivity-impulsivity, in line with the DSM-5 (American Psychiatric Association, 2013). This model exhibited acceptable fit to the data. Although this model exhibited acceptable fit to the data, it is important to point out this is not necessarily the best model of ADHD; rather, it happens to be consistent with DSM-5's implied conceptualization of ADHD. Other work, including our own (Martel, von Eye, & Nigg, 2010) and others (Toplak et al., 2009) suggests that ADHD may be best described using a bifactor model (vs. one-, two-, three-factor or second-order factor models) with a general ADHD factor and partially distinct specific factors of inattention and hyperactivity-impulsivity. Yet, such a model has not yet demonstrated clinical utility and so for simplicity was omitted here.

We then examined how much variance this general ADHD factor explained in manifest measures of ADHD that used different scoring methods. An unweighted average of nine inattentive and hyperactive-impulsive items rated by all three informants was most strongly related to the inattention and hyperactivity-impulsivity latent factors. A shorter scale with four inattentive and four hyperactive-impulsive items per rater was also strongly related to the inattentive and hyperactive-impulsive latent factors. The short average performed as well or better than the or algorithm in diagnostic sensitivity and specificity. This likely happened for two reasons. First, the or scoring method cannot reduce rater bias,

Table 3
Comparison of Scoring Algorithm Sensitivity and Specificity With Best Estimate Diagnostic Team ADHD Diagnosis

	Or algorithm diagnosis		Short scale diagnosis	
Diagnostic team diagnosis	No ADHD	ADHD	No ADHD	ADHD
No ADHD	235	22	260	26
ADHD	82	178	56	268
Sensitivity	.68		.83	
False negative rate	.32		.17	
Specificity	.91		.91	
False positive rate	.09		.09	
Positive predictive power (PPP)	.89		.92	
Negative predictive power (NPP)	.74		.83	

whereas averaging reduces rating biases that are unique to a single rater. Second, the and algorithm assumes that raters agree on specific items that are treated like unique symptoms, but our psychometric model shows that raters agree predominantly on general factors underlying symptom ratings and that many items show no agreement between teacher and parent ratings after controlling for general agreement. Based on these findings, we recommend consideration of utilization of an average composite in quantification of symptom counts in diagnosis of ADHD. Such an approach would advance research and clinical practice in the field by standardizing diagnostic practice (Voigt et al., 2007), although—of course—our results first need to be replicated and validated vis a vis impairment and—of course—may not always operate perfectly at the level of the individual child. Namely, we advocate symptoms to be averaged (or summed which is equivalent) at the symptom domain (inattention, hyperactivity-impulsivity) or overall diagnostic category (ADHD) level within reporter, and then an average taken across reporters to determine child symptom counts and diagnostic status.

Our study also provided new insights into other issues in the diagnosis of ADHD. First, in line with DSM-5, our model suggests that ADHD is a multidimensional construct. At the same time, our model showed that these two factors are strongly correlated with each other, suggesting a common etiology, consistent with recently supported bifactor models of ADHD (Martel et al., 2010; Toplak et al., 2009). Second, our model showed strong loadings of ratings by mothers, fathers, and teachers on the general factor, indicating that all three raters provide valid information about ADHD. Loadings for mother and father were high, even after allowing for shared rating biases between them, indicating that agreement is not just a method artifact. Because of the high agreement between mothers and fathers and shared method variance between them, it may be sufficient to obtain ratings from only one parent; either the mother or father. Furthermore, based on our results, teachers may be slightly better raters of inattentive symptoms than parents, perhaps because they are more easily able to compare individual child behavior with same-age peers in the classroom.

Third, we found rather modest evidence for agreement in ratings of specific items after controlling for agreement on the general factors. This finding raises concern about the interpretation of items as symptoms and scoring methods that add up individual symptoms. Instead, it seems most accurate to consider individual symptoms as probabilistic behavioral manifestations of ADHD severity that can be usefully summarized as a composite rating within symptom domain. It is, of course, possible that ADHD is a broad disorder that has distinct manifestations across settings via different constellations of symptoms (e.g., see Barkley, 1997; Nigg & Casey, 2005; Sonuga-Barke, Bitsakou, & Thompson, 2010). However, to examine this hypothesis, it will be necessary to develop new measures that demonstrate convergent validity in ratings of specific symptoms. A single item is likely to be insufficient to achieve this goal.

Of course, our study has a number of limitations that should be addressed in future research. Although the sample used here had substantial advantages in avoiding both the inferential biases of clinic referred samples and the lack of depth of population surveys, it will be important to extend these results to other populations and samples and to examine adults when self-report becomes more important. Because ADHD is a neurodevelopmental disorder, it will be particularly important for future work and replications to evaluate possible developmental change in this model across the life span, particularly due to recently expanded symptom item content in DSM-5 designed to account for age-related variability in symptoms. We did not have that content available for the present study. Although helpful for model stabilization and testing of relative importance of raters, the inclusion of executive function (EF) in the ADHD model might be considered a limitation since EF is not part of the diagnostic criteria for ADHD, EF tasks have shared variance, and can influence and be influenced by other child clinical comorbidities such as learning disorders (Miyake et al., 2000). SEM helps validate correlational structure and also helps to estimate reliability-corrected correlations; however, it is not comprehensive in that it assumes linearity and, like all correlational methods, does not demonstrate causality (Kline, 2011). In addition, our study had missing data, although we conducted secondary checks on this point with little effect on model fit. The diagnostic team saw parent and teacher ratings on the ADHD Rating Scale as one of several factors in determining gold standard ADHD diagnosis; this could have inflated correlations between rater report and gold standard case identification. An important direction for future work is further validation research with additional validation criteria (e.g., observational ratings of ADHD and/or impairment; treatment effectiveness; neurological correlates) and in clinical settings at the individual child level. It will also be important to compare a simple averaging algorithm with more complex scoring algorithms that take symptom profiles into account (De Los Reyes et al., 2009; Dirks et al., 2012). Finally, it is important to continue to critically evaluate choice of and utility of use of multiple informants (Kraemer et al., 2003). For example, we did not include child report, which may become important particularly during adolescence.

Despite these limitations, our study utilizes a large, community-based sample of well-characterized ADHD and non-ADHD youth that represents the spectrum of ADHD severity in the community and provides the first evaluation as to the utility of psychometric analysis to examine and increase the validity of ADHD diagnosis. The results provide empirical evidence against the use of traditional scoring methods of *or* or *and* algorithms. Given the need for valid diagnosis, practitioners are advised to use our scoring algorithm at least in combination with the traditional scoring algorithm at present. Discrepancies between reporters in ratings of ADHD symptoms are likely, at least in part, due to error variance and are best treated using an averaging approach, at least given currently available diagnostic measures.

References

Achenbach, T. M. (2011). Commentary: Definitely more than measurement error: But how should we understand and deal with informant discrepancies? *Journal of Clinical Child and Adolescent Psychology, 40,* 80–86. http://dx.doi.org/10.1080/15374416.2011.533416

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: American Psychiatric Association.

Barkley, R. A. (1997). Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of ADHD. *Psychological Bulletin, 121,* 65–94. http://dx.doi.org/10.1037/0033-2909.121.1.65

Bird, H. R., Gould, M. S., & Staghezza, B. (1992). Aggregating data from multiple informants in child psychiatry epidemiological research. *Journal of the American Academy of Child & Adolescent Psychiatry, 31,* 78–85. http://dx.doi.org/10.1097/00004583-199201000-00012

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105. http://dx.doi.org/10.1037/h0046016

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136,* 1092–1122. http://dx.doi.org/10.1037/a0021212

Cornblatt, B. A., Risch, N. J., Faris, G., Friedman, D., & Erlenmeyer-Kimling, L. (1988). The Continuous Performance Test, identical pairs version (CPT-IP): I. New findings about sustained attention in normal families. *Psychiatry Research, 26,* 223–238. http://dx.doi.org/10.1016/0165-1781(88)90076-5

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668–1674. http://dx.doi.org/10.1126/science.2648573

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan Executive Function System (DKEFS*). San Antonio, TX: The Psychological Corporation.

De Los Reyes, A. (2013). Strategic objectives for improving understanding of informant discrepancies in developmental psychopathology research. *Development and Psychopathology, 25,* 669–682. http://dx.doi.org/10.1017/S0954579413000096

De Los Reyes, A., Henry, D. B., Tolan, P. H., & Wakschlag, L. S. (2009). Linking informant discrepancies to observed variations in young children's disruptive behavior. *Journal of Abnormal Child Psychology, 37,* 637–652. http://dx.doi.org/10.1007/s10802-009-9307-3

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 131,* 483–509. http://dx.doi.org/10.1037/0033-2909.131.4.483

De Los Reyes, A., Youngstrom, E. A., Pabón, S. C., Youngstrom, J. K., Feeny, N. C., & Findling, R. L. (2011). Internal consistency and associated characteristics of informant discrepancies in clinic referred youths age 11 to 17 years. *Journal of Clinical Child and Adolescent Psychology, 40,* 36–53. http://dx.doi.org/10.1080/15374416.2011.533402

Dirks, M. A., Boyle, M. H., & Georgiades, K. (2011). Psychological symptoms in youth and later socioeconomic functioning: Do associations vary by informant? *Journal of Clinical Child and Adolescent Psychology, 40,* 10–22. http://dx.doi.org/10.1080/15374416.2011.533403

Dirks, M. A., De Los Reyes, A., Briggs-Gowan, M., Cella, D., & Wakschlag, L. S. (2012). Annual research review: Embracing not erasing contextual variability in children's behavior—Theory and utility in the selection and use of methods and informants in developmental psychopathology. *Journal of Child Psychology and Psychiatry, 53,* 558–574. http://dx.doi.org/10.1111/j.1469-7610.2012.02537.x

DuPaul, G. J., Power, T. J., Anastopolous, A. D., & Reid, R. (1998). *ADHD Rating Scale-IV: Checklists, norms, and clinical interpretation*. New York, NY: Guilford Press.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11,* 19–23. http://dx.doi.org/10.1111/1467-8721.00160

Froehlich, T. E., Lanphear, B. P., Epstein, J. N., Barbaresi, W. J., Katusic, S. K., & Kahn, R. S. (2007). Prevalence, recognition, and treatment of attention-deficit/hyperactivity disorder in a national sample of U.S. children. *Archives of Pediatrics & Adolescent Medicine, 161,* 857–864. http://dx.doi.org/10.1001/archpedi.161.9.857

Gomez, R. (2007). Australian parent and teacher ratings of the *DSM–IV* ADHD symptoms: Differential symptom functioning and parent-teacher agreement and differences. *Journal of Attention Disorders, 11,* 17–27. http://dx.doi.org/10.1177/1087054706295665

Gomez, R. (2008). Item response theory analyses of the parent and teacher ratings of the *DSM–IV* ADHD rating scale. *Journal of Abnormal Child Psychology, 36,* 865–885. http://dx.doi.org/10.1007/s10802-008-9218-8

Gomez, R., Vance, A., & Gomez, A. (2011). Item response theory analyses of parent and teacher ratings of the ADHD symptoms for recoded dichotomous scores. *Journal of Attention Disorders, 15,* 269–285. http://dx.doi.org/10.1177/1087054709356404

Halperin, J. M., Sharma, V., Greenblatt, E., & Schwartz, S. T. (1991). Assessment of the continuous performance test: Reliability and validity in a non-referred sample. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3,* 603–608. http://dx.doi.org/10.1037/1040-3590.3.4.603

Holmbeck, G. N., Thill, A. W., Bachanas, P., Garber, J., Miller, K. B., Abad, M., . . . Zukerman, J. (2008). Evidence-based assessment in pediatric psychology: Measures of psychosocial adjustment and psychopathology. *Journal of Pediatric Psychology, 33,* 958–980. http://dx.doi.org/10.1093/jpepsy/jsm059

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112,* 165–172.

Kessler, R. C., Green, J. G., Adler, L. A., Barkley, R. A., Chatterji, S., Faraone, S. V., . . . Van Brunt, D. L. (2010). Structure and diagnosis of adult attention-deficit/hyperactivity disorder: Analysis of expanded symptom criteria from the Adult ADHD Clinical Diagnostic Scale. *Archives of General Psychiatry, 67,* 1168–1178. http://dx.doi.org/10.1001/archgenpsychiatry.2010.146

Kline, R. B. (2011). *Principles and practice of structural equation modelling* (3rd ed.). New York, NY: Guilford Press.

Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *The American Journal of Psychiatry, 160,* 1566–1577. http://dx.doi.org/10.1176/appi.ajp.160.9.1566

Lahey, B. B., Applegate, B., McBurnett, K., Biederman, J., Greenhill, L., Hynd, G. W., . . . Shaffer, D. (1994). DSM–IV field trials for attention deficit hyperactivity disorder in children and adolescents. *The American Journal of Psychiatry, 151,* 1673–1685. http://dx.doi.org/10.1176/ajp.151.11.1673

Loeber, R., Green, S. M., Lahey, B. B., & Stouthamer-Loeber, M. (1989). Optimal informants on childhood disruptive behaviors. *Development and Psychopathology, 1,* 317–337. http://dx.doi.org/10.1017/S095457940000050X

Logan, G. D. (1994). A user's guide to the stop signal paradigm. In D. Dagenbach & T. H. Carr (Eds.), *Inhibitory processes in attention, memory, and language* (pp. 189–239). San Diego, CA: Academic Press.

Marcus, D. K., & Barry, T. D. (2011). Does attention-deficit/hyperactivity disorder have a dimensional latent structure? A taxometric analysis. *Journal of Abnormal Psychology, 120,* 427–442. http://dx.doi.org/10.1037/a0021405

Martel, M. M., von Eye, A., & Nigg, J. T. (2010). Revisiting the latent structure of ADHD: Is there a "g" factor? *Journal of Child Psychology and Psychiatry, 51,* 905–914. http://dx.doi.org/10.1111/j.1469-7610.2010.02232.x

Martinussen, R., & Tannock, R. (2006). Working memory impairments in children with attention-deficit hyperactivity disorder with and without comorbid language learning disorders. *Journal of Clinical and Experimental Neuropsychology, 28,* 1073–1094. http://dx.doi.org/10.1080/13803390500205700

McDonald, R. P., & Ho, M. H. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7,* 64–82. http://dx.doi.org/10.1037/1082-989X.7.1.64

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41,* 49–100. http://dx.doi.org/10.1006/cogp.1999.0734

Muthen, L. K., & Muthen, B. O. (2013). *Mplus user's guide* (4th ed.). Los Angeles, CA: Author.

Nigg, J. T. (2006). *What causes ADHD? Understanding what goes wrong and why*. New York, NY: Guilford Press.

Nigg, J. T., & Casey, B. J. (2005). An integrative theory of attention-deficit/ hyperactivity disorder based on the cognitive and affective neurosciences. *Development and Psychopathology, 17,* 785–806. http://dx.doi.org/10.1017/S0954579405050376

Nikolas, M. A., & Nigg, J. T. (2013). Neuropsychological performance and attention-deficit hyperactivity disorder subtypes and symptom dimensions. *Neuropsychology, 27,* 107–120. http://dx.doi.org/10.1037/a0030685

Pelham, W. E., Jr., Fabiano, G. A., & Massetti, G. M. (2005). Evidence-based assessment of attention deficit hyperactivity disorder in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 34,* 449–476. http://dx.doi.org/10.1207/s15374424jccp3403_5

Pennington, B. F., & Ozonoff, S. (1996). Executive functions and developmental psychopathology. *Child Psychology & Psychiatry & Allied Disciplines, 37,* 51–87. http://dx.doi.org/10.1111/j.1469-7610.1996.tb01380.x

Piacentini, J. C., Cohen, P., & Cohen, J. (1992). Combining discrepant diagnostic information from multiple sources: Are complex algorithms better than simple ones? *Journal of Abnormal Child Psychology, 20,* 51–63. http://dx.doi.org/10.1007/BF00927116

Puig-Antich, J., & Ryan, N. (1986). *Kiddie schedule for affective disorders and schizophrenia*. Pittsburgh, PA: Western Psychiatric Institute.

Schimmack, U. (2010). What multi-method data tell us about construct validity. *European Journal of Personality, 24,* 241–257. http://dx.doi.org/10.1002/per.771

Solanto, M. V., & Alvir, J. (2009). Reliability of *DSM–IV* symptom ratings of ADHD: Implications for DSM-V. *Journal of Attention Disorders, 13,* 107–116. http://dx.doi.org/10.1177/1087054708322994

Sonuga-Barke, E., Bitsakou, P., & Thompson, M. (2010). Beyond the dual pathway model: Evidence for the dissociation of timing, inhibitory, and delay-related impairments in attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 49,* 345–355.

Toplak, M. E., Dockstader, C., & Tannock, R. (2006). Temporal information processing in ADHD: Findings to date and new methods. *Journal of Neuroscience Methods, 151,* 15–29. http://dx.doi.org/10.1016/j.jneumeth.2005.09.018

Toplak, M. E., Pitch, A., Flora, D. B., Iwenofu, L., Ghelani, K., Jain, U., & Tannock, R. (2009). The unity and diversity of inattention and hyperactivity/impulsivity in ADHD: Evidence for a general factor with separable dimensions. *Journal of Abnormal Child Psychology, 37,* 1137–1150. http://dx.doi.org/10.1007/s10802-009-9336-y

Valo, S., & Tannock, R. (2010). Diagnostic instability of *DSM–IV* ADHD subtypes: Effects of informant source, instrumentation, and methods for combining symptom reports. *Journal of Clinical Child and Adolescent Psychology, 39,* 749–760. http://dx.doi.org/10.1080/15374416.2010.517172

Voigt, R. G., Llorente, A. M., Jensen, C. L., Fraley, J. K., Barbaresi, W. J., & Heird, W. C. (2007). Comparison of the validity of direct pediatric developmental evaluation versus developmental screening by parent report. *Clinical Pediatrics, 46,* 523–529. http://dx.doi.org/10.1177/0009922806299100

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V., & Pennington, B. F. (2005). Validity of the executive function theory of attention-deficit/hyperactivity disorder: A meta-analytic review. *Biological Psychiatry, 57,* 1336–1346. http://dx.doi.org/10.1016/j.biopsych.2005.02.006

Zou, C., Schimmack, U., & Gere, J. (2013). The validity of well-being measures: A multiple-indicator-multiple-rater model. *Psychological Assessment, 25,* 1247–1254. http://dx.doi.org/10.1037/a0033902

(*Appendix follows*)

Appendix

Annotated Mplus Modeling Syntax

```
Title: Final Model of ADHD symptom ratings by multiple raters;
m = mother
f = father
t = teacher
Data:
! read the input file from the same directory as the syntax file
FILE IS inputinc.dat;
variable:
! List of names of the variables in the raw data file
Names are
famid ! family ID; used for cluster command to allow for intra-class correlation among children from the same family
agev1 ! age
m1-m18 f1-f18 t1-t18 ! ratings of the 18 items by mother (m) father (f) teacher (t)
or1 or2 and1 and2 ! scale scores based on OR or AND scoring methods
wmem speed inhib ! scores on three working memory tasks
incgr ! income group
MISSING IS famid-inhib (-99)
! missing value for all variables is −99
USEVAR
agev1
m1 m3 m5 m7 m9 m11 m13 m15 m17
f1 f3 f5 f7 f9 f11 f13 f15 f17
t1 t3 t5 t7 t9 t11 t13 t15 t17
m2 m4 m6 m8 m10 m12 m14 m16 m18
f2 f4 f6 f8 f10 f12 f14 f16 f18
t2 t4 t6 t8 t10 t12 t14 t16 t18
wmem speed inhib;
! exf is a composite of the working memory tasks that is created with the DEFINE command below
! exf has to be listed last because it was created with the DEFINE command
CLUSTER = famid; ! specifies the dependency among family members
CATEGORICAL m1-t18; ! specifies that all symptom ratings are categorical variables
ANALYSIS:
model = nocov; ! suppresses any default correlations among factors; all correlations are fixed at zero unless specified as free parameters
type = complex; ! complex data because data are not independent
PARAMETERIZATION = THETA; ! THETA parameterization for categorical variables (see MPLUS website for further details)
!define:
!exf = wmem + speed + inhib; ! create working memory measure by averaging scores on three tasks; gives same results as creating a latent variable
model:
exf by wmem speed inhib;
!Factor 1
mf1 by m1-m17*2(il1); ! create factor for mothers' ratings of attention-deficit items; constrain loadings across items = estimate single parameter (iL1)
ff1 by f1-f17*2(il1); ! create factor for fathers' ratings of attention-deficit items; constrain loadings across items = estimate single parameter (iL1)
tf1 by t1-t17*2(il1); ! create factor for teachers' ratings of attention-deficit items; constrain loadings across items = estimate single parameter (iL1)
! same parameter for mothers, fathers, and teachers means constrained loadings for all raters and items
m1-m17 pwith f1-f17*0(pp11-pp19); ! allow for correlations among residuals of the same item between mothers and fathers
m1-m17 pwith t1-t17*0(pt11-pt19); ! allow for correlations among residuals of the same item between mothers and teachers
f1-f17 pwith t1-t17*0(pt11-pt19); ! allow for correlations among residuals of the same item between fathers and teachers
! significant parameter estimates indicate convergent validity for the unique variance in specific items above and beyond convergent validity at the factor level
!Factor 2
mf2 by m2-m18*1.7(il2); ! same as for factor 1
ff2 by f2-f18*1.7(il2); ! same as for factor 1
tf2 by t2-t18*1.7(il2); ! same as for factor 1
m2-m18 pwith f2-f18*0(pp21-pp29); ! same as for factor 1
m2-m18 pwith t2-t18*0(pt21-pt29); ! same as for factor 1
f2-f18 pwith t2-t18*0(pt21-pt29); ! same as for factor 1
! same as for factor 1
! Create ADHD factors
! hierarchical model with ADF1 and ADF2 as a higher order factor based on rater-specific factors
! ADF1 = Attention-related symptoms, ADF2 = hyperactivity related factors
ADF1 by mf1*1(pl1); ! define the attention-deficit factor by the mothers' factor; loading constrained across parents
ADF1 by ff1*1(pl1); ! define the attention-deficit factor by the fathers' factor; loadings constrained across parents
ADF1 by tf1*1(tl1); ! define the attention-deficit factor by the teachers' factor; loading can be different for teacher
ADF2 by mf2*1(pl2); ! same as factor 1
ADF2 by ff2*1(pl2); ! same as factor 1
ADF2 by tf2*1(tl2); ! same as factor 1
mf1 ff1*(pf1res); ! name residual variances of parents' factors, mf1, df1 and constrain residual variances
```

(*Appendix continues*)

mf2 ff2*(pf2res); ! name residual variances of parents' factors, mf2, df2, and constrain residual variances
tf1*(tf1res); ! name residual variances of teacher factor, tf1.
tf2*(tf2res); ! name residual variances of teacher factor, tf2.
mf1 with mf2*(srb); ! shared rater bias for attention and hyperactivity symptoms for mothers;
ff1 with ff2*(srb); ! shared rater bias for attention and hyperactivity symptoms for fathers;
tf1 with tf2*(tsrb); ! shared rater bias for attention and hyperactivity symptoms for teachers;
! same parameter label means that this is constrained.
mf1 mf2 with ff1 ff2; ! allow for correlated rating biases between parents' ratings of ADHD
ADF1 with ADF2 *.7; ! allow for correlation between the two ADHD factors
adf1 on agev1*-.1(age1); ! regress attention factor on age
adf2 on agev1*-.1(age2); ! regress hyperactivity factor on age
exf on agev1*-.5; ! regress executive functioning factor on age
agev1*(agevar); ! name age variance
ADF1*(adf1res); ! name residual variances in attention-factor
ADF2*(adf2res); ! name residual variances in hyperactivity factor
ADF1 with exf*; ! allow for correlation between attention factor and executive functioning factor; correlation does not make assumptions about causality
ADF2 with exf*; ! based on non-significant relationship in a previous model, this parameter is fixed to zero
MODEL CONSTRAINT:
0 = agevar*age1**2 + adf1res – 1;
! agevar*age1**2 is the variance in adf1 that is explained by age;
! adf1res is the residual variance in adf1 that is not explained by age;
! the sum of agevar*age1**2 + adf1res is the total variance in adf1;
! 0 = total variance - 1 is used to scale the variance in adf1 to 1;
0 = agevar*age2**2 + adf2res - 1;
! same as for adf1
0 = pl1**2 + pf1res - 1;
! scaling the variances in mf1 and df1 to 1
0 = pl2**2 + pf2res - 1;
! scaling the variances in mf2 and df2 to 1
0 = tl1**2 + tf1res - 1;
! scaling the variances in tf1 to 1
0 = tl2**2 + tf2res - 1;
! scaling the variances in tf2 to 1
output: SAMP MOD(all 10) RESIDUAL STANDARDIZED CINT TECH1 TECH4

Received March 21, 2014
Revision received December 9, 2014
Accepted December 15, 2014 ■