Title
Item Response Theory Analysis of ADHD Symptoms in Children With and Without ADHD

Journal
Assessment, 23(6)

ISSN
1073-1911

Authors
Li, James J
Reise, Steven P
Chronis-Tuscano, Andrea
et al.

Publication Date
2016-12-01

DOI
10.1177/1073191115591595

Peer reviewed

Published in final edited form as:
Assessment. 2016 December ; 23(6): 655–671. doi:10.1177/1073191115591595.

Item Response Theory Analysis of ADHD Symptoms in Children With and Without ADHD

James J. Li[1], **Steven P. Reise**[2], **Andrea Chronis-Tuscano**[3], **Amori Yee Mikami**[4], and **Steve S. Lee**[2]

[1]University of Wisconsin, Madison, WI, USA

[2]University of California, Los Angeles, CA, USA

[3]University of Maryland, College Park, MD, USA

[4]University of British Columbia, Vancouver, British Columbia, Canada

Abstract

Item response theory (IRT) was separately applied to parent- and teacher-rated symptoms of attention-deficit/hyperactivity disorder (ADHD) from a pooled sample of 526 six- to twelve-year-old children with and without ADHD. The dimensional structure ADHD was first examined using confirmatory factor analyses, including the bifactor model. A general ADHD factor and two group factors, representing inattentive and hyperactive/impulsive dimensions, optimally fit the data. Using the graded response model, we estimated discrimination and location parameters and information functions for all 18 symptoms of ADHD. Parent- and teacher-rated symptoms demonstrated adequate discrimination and location values, although these estimates varied substantially. For parent ratings, the test information curve peaked between −2 and +2 *SD*, suggesting that ADHD symptoms exhibited excellent overall reliability at measuring children in the low to moderate range of the general ADHD factor, but not in the extreme ranges. Similar results emerged for teacher ratings, in which the peak range of measurement precision was from −1.40 to 1.90 *SD*. Several symptoms were comparatively more informative than others; for example, *is often easily distracted* ("Distracted") was the most informative parent- and teacher-rated symptom across the latent trait continuum. Clinical implications for the assessment of ADHD as well as relevant considerations for future revisions to diagnostic criteria are discussed.

Keywords

item response theory; attention-deficit/hyperactivity disorder; psychometrics

---

Attention-deficit/hyperactivity disorder (ADHD) is a highly heterogeneous, prevalent youth-onset disorder characterized by developmentally aberrant and impairing levels of inattention

---

Reprints and permissions: sagepub.com/journalsPermissions.nav

Corresponding Author: James J. Li, Department of Psychology, Brogden Hall, 1202 West Johnson Street, Madison, WI 53706, USA. james.li@wisc.edu.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

and/or hyperactivity-impulsivity. ADHD affects approximately 8% to 12% of school-aged children worldwide (Froehlich et al., 2007), and prospectively predicts a wide range of negative adolescent and adult outcomes including academic and social failure, criminality, substance abuse, suicidality, and neuropsychological impairment (Barkley & Fischer, 2010; Chronis-Tuscano et al., 2010; Langley et al., 2010).

According to the *Diagnostic and Statistical of Mental Disorders* (5th ed.; *DSM-5*; American Psychiatric Association, 2013), there are two putative dimensions of ADHD: hyperactivity/impulsivity (HI), represented by symptoms such as poor impulse control, difficulty sitting still, and fidgeting or squirming; and inattention (IA), represented by symptoms including difficulty sustaining attention, carelessness, and disorganization. These dimensions have been studied extensively across multiple empirical models (Nigg, 2012). For example, factor analytic studies of parent ratings of ADHD symptoms fit a two-factor model, reflecting dimensions of IA and HI that were invariant across gender (Collett, Crowley, Gimpel, & Greenson, 2000). These dimensions have been consistently replicated across different samples (e.g., Gomez, Burns, Walsh, & De Moura, 2003; Pillow, Pelham, Hoza, Molina, & Stultz, 1998), including in clinic-referred and nonreferred samples of children (Lahey et al., 1988). Furthermore, twin designs suggest that IA and HI may be etiologically distinct (Nikolas & Burt, 2010; Willcutt, Pennington, & DeFries, 2000), with genetic influences varying according to whether dominant or additive genetic models were specified (Nikolas & Burt, 2010). Despite considerable evidence on the separability of these ADHD dimensions, emerging studies have demonstrated that the covariation between IA and HI may in fact be better accounted for by a general ADHD factor, with specific factors that contribute independent covariance related to the individual dimensions (Normand, Flora, Toplak, & Tannock, 2012; Toplak et al., 2012). For example, Normand et al. (2012) assessed 6- to 9-year-old school children across two time points and found evidence of a general factor of ADHD as well as two orthogonal factors represented by IA and HI. Moreover, these factors were invariant to parent versus teacher ratings, sex, and time of the assessment. The “bifactor” characterization of ADHD has become increasingly popular as it represents a promising approach toward understanding the heterogeneous presentation of ADHD. More studies are clearly required however, particularly given that previous bifactor studies of ADHD have focused exclusively on school-based populations.

Furthermore, there is evidence that ADHD exists on a continua of severity (Frazier, Youngstrom, & Naugle, 2007; Larsson, Anckarsater, Råstam, Chang, & Lichtenstein, 2012; Willcutt et al., 2012). Multiple data analytic techniques have uncovered empirically distinct groups of individuals based on their response probabilities to ADHD symptoms. For example, latent class analysis (LCA) of ADHD symptoms was reported in the National Longitudinal Study of Adolescent Health (Li & Lee, 2012b), with clinic-referred children (de Nijs, van Lier, Verhulst, & Ferdinand, 2007; Elia et al., 2009), as well as with U.S. and Australian twins (Rasmussen et al., 2002; Todd et al., 2001). These studies uncovered class distinctions based on severity (i.e., mild, moderate, and severe) and that also approximated various dimensions of ADHD. LCA of a large school sample of Brazilian youth yielded six classes (eight total) for boys and girls that were distinguished largely according to differences in the severity of combined-type ADHD (Rohde et al., 2001). IA and combined classes (mild and severe), as well as a purely hyperactive class were identified through LCA of ADHD using a population-based twin sample in the United States (Volk, Neuman, & Todd, 2005). Similarly, Todd et al. (2001) reported three mild and severe classes that overlapped with *DSM-IV* subtypes of ADHD among adolescent female twins. Overall, LCA studies of ADHD converge around the centrality of a severity continuum with respect to ADHD in addition to the qualitatively distinct symptom profiles based on the dimensions of ADHD.

The conceptualization of ADHD as a continuous latent trait results in a number of important research and clinical questions. For example, which ADHD symptoms optimally (i.e., precisely) measure an individual's standing on the trait continuum? Should each symptom be weighted equally, as currently practiced according to *DSM*, or should psychometrically superior symptoms be prioritized? Crucially, identifying psychometrically sound symptoms and simultaneously eliminating unreliable symptoms are likely to improve diagnostic precision, enhance treatment effectiveness, and improve the efficiency of assessment procedures (i.e., cost, time). Despite its significant clinical implications, the question of evaluating individual ADHD symptoms relative to the latent trait continuum has rarely been investigated using modern quantitative methods. One suitable analytic method to address these gaps in knowledge is item response theory (IRT; Embretson & Reise, 2000), a model-based, item-level analysis that measures the relationship between an item and its latent trait. The current study used IRT to directly compare the psychometric quality of each ADHD symptom.

Traditional methods for assessing the psychometric properties of rating scales have important limitations (see Embretson & Reise, 2000, for a detailed discussion of the advantages of IRT relative to classical test theory). Despite these limitations and the unique advantages afforded by IRT, there are few studies consisting of IRT analyses of ADHD symptoms. Gomez (2008) employed IRT to evaluate the item functioning of parent and teacher ratings on a *DSM-IV* ADHD rating scale in a large sample of school-aged children in Australia. Two symptoms (i.e., "loses things" and "fidgets") poorly discriminated across the latent trait continuum according to parents *and* teachers. Furthermore, although every ADHD symptom demonstrated good discrimination and threshold values (relative to the latent trait), the overall precision of the measure (i.e., test information) was poor for latent scores below the mean (i.e., −1 *SD*) and well above the mean (i.e., +3 *SD*; Gomez, 2008). Purpura, Wilson, and Lonigan (2010) reported similar results among typically developing preschoolers: although all the items demonstrated acceptable discrimination parameters, they were less informative at high levels of the latent trait. Poor precision at high and low latent trait scores were also reported in preschool and adult samples (Gomez, 2011; Purpura et al., 2010). Importantly, these studies were based on samples of typically developing youth. It is unknown whether ADHD rating scales should be interpreted similarly in high-risk versus population-based samples, a critical consideration given that most children who are referred for ADHD problems are likely to have latent scores well above the mean (Gomez, 2008).

The present investigation substantively improves the literature by using IRT to examine the psychometric properties of ADHD symptoms, with the important consideration that its latent structure may be characterized by a common general factor. The study separately analyzed parent- and teacher-ratings of ADHD from four independent studies that consisted of highly parallel ascertainment procedures, thereby yielding a large well-characterized sample of children with and without ADHD. We compared item discrimination (symptoms with high discrimination values are better at differentiating individuals at a certain point in the trait range than symptoms with lower values), location (or how much of the latent trait is needed to have a 50% probability of endorsing the symptoms at or above the "threshold"), and item and test information (or the degree to which symptoms discriminate between individuals at a certain range or value on the latent trait) for each ADHD symptom.

Method

Participants

The current study used data from four samples of children with and without ADHD (Chronis-Tuscano et al., 2008; Li & Lee, 2012a; Mikami, Jack, Emeh, & Stephens, 2010), yielding a full sample of 526 children. Teacher data were available for 385 of the total sample. Table 1 summarizes demographic information available across the pooled sample. Children and families spanned three diverse urban, suburban, and rural areas in the United States. They were recruited using mailings and presentations to local physicians, ADHD self-help groups, schools, and mental health professionals. Expanded details about each sample's recruitment, ascertainment, and participant characteristics are explained elsewhere (Chronis-Tuscano et al., 2008; Li & Lee, 2012a; Mikami, Jack, Emeh, & Stephens, 2010). The pooled sample was 71% male, 6 to 12 years of age (M = 7.4; SD = 1.1), and ethnically diverse: 48.4% Caucasian, 22.1% African American, 6.0% Latino, 13.9% mixed-race, and 9.6% "other."

Clinical Assessment of Child ADHD

Procedures used to assess ADHD aligned with established clinical standards, which were nearly identical across each study that contributed to the pooled sample. Diagnoses were determined in part by parent and teacher reports on the child's ADHD symptoms using well-validated rating scales (described in greater detail below), and confirmed by fully structured or semistructured clinical interviews conducted with a parent (Study 1: Diagnostic Interview Schedule for Children, DISC; Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000; Studies 2-4: Kiddie-Schedule for Affective Disorders and Schizophrenia, K-SADS; Kaufman et al., 1997). Children met diagnostic criteria for ADHD if the respondent endorsed at least six symptoms of IA and/or six symptoms of HI for the child, while also endorsing criteria for age-of-onset and cross-situational impairment as specified in the *DSM*. On self-reported measures of ADHD, parents and teachers were asked to report on the child's behavior while off medication. In accordance with *DSM-IV* field trails (Lahey et al., 1994), each study employed the "or" algorithm for determining whether a symptom was considered present if either the parent or the teacher endorsed it as occurring "often" or "very often." All clinical interviews were administered by highly trained staff and doctoral students in clinical psychology. A total of 360 out of 526 (68.4%) children met full diagnostic criteria for ADHD.

Parent and Teacher Rating Scales for ADHD

Parent and teacher rating scales of childhood psychopathology were used across the four studies: the Disruptive Behavior Disorders Rating Scale, Parent and Teacher Versions (DBDRS; Pelham et al., 1992) in Studies 1, 2, and 3, and the Child Symptom Inventory–4, Parent and Teacher Checklists (CSI-4; Gadow & Sprafkin, 2002) in Study 4. The DBDRS and CSI-4 contain identical items and response options. For both rating scales, parents were asked whether each of the 18 ADHD symptoms occurred 0 = *not at all*, 1 = *just a little*, 2 = *pretty much*, or 3 = *very much*. The symptoms and response categories were identical for both measures, and were scored according to this polytomous scale (i.e., 0, 1, 2, and 3). Cronbach's alpha was .87 and .88 for IA and HI, respectively, and .90 for all 18 symptoms.

Statistical Analyses

In IRT, a latent trait (or *theta*, θ) is assumed to underlie responses for each item on the rating scale (Reise, Ainsworth, & Haviland, 2005). Thus, the probability that an individual will give a particular response to an item is influenced by θ. For example, parents are likely to positively endorse the item "has difficulty organizing tasks and activities" for their child if he/she is high on θ. This relationship is called the item response function, and is the fundamental unit in IRT used to evaluate item functioning (Reise et al., 2005). The item response function is derived from two parameters. L*ocations* are a set of item thresholds that correspond to value of θ at which there is a 50% probability of endorsing the item at the threshold, and *discrimination* reflects the item's ability to differentiate individuals (i.e., high vs. low ADHD) at the thresholds. In IRT, the reliability of an item can be judged by its *information function*, which represents the item's ability to differentiate individuals at each θ level. Information functions are summed together to form a *test information curve*, which estimates how well the measure functions along the entire θ continuum (Reise et al., 2005). Taken together, these estimates allow direct comparisons of the psychometric quality of each item relative to a trait continuum.

Unidimensionality and local independence must be established to ensure that the relationship between ADHD symptoms is fully characterized by the IRT model (Embretson & Reise, 2000). Unidimensionality was examined in Mplus 6.11 (Muthén & Muthén, 2007) by first conducting a confirmatory factor analysis (CFA). A bifactor model was subsequently fit using the best fitting factor model from the CFA. The bifactor model allows each item to have a positive loading on the general trait (which is assumed to underlie all items) as well as loadings on one or more "group" factors (Reise, Morizot & Hays, 2007). Fit statistics, including the chi-squared estimates, the comparative fit index (Bentler, 1990), the Tucker–Lewis index (Bentler, 1990), and the root-mean-square error of approximation (Browne & Cudeck, 1993) were examined to determine the best fitting overall model. Local dependence was assessed in IRTPRO (Cai, du Toit, & Thissen, 2011) using χ^2 statistics (Chen & Thissen, 1997) from the observed versus expected frequencies in each of the two-way cross tabulations between responses to each item. Chi-square values are standardized (i.e., z scores) and computed by comparing the observed and expected frequencies in each of the two-way cross-tabulations between responses to each item and other items. Excessively large χ^2 values (i.e., >10) indicate a violation of the local independence assumption (Chen & Thissen, 1997).

After assessing dimensionality and local dependence, we used the graded response model (GRM) analysis (Samejima, 1969) to analyze scales with ordered categories (i.e., 0, 1, 2, and 3). We calculated discrimination (α) and location parameters (β_i) from the GRM from the following formula:

$$P^*{}_{ix}(\theta) = \frac{\exp\left(\alpha_i\left(\theta-\beta_{ij}\right)\right)}{1+\exp\left(\alpha_i\left(\theta-\beta_{ij}\right)\right)},$$

where $P^*{}_{ix}(\Theta)$ is the probability of an examinee's raw item response (*x*) falling in or above a given category threshold (*j*) conditional on trait level (θ). Discrimination reflects the symptom's ability to differentiate individuals at different locations of θ. In other words, symptoms with high α are better at differentiating individuals on the trait continuum than symptoms with lower α values. For symptoms with multiple response formats, the item locations are a set of item thresholds (β_i) that correspond to how much of θ is needed to have a 50% probability of endorsing the item at the "threshold." Thus, for a four category item, four category response curves must be estimated, each reflecting the probability of selecting a given category conditional on the latent trait. These are derived from $K-1$ threshold response curves that reflect the probability of responding above a particular threshold as a function of the latent trait (Embretson & Reise, 2000). Symptoms with higher thresholds require more θ for a positive endorsement in the category than symptoms with lower thresholds. Parameter estimates were then used to derive item response curves and item information curves. Symptoms with high information across θ are more desirable than symptoms with low information, and more discriminating items tend to provide more information. Additionally, item information curves can be summed to create a test information curve, which displays information across θ. Finally, we converted the summed scores to IRT scaled scores using the summed score *expected a posterior* recursive method.

Results

Unidimensionality and Local Independence

To assess dimensionality, we compared three separate confirmatory factor models that were based on established theories regarding the latent structure of ADHD: unidimensional and two-dimensional CFA models (i.e., representing IA and HI), and a two-factor bifactor model. Fit statistics are presented in Table 2, and factor loadings for these models are presented in Tables 3 (parent ratings) and 4 (teacher ratings). For both parent- and teacher-ratings, the best fitting model was the two-factor bifactor model, where symptom-level covariation was attributed to a general ADHD factor as well as two group factors (i.e., representing the IA and HI symptom clusters). Factor loadings for the symptoms were significantly higher on the general factor of ADHD than the subfactors (IA and HI), suggesting that a unidimensional model could be assumed.

Standardized local dependence statistics for each pair of items were calculated in IRTPRO (results available on request). Symptoms with high local dependence suggest overlap to some degree, after controlling for the latent variable. For parent ratings, pairs of items that exhibited the highest degrees of local dependence values (i.e., χ^2 between 5 and 10) included (1) “Forgetful” and “Disorganized,” (2) “Forgetful” and “Loses,” (3) “Attention” and “Sitting,” (4) “Motor” and “Climbs,” (5) “Blurts” and “Attention,” (6) “Waiting” and “Attention,” and (7) “Interrupts” and “Blurts.” However, all local dependence values were less than 10, indicating local dependence violations were unlikely. For teacher ratings, several pairs of symptoms exhibited the excessively high degrees of local dependence values (i.e., $\chi^2 > 10$), including (1) “loses” and “distracted,” (2) “forgetful” and disorganized,” (3) “forgetful” and “loses,” (4) “motor” and “disorganized,” (5) “motor” and “forgetful,” (6) “blurts” and “forgetful,” (7) “blurts” and “talkative,” (8) “blurts” and “waiting,” (9) “blurts” and “interrupts,” and (10) “waiting” and “interrupts.” These results strongly suggest that there may be significant diagnostic overlap with respect to ADHD symptoms as rated by teachers, such that certain pairs of symptoms may constitute as “couplets” (i.e., individuals who blurt out are also highly likely to interrupt others). Thus, the teacher IRT analyses should be interpreted with some caution.

Item Discrimination, Location, and Information: Parent Ratings of ADHD Symptoms

Table 5 shows the discrimination and location parameter estimates from the GRM, and Figure 1 represents item response and item information curves for the most and least informative symptoms. Discrimination estimates, which represent an item’s ability to discriminate between individuals high and low on the latent trait, ranged from 1.65 to 3.21. Two symptoms demonstrated relatively high discrimination values ($\alpha > 3.00$): *has trouble keeping attention* (“Attention”; $\alpha = 3.21$, *SE* = .23) and does not follow through on instructions or fails to finish schoolwork, chores, or duties (“Finish”; $\alpha = 3.08$, *SE* = .23). The least discriminative items ($\alpha > 2.00$) were *often talks excessively* (“Talkative”; $\alpha = 1.65$, *SE* = .13) and often blurts out answers before questions have been finished (“Blurts”; $\alpha = 1.89$, *SE* = .15).

Location parameter estimates at the first threshold (β_1) ranged between −1.41 and −0.23, indicating that children who were estimated to be 0.23 to 1.41 *SD* below the mean of the latent trait were those who were 50% likely to at least endorse the second response option (i.e., *just a little*). This was expected, given our sample included both ADHD cases and non-ADHD controls. Location parameters estimates at the second (β_2) and third (β_3) thresholds ranged from −0.42 to 0.67 and 0.38 to 1.42, respectively. The “easiest” symptom to endorse at the thresholds of *pretty much* and *very much* (i.e., requiring the lowest θ to be 50% likely to endorse either option) was *is often easily distracted* (“Distracted”; $\beta_2 = 0.42$, *SE* = .06; $\beta_3 = 0.38$, *SE* = .06), whereas the most “difficult” symptom was often has trouble playing or doing leisure activities quietly (“Loud”; $\beta_2 = 0.67$, *SE* = .07; $\beta_3 = 1.42$, *SE* = .11).

The test information curve (i.e., sum of the individual information values for each symptom; Figure 2) indicated that information was normally distributed around the mean of the latent trait, with the peak information value at $\theta = .20$ (information = 31.51, *SE* = .18). The range of highest measurement precision was between ±2 *SD* of the 0, where the information values were >10 and the standard errors were less than .30. However, the precision of the measure declined precipitously at the margins, specifically beyond ±2 *SD* from 0, where information values were <10 and standard errors were greater than .37. For individual symptoms, peak information values were also normally distributed around mean of the latent trait ($\theta = .20$).

The symptoms that provided the highest amount of information around the mean (i.e., information values >2.0) were often has trouble keeping attention on tasks or play activities ("Attention"), often does not follow through on instructions and fails to finish schoolwork, chores, or duties in the workplace (not due to oppositional behavior or failure to understand directions ("Finish"), and *is often easily distracted* ("Distracted"), while the symptom that had the lowest amount of peak information around the mean (i.e., information value < 1.0) was *often talks excessively* ("Talkative").

Item Discrimination, Location, and Information: Teacher Ratings of ADHD Symptoms

The GRM was conducted on all 18 teacher-rated symptoms of ADHD. Discrimination and location parameter estimates are presented on Table 6. Teacher-reported ADHD symptoms were generally more discriminative along the latent trait continuum than parent ratings. The most discriminative symptoms (where $\alpha > 3$) were "Attention" ($\alpha = 3.26$, *SE* = .29), "Distracted" ($\alpha = 3.48$, *SE* = .32), "Fidgets" ($\alpha = 3.47$, *SE* = .32), "Sitting" ($\alpha = 3.33$, *SE* = .32), "Climbs" ($\alpha = 3.00$, *SE* = .34), and "Motor" ($\alpha = 3.19$, *SE* = .34). The least discriminative symptoms were "Careless" ($\alpha = 1.94$, *SE* = .18), "Forgetful" ($\alpha = 1.99$, *SE* = .19), and "Blurts" ($\alpha = 1.91$, *SE* = .20). Location parameter estimates at the first threshold (β_1, *just a little*) ranged between −1.21 and 0.42. Consistent with IRT estimates from the parent ratings, "Distracted" required the least amount of the latent trait to be 50% likely to be positively endorsed (either *pretty much* or *very much*), $\beta_2 = -.31$, *SE* = .08; $\beta_3 = .32$, *SE* = .06. The most "difficult" symptom at the higher thresholds was "Climbs" ($\beta_2 = .95$, *SE* = .07; $\beta_3 = 1.38$, *SE* = .09).

With respect to the individual item information curves, the most and least informative items are presented in Figure 3. Peak information values for most symptoms were normally distributed around mean of the latent trait ($\theta = .40$). The amount of information at the peak ranged from 1.04 ("Blurts") to 3.44 ("Fidgets"). The symptoms that provided the highest amount of information around the mean (i.e., information values > 3.0) were "Distracted," "Fidgets," and "Sitting," whereas the least informative symptoms (i.e., information values ~ 1.0) were "Careless" and "Blurts." The test information curve (Figure 4) indicated that, similar to parent ratings, information for teacher ratings of ADHD was normally distributed around the mean of the latent trait, with the peak information value at $\theta = .40$ (information = 38.38, *SE* = .16). The range of highest measurement precision was between $\theta = -1.40$ and $\theta = 1.90$, where the information values were >10 and the standard errors were less than .30. This range is narrower compared with parent ratings, particularly at the lower end of the latent trait continuum.

Converting Summed Scores Into θ Estimates

Finally, we computed θ from the summed scores using *estimated a posterior* recursive estimation in IRTPRO (Figure 5 and 6 for parent and teacher ratings, respectively). For the parent ratings, $\theta = 0$ corresponded to a summed score of approximately 24. In other words, individuals who had a summed score of 24 were likely to score at or very near the mean of the latent scale. Individuals who scored between 9 and 43 were within 1 *SD* of $\theta = 0$ [−.95, .98]. Individuals who scored between 0 and 8 were in the lowest range ($\theta = [-2.30, -1.03]$), falling below 1 *SD* of $\theta = 0$, while those who scored between 44 and 52 fell 1 *SD* above $\theta =$

0 [1.05, 1.84]. The most clinically severe group of individuals were those who scored 53 and above (θ = [2.03, 2.38]), as their scores were 2 *SD* above θ = 0.

Conversions from the summed raw scores derived from teacher ratings of ADHD symptoms approximated those derived from parent ratings. θ = 0 corresponded to a summed score of approximately 20. In other words, individuals with a summed score of 20 were likely to score at or very near the mean of the latent scale. Individuals who scored between 5 and 41 were within 1 *SD* of θ = 0 [−.98, .98]. Individuals who scored between 0 and 4 were in the lowest range (θ = [−2.03, −1.09]), falling below 1 *SD* of θ = 0, while those who scored between 42 and 52 fell 1 *SD* above θ = 0 [1.04, 1.90]. The most clinically severe group of individuals were those who scored 53 and above (θ = [2.07, 2.42]), as their scores were 2 *SD* above θ = 0.

Discussion

Based on a large, pooled sample of children with and without ADHD, we applied the GRM to estimate the discrimination, location, and information functions for ADHD symptoms using parent- and teacher-reported DSM symptom scales. As expected, the 18 symptoms of ADHD largely measured a single general factor of ADHD, as well as orthogonal factors that were composed of IA and HI symptoms. All symptoms demonstrated adequate discrimination and location values, although certain symptoms better discriminated different levels of the latent trait based on parent or teacher ratings. The reliability of the measure peaked between −2 *SD* and +2 *SD* for the parent ratings and between −1.40 *SD* and +1.90 *SD* for the teacher ratings, suggesting that the scale has excellent precision in the low to moderate range of ADHD, but may be less precise at the extreme ranges (particularly at the lower range). Finally, results from the summed score to IRT scale score conversions indicated that there was significant variability on the latent trait continuum among individuals who had a summed score between 9 and 43 on the parent measure and between 5 and 41 on the teacher measure, as total raw scores within these respective ranges fell within 1 *SD* of the mean of the latent trait.

These results must be interpreted relative to the metric, which is based on a pooled sample that included children with and without ADHD. The inclusion of ADHD cases and non-ADHD controls resulted in a sample mean (on the latent trait) that was likely higher than what was expected in a population-based sample, but there was also greater variation across the latent trait. Furthermore, most clinical measures provide peak measurement precision in the most extreme range of the trait continuum (Reise & Waller, 2009) and yet previous IRT studies of ADHD have focused exclusively on populations that fall out of this range (e.g., school-based samples). For parent ratings, the least informative symptom was *often talks excessively* (“Talkative”), while the most informative symptoms were often has trouble keeping attention on tasks or play activities (“Attention”), often does not follow through on instructions and fails to finish schoolwork, chores, or duties in the workplace (not due to oppositional behavior or failure to understand directions (“Finish”), and *is often easily distracted* (“Distracted”). The least informative teacher symptoms were often does not give close attention to details or makes careless mistakes (“Careless”) and often blurts out answers before questions have been finished (“Blurts”), while the most informative symptoms were *is often easily distracted* ("Distracted"), often fidgets with hands or feet or squirms in seat when sitting still is expected ("Fidgets"), and often gets up from seat when remaining in seat is expected ("Sitting"). Symptoms that are more informative than others may be more useful in terms of reliably assessing ADHD. For example, several studies have also reported that "Attention" is a more informative symptom than others (Gomez, 2008; Lindheim, Yu, Grasso, Kolko, & Youngstrom, 2015). This is notable because these studies used markedly different samples, including school-based and clinical samples, suggesting that this symptom may be diagnostically sensitive across settings. More replication of these findings are needed, especially in clinical samples, but accumulating evidence suggests that ADHD symptoms are not psychometrically equivalent and that certain symptoms may deserve greater weight than others.

Interpretations from the IRT models for teachers should be made cautiously, as several of the symptoms that were rated by teachers showed redundancy in the form high local dependence. This was not unique to our sample however, as high local dependence between teacher-rated ADHD items were also reported in an IRT study on a sample of Danish school children (Makransky & Bilenberg, 2014), suggesting that certain symptoms may need to be revised. For example, "Talkative" and "Blurts" were not only highly correlated, but they also demonstrated relatively poor psychometric properties in other IRT samples (e.g., Gomez, 2008; Purpura et al., 2010). Both items appeared to assess verbal impulsivity broadly and their inclusion as separate symptoms did not uniquely improve the reliability of the scale. Instead, one of the symptoms may be eliminated with little loss in measurement precision, while the language of the remaining symptom can be revised to emphasize its problematic nature across settings (e.g., often talks excessively in situations in which talking is undesirable). Poor psychometric properties may also be indicative of symptoms that measure a trait different from ADHD altogether. For example, "Talkative" overlaps with other disorders, including bipolar disorder and schizophrenia, as well as for cognitive dimensions such as intelligence and linguistic abilities (Wardle, Cederbaum, & de Wit, 2011). In fact, many Western parents often do not perceive excessive talking as problematic because it is associated with assertiveness and leadership (Chen & French, 2008) and school teachers perceive talkative children as being more intelligent and more likely to perform better academically compared to quiet children (Coplan, Hughes, Bosacki, & Rose-Krasnor, 2011). Thus, item responses on this symptom may have been influenced by cognitive or social factors, rather than the latent ADHD trait per se.

Another important finding was that the test information functions from the parent ratings were normally distributed near the mean of the latent trait, with the highest range of measurement precision between −2 *SD* and +2 *SD*. Compared with the parent ratings, teacher ratings provided relatively more measurement precision, but within a slightly narrower range on the latent trait continuum, between −1.40 *SD* and +1.90 *SD* of the mean. The results also indicated that both parent and teacher ratings can reliably measure the latent trait across a fairly wide range on the continuum. This is in contrast to most IRT studies based on epidemiological samples, in which similar clinical measures of ADHD provided peak measurement precision at a narrow (e.g., extreme) portion of the trait continuum (Gomez, 2008; Purpura et al., 2010; see Makransky & Bilenberg, 2014, for an exception). Given the disproportionately high number of low trait individuals in epidemiological samples, the poor reliability at the lower range of the continuum likely reflects sampling effects (relatively little variance along the latent trait) rather than poor psychometric properties of the scale in general.

These findings also have potentially important implications for the assessment of ADHD. First, the findings generally support the view that ADHD may be better characterized dimensionally, rather than categorically (Hudziak, Achenbach, Altoff, & Pine, 2007; Krueger, Markon, Patrick, & Iacono, 2005). This was evidenced by the fact that the relationship between IRT scores and summed scores was nearly linear and that information curves were normally distributed at the mean. However, a long-standing criticism of the dimensionality perspective for ADHD is that it complicates clinical decision making, such as determining whether treatment is warranted. Using IRT, we showed that cut-points can be applied empirically, rather than a priori, based on where an individual falls along the latent trait continuum. A recent study demonstrated that IRT and Bayesian methods used to calculate latent trait scores of ADHD (i.e., a "Posterior Probability of Diagnosis Index") were comparable and more diagnostically efficient than traditional clinical decision-making methods in terms of specificity and sensitivity (Lindhiem et al., 2015). Importantly, their study was also based on a clinical sample of children and adolescents, and item weights were calibrated according to their specific population (Lindhiem et al., 2015). Thus, quantitative methods for decision making can be helpful in determining clinical severity or diagnosis in specific study or patient samples, but may not apply generally across samples. More research is needed to determine the appropriate item weights across samples, especially given recent evidence of differential item functioning in ADHD across age and gender (Makransky & Bilenberg, 2014).

Second, mean parent ratings were higher than mean teacher ratings with respect to the latent trait, although teacher ratings were generally more precise across a narrower trait range. This is entirely consistent with a recent IRT investigation (Makransky & Bilenberg, 2014) and with other studies that show low concordance rates between parent and teacher ratings of ADHD (Antrop, Roeyers, Oosterlaan, & Van Oost, 2002; Gomez, 2007). Teachers may be less likely to endorse certain symptoms than parents due to having different perceptions (or perhaps greater leniency) regarding the child's problem behaviors (Molina, Smith, & Pelham, 2001), as teachers interact daily with a large classroom of typically developing age-matched students, whereas parents often lack such a reference with which to rate their own child's behavior. Different clinical thresholds and interpretations may be necessary when scoring and comparing teacher data with parent data, particularly if there is a lack of invariance across informants (Gomez, 2007). Alternatively, differences between raters may be indicative of children exhibiting situational specificity of their problem behaviors at home and at school, which would be evident in the case where the measure is psychometrically equivalent across informants (Gomez, 2007). Research on measurement invariance from an IRT perspective is emerging (see Gomez, 2007; Makransky & Bilenberg, 2014, for examples) given the importance of accurately assessing cross-situational impairment in the assessment of ADHD.

Finally, there was variability in the symptom parameters; some symptoms were more discriminating than others and certain symptoms may deserve greater weight toward diagnostic ascertainment or for initial screener items for ADHD. In other words, an individual with more symptoms of ADHD did not necessarily mean he/she also had more severe ADHD than the person with fewer symptoms, particularly if the symptoms that were positively endorsed had low discrimination values. In fact, the item discrimination parameters discerned from the GRM suggest that certain symptoms (e.g., “Talkative”) could have been excluded from the measure altogether, with very little loss of measurement precision. This could also increase the efficiency of assessing ADHD from self-report measures (i.e., fewer items) and enhance clinicians’ ability to screen routinely in primary care or even school settings. Future studies could also evaluate the relative weight of each ADHD symptom by validating these symptoms against independent measures of academic and social functioning.

These findings should be considered in light of several limitations. First, developmental aspects of ADHD could not be fully discerned given the limited age range of the sample. However, studies based on samples with wide age ranges must also be interpreted cautiously given that ADHD symptoms (particularly HI symptoms) diminish with age (Molina et al., 2009; Willoughby, 2003). Differential item functioning may be indicated if the latent trait changes with informant (i.e., teachers), age, or by variables such as sex, race-ethnicity, and so on. This is an important issue for future research, as the presence of differential item functioning may indicate that the symptoms may have utility for only specific populations. Additionally, as is typically the case with ADHD, this sample had other comorbidities that were not accounted for in the analyses. In a population-based twin study, approximately 90% of children with ADHD were affected by a comorbid disorder, with oppositional defiant and conduct disorders being the most prevalent (Willcutt, Pennington, Chhabildas, Friedman, & Alexander, 1999).

IRT has potential to significantly affect clinical decision making in ADHD. First, a symptom count cutoff score (i.e., having six or more symptoms) may have limited utility given that symptoms likely reflect a dimensional spectra rather than discrete categories. Instead, the ADHD subtypes could be supplemented or replaced by levels of severity, including high, moderate, and low trait levels along the general ADHD continuum (see Willcutt et al., 2012, for recommendations on the inclusion of dimensional modifiers). This could be further validated using impairment data to determine whether certain (or amounts of) treatments can be specifically tailored according to the severity level. Others have suggested the inclusion of context-specific dimensions, such as age, sex, and informant (Hudziak et al., 2007), although it is unknown whether symptoms vary psychometrically according to these different factors. Next, *DSM* currently weights all ADHD symptoms equally such that any combination of six symptoms (within IA and HI) satisfied criteria for ADHD (thus resulting in thousands of different symptom combinations). However, these data suggest that ADHD symptoms exhibited different discrimination and location parameters and provided different amounts of information along the continua. Thus, a revised diagnostic algorithm may consist of weighting ADHD symptoms according to their discrimination of individuals along the trait continua (i.e., items with better discrimination and information values are more highly weighted). For example, a major depressive episode *must* be accompanied by either depressed mood or anhedonia, thus prioritizing these symptoms in the diagnosis. ADHD may also be characterized by certain “hallmark” symptoms, although this requires further investigation in larger samples using IRT. Finally, IRT can play a significant role in streamlining screening interviews for the clinical assessment of ADHD (such as the K-SADS) by improving diagnostic accuracy and efficiency (e.g., reductions in cost, interviewing time, etc.). IRT can be used for computer-adaptive testing in which specific symptoms (with known item response functions) are administered to optimally match an individual's trait level (Reise et al., 2005; also see Lindhiem et al., 2015, for an application to ADHD). Thus, an ADHD diagnosis can potentially be determined with significantly fewer symptoms, and with extremely high precision, because the symptoms presented will match the latent trait of the individual. Although these results must be replicated across additional, larger samples, the prospect of combining quantitative and evidenced-based approaches to classification may facilitate the search for underlying causes of ADHD, improve diagnostic assessment, and potentially enhance treatment efficacy.

Acknowledgments

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Consortium of Neuropsychiatric Phenomics (CNP; NIH Roadmap for Medical Research Grant UL1-DE019580, RL1DA024853) to SSL and the National Institute of Mental Health to AYM (1R03MH12838), ACT (R03MH070666-1), and JJL (T32MH20030-14).

References

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 5th. Author; Washington, DC: 2013.

Antrop I, Roeyers H, Oosterlaan J, Van Oost P. Agreement between parent and teacher ratings of disruptive behavior disorders in children with clinically diagnosed ADHD. Journal of Psychopathology and Behavioral Assessment. 2002; 24:67–73. doi:10.1023/A:1014057325752.

Barkley RA, Fischer M. The unique contribution of emotional impulsiveness to impairment in major life activities in hyperactive children as adults. Journal of the American Academy of Child & Adolescent Psychiatry. 2010; 49:503–513. doi:10.1016/j.jaac.2010.01.019. [PubMed: 20431470]

Bentler PM. Comparative fit indexes in structural models. Psychological Bulletin. 1990; 107:238–246. doi:10.1037/0033-2909.107.2.238. [PubMed: 2320703]

Browne, MW.; Cudeck, R. Alternative ways of assessing model fit. In: Bollen, KA.; Long, JS., editors. Testing structural equation models. Sage; Newbury Park, CA: 1993. p. 136-162.

Cai, L.; du Toit, S.; Thissen, D. IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling. Scientific Software International; Chicago, IL: 2011.

Chen WH, Thissen D. Local dependence indexes for item pairs using item response theory. Journal of Educational and Behavioral Statistics. 1997; 22:265–289. doi:10.3102/10769986022003265.

Chen X, French DC. Children's social competence in cultural context. Annual Review of Psychology. 2008; 59:591–616. doi:10.1146/annurev.psych.59.103006.093606.

Chronis-Tuscano A, Molina BSG, Pelham WE, Applegate B, Dahlke A, Overmyer M, Lahey BB. Very early predictors of adolescent depression and suicide attempts in children with attention-deficit/hyperactivity disorder. Archives of General Psychiatry. 2010; 67:1044–1051. doi:10.1001/archgenpsychiatry.2010.127. [PubMed: 20921120]

Chronis-Tuscano A, Raggi VL, Clarke TL, Rooney ME, Diaz Y, Pian J. Associations between maternal attention-deficit/hyperactivity disorder symptoms and parenting. Journal of Abnormal Child Psychology. 2008; 36:1237–1250. doi:10.1007/s10802-008-9246-4. [PubMed: 18553132]

Collett BR, Crowley SL, Gimpel GA, Greenson JN. The factor structure of DSM-IV attention deficithyperactivity symptoms: A confirmatory factor analysis of the ADHD-SRS. Journal of Psychoeducational Assessment. 2000; 18:361–373. doi:10.1177/073428290001800406.

Coplan RJ, Hughes K, Bosacki S, Rose-Krasnor L. Is silence golden? Elementary school teachers' strategies and beliefs regarding hypothetical shy/quiet and exuberant/ talkative children. Journal of Educational Psychology. 2011; 103:939–951. doi:10.1037/a0024551.

de Nijs PF, van Lier PA, Verhulst FC, Ferdinand RF. Classes of disruptive behavior problems in referred adolescents. Psychopathology. 2007; 40:440–445. [PubMed: 17709974]

Elia J, Arcos-Burgos M, Bolton KL, Ambrosini PJ, Berrettini W, Muenke M. ADHD latent class clusters: DSM-IV subtypes and comorbidity. Psychiatry Research. 2009; 170:192–198. doi: 10.1016/j.psychres.2008.10.008. [PubMed: 19900717]

Embretson, SE.; Reise, SP. Item response theory for psychologists. Erlbaum; Mahwah, NJ: 2000.

Frazier TW, Youngstrom EA, Naugle RI. The latent structure of attention-deficit/hyperactivity disorder in a clinic-referred sample. Neuropsychology. 2007; 21:45–64. doi:10.1037/0894-4105.21.1.45. [PubMed: 17201529]

Froehlich TE, Lanphear BP, Epstein JN, Barbaresi WJ, Katusic SK, Kahn RS. Prevalence, recognition, and treatment of attention-deficit/hyperactivity disorder in a national sample of US children. Archives of Pediatric and Adolescent Medicine. 2007; 161:857–864. doi:10.1001/archpedi. 161.9.857.

Gadow, KD.; Sprafkin, J. Child Symptom Inventory 4: Screening and norms manual. Checkmate Plus; Stony Brook, NY: 2002.

Gomez R. Australian parent and teacher ratings of the DSM-IV ADHD symptoms: Differential symptom functioning and parent-teacher agreement and differences. Journal of Attention Disorders. 2007; 11:17–27. doi:10.1177/1087054706295665. [PubMed: 17606769]

Gomez R. Item response theory analyses of the parent and teacher ratings of the DSM-IV ADHD rating scale. Journal of Abnormal Child Psychology. 2008; 36:865–885. doi:10.1007/ s10802-008-9218-8. [PubMed: 18266103]

Gomez R. Item response theory analyses of adult self-ratings of the ADHD symptoms in the current symptoms scale. Assessment. 2011; 18:476–486. doi:10.1177/1073191110386341. [PubMed: 20966343]

Gomez R, Burns GL, Walsh JA, De Moura MA. Multitrait-multisource confirmatory factor analytic approach to the construct validity of ADHD rating scales. Psychological Assessment. 2003; 15:3–16. doi:10.1037/1040-3590.15.1.3. [PubMed: 12674720]

Hudziak JJ, Achenbach TM, Althoff RR, Pine DS. A dimensional approach to developmental psychopathology. International Journal of Methods in Psychiatric Research. 2007; 16:16–23. doi: 10.1002/mpr.217.

Kaufman J, Birmaher B, Brent D, Rao UMA, Flynn C, Moreci P, Ryan N. Schedule for Affective Disorders and Schizophrenia for School-Age Children–Present and Lifetime Version (K-SADS-PL): Initial reliability and validity data. Journal of the American Academy of Child & Adolescent Psychiatry. 1997; 36:980–988. doi:10.1097/00004583-199707000-00021. [PubMed: 9204677]

Krueger RF, Markon KE, Patrick CJ, Iacono WG. Externalizing psychopathology in adulthood: A dimensional-spectrum conceptualization and its implications for DSM-V. Journal of Abnormal Psychology. 2005; 114:537–550. doi:10.1037/0021-843X.114.4.537. [PubMed: 16351376]

Lahey BB, Applegate B, McBurnett K, Biederman J, Greenhill LL, Hynd GW, Shaffer D. DSM-IV field trials for attention deficit/hyperactivity disorder in children and adolescents. American Journal of Psychiatry. 1994; 151:1673–1685. [PubMed: 7943460]

Lahey BB, Pelham WE, Schaughency EA, Atkins MS, Murphy A, Hynd G, Lorys-Vernon A. Dimensions and types of attention deficit disorder. American Academy of Child and Adolescent Psychiatry. 1988; 27:330–335.

Langley K, Fowler T, Ford T, Thapar AK, van den Bree M, Harold G, Thapar A. Adolescent clinical outcomes for young people with attention-deficit hyperactivity disorder. British Journal of Psychiatry. 2010; 196:235–240. doi:10.1192/bjp.bp.109.066274. [PubMed: 20194547]

Larsson H, Anckarsater H, Råstam M, Chang Z, Lichtenstein P. Childhood attention-deficit hyperactivity disorder as an extreme of a continuous trait: A quantitative genetic study of 8,500 twin pairs. Journal of Child Psychology and Psychiatry. 2012; 53:73–80. doi:10.1111/j. 14697610.2011.02467.x. [PubMed: 21923806]

Li JJ, Lee SS. Association of positive and negative parenting behavior with childhood ADHD: Interactions with offspring monoamine oxidase A (MAO-A) genotype. Journal of Abnormal Child Psychology. 2012a; 40:165–175. doi:10.1007/ s10802-011-9553-z. [PubMed: 21826446]

Li JJ, Lee SS. Interaction of dopamine transporter (DAT1) genotype and maltreatment for ADHD: A latent class analysis. Journal of Child Psychology and Psychiatry. 2012b; 53:997–1005. doi: 10.1111/j.1469-7610.2012.02563.x. [PubMed: 22646917]

Lindhiem O, Yu L, Grasso DJ, Kolko DJ, Youngstrom EA. Adapting the posterior probability of diagnosis index to enhance evidence-based screening an application to ADHD in primary care. Assessment. 2015; 22:198–201. doi:10.1177/1073191114540748. [PubMed: 25000935]

Makransky G, Bilenberg N. Psychometric properties of the parent and teacher ADHD Rating Scale (ADHD-RS): Measurement invariance across gender, age, and informant. Assessment. 2014; 21:694–705. doi:10.1177/1073191114535242. [PubMed: 24852496]

Mikami A, Jack A, Emeh C, Stephens H. Parental influences on children with ADHD: Parental behaviors associated with children's peer relationships. Journal of Abnormal Child Psychology. 2010; 38:721–736. [PubMed: 20339912]

Molina BSG, Hinshaw SP, Swanson JM, Arnold LE, Vitiello B, Jensen PS, Houck PR. The MTA at 8 years: Prospective follow-up of children treated for combined-type ADHD in a multisite study. Journal of the American Academy of Child & Adolescent Psychiatry. 2009; 48:484–500. doi: 10.1097/CHI.0b013e31819c23d0. [PubMed: 19318991]

Molina BSG, Smith BH, Pelham WE. Factor structure and criterion validity of secondary school teacher ratings of ADHD and ODD. Journal of Abnormal Child Psychology. 2001; 29:71–82. doi: 10.1023/A:1005203629968. [PubMed: 11316336]

Muthén, LK.; Muthén, BO. Mplus User's Guide. Sixth. Muthén & Muthén; Los Angeles, CA: 1998-2011.

Nigg JT. Future directions in ADHD etiology research. Journal of Clinical Child & Adolescent Psychology. 2012; 41:524–533. doi:10.1080/15374416.2012.686870. [PubMed: 22642834]

Nikolas MA, Burt SA. Genetic and environmental influences on ADHD symptom dimensions of inattention and hyperactivity: *A meta-analysis*. Journal of Abnormal Psychology. 2010; 119:1–17. doi: 10.1037/a0018010. [PubMed: 20141238]

Normand S, Flora DB, Toplak ME, Tannock R. Evidence for a general ADHD factor from a longitudinal general school population study. Journal of Abnormal Child Psychology. 2012; 40:555–567. doi:10.1007/s10802-011-9584-5. [PubMed: 22033884]

Pelham WE, Gnagy EM, Greenslade KE, Milich R. Teacher ratings of DSM-III-R symptoms for the disruptive behavior disorders. Journal of the American Academy of Child & Adolescent Psychiatry. 1992; 31:210–218. doi: 10.1097/00004583-199203000-00006. [PubMed: 1564021]

Pillow DR, Pelham WE Jr. Hoza B, Molina BS, 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. doi:10.1023/A:1022658618368. [PubMed: 9700521]

Purpura DJ, Wilson SB, Lonigan CJ. Attention-deficit/hyperactivity disorder symptoms in preschool children: Examining psychometric properties using item response theory. Psychological Assessment. 2010; 22:546–558. doi:10.1037/ a0019581. [PubMed: 20822267]

Rasmussen ER, Neuman RJ, Heath AC, Levy F, Hay DA, Todd RD. Replication of the latent class structure of attention-deficit/hyperactivity disorder (ADHD) subtypes in a sample of Australian twins. Journal of Child Psychology and Psychiatry. 2002; 43:1018–1028. doi: 10.1111/1469-7610.00229. [PubMed: 12455923]

Reise SP, Ainsworth AT, Haviland MG. Item response theory fundamentals, applications, and promise in psychological research. Current Directions in Psychological Science. 2005; 14:95–101. doi: 10.1111/j.0963-7214.2005.00342.x.

Reise SP, Morizot J, Hays RD. The role of the bifactor model in resolving dimensionality issues in health outcomes measures. Quality of Life Research. 2007; 16:19–31. doi: 10.1007/ s11136-007-9183-7. [PubMed: 17479357]

Reise SP, Waller NG. Item response theory and clinical measurement. Annual Review of Clinical Psychology. 2009; 5:27–48. doi:10.1146/annurev.clinpsy.032408.153553.

Rohde LA, Barbosa G, Polanczyk G, Eizirik M, Rasmussen ER, Neuman RJ, Todd RD. Factor and latent class analysis of DSM-IV ADHD symptoms in a school sample of Brazilian adolescents. Journal of the American Academy of Child & Adolescent Psychiatry. 2001; 40:711–718. doi: 10.1097/00004583-200106000-00017. [PubMed: 11392350]

Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement. 1969; 43:100. Retrieved from http://psycnet.apa.org/psycinfo/.

Shaffer D, Fisher P, Lucas CP, Dulcan MK, Schwab-Stone ME. NIMH Diagnostic Interview Schedule for Children Version IV (NIMH DISC-IV): Description, differences from previous versions, and reliability of some common diagnoses. Journal of the American Academy of Child & Adolescent Psychiatry. 2000; 39:28–38. doi:10.1097/00004583-200001000-00014. [PubMed: 10638065]

Todd RD, Rasmussen ER, Neuman RJ, Reich W, Hudziak JJ, Bucholz KK, Heath A. Familiality and heritability of subtypes of attention deficit hyperactivity disorder in a population sample of adolescent female twins. American Journal of Psychiatry. 2001; 158:1891–1898. doi:10.1176/appi.ajp.158.11.1891. [PubMed: 11691697]

Toplak ME, Sorge GB, Flora DB, Chen W, Banaschewski T, Buitelaar J, Gill M. The hierarchical factor model of ADHD: Invariant across age and national groupings? Journal of Child Psychology and Psychiatry. 2012; 53:292–303. doi:10.1111/j.1469-7610.2011.02500.x. [PubMed: 22084976]

Volk HE, Neuman RJ, Todd RD. A systematic evaluation of ADHD and comorbid psychopathology in a population-based twin sample. Journal of the American Academy of Child & Adolescent Psychiatry. 2005; 44:768–775. doi:10.1097/01.chi.0000166173.72815.83. [PubMed: 16034278]

Wardle M, Cederbaum K, de Wit H. Quantifying talk: Developing reliable measures of verbal productivity. Behavior Research Methods. 2011; 43:168–178. doi:10.3758/ s13428-010-0019-y. [PubMed: 21287128]

Willcutt EG, Nigg JT, Pennington BF, Solanto MV, Rohde LA, Tannock R, Lahey BB. Validity of DSM-IV attention deficit/hyperactivity disorder symptom dimensions and subtypes. Journal of Abnormal Psychology. 2012; 121:991–1010. [PubMed: 22612200]

Willcutt EG, Pennington BF, Chhabildas NA, Friedman MC, Alexander J. Psychiatric comorbidity associated with DSM-IV ADHD in a nonreferred sample of twins. Journal of the American Academy of Child & Adolescent Psychiatry. 1999; 38:1355–1362. doi: 10.1097/00004583-199911000-00009. [PubMed: 10560221]

Willcutt EG, Pennington BF, DeFries JC. Etiology of inattention and hyperactivity/impulsivity in a community sample of twins with learning difficulties. Journal of Abnormal Child Psychology. 2000; 28:149–159. doi:10.1023/A:1005170730653. [PubMed: 10834767]

Willoughby MT. Developmental course of ADHD symptomatology during the transition from childhood to adolescence: A review with recommendations. Journal of Child Psychology and Psychiatry. 2003; 44:88–106. doi:10.1111/14697610.t01-1-00104. [PubMed: 12553414]

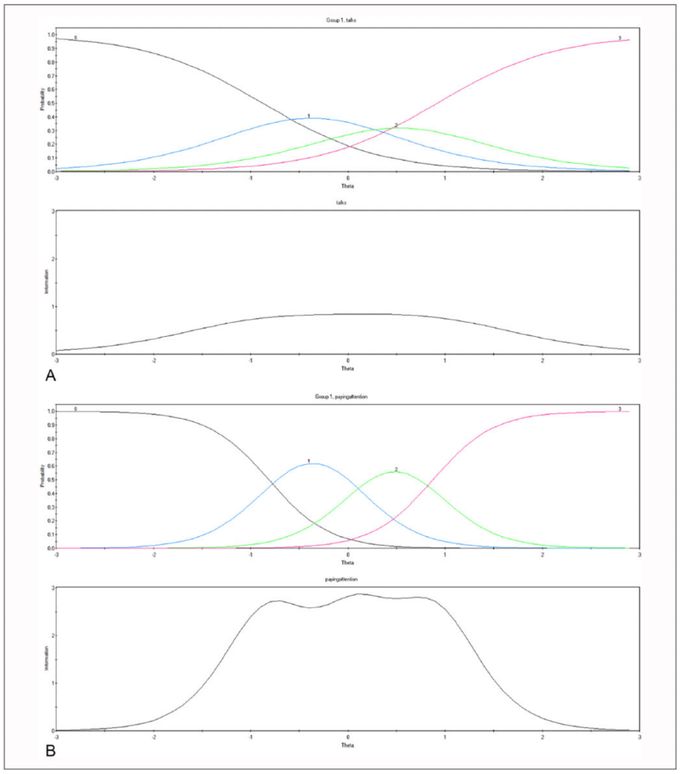

Figure 1.

Item response and item information curves teacher-rated (A) "Talkative" and (B) "Attention."

Note. Figure above shows item response curves, where response categories are rated 0 (black) = *not at all*, 1 (blue) = *just a little*, 2 (green) = *pretty much*, and 3 (red) = *very much*; Figure below shows item information curve.

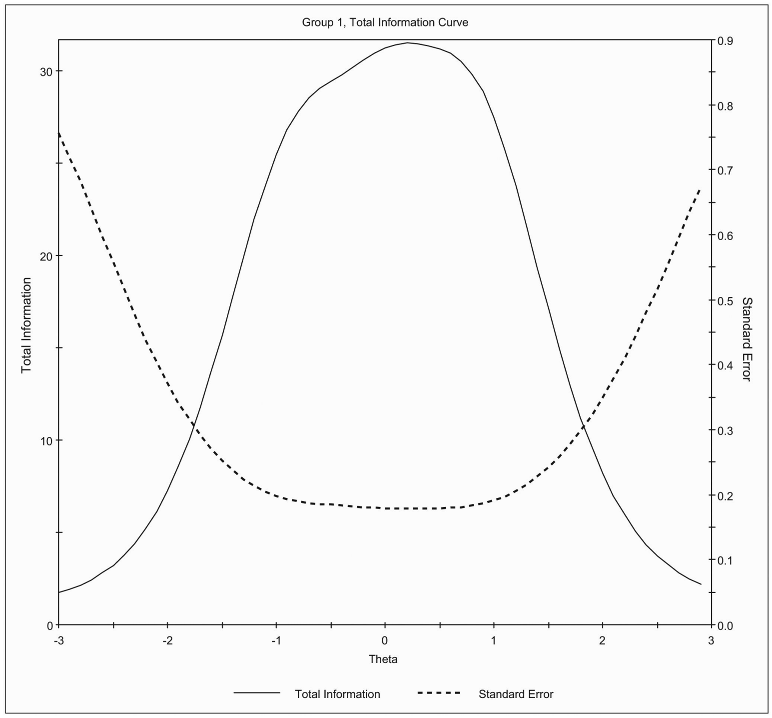

Figure 2.
Test information curve of parent ratings.

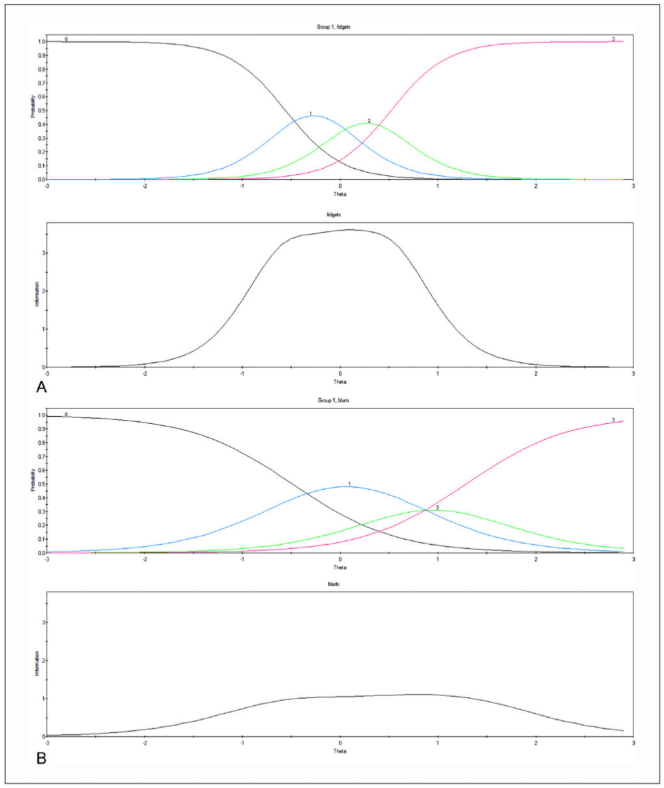

Figure 3.

Item response and item information curves teacher-rated (A) “Fidgets” and (B) “Blurts”.

Note. Figure above shows item response curves, where response categories are rated 0 (black) = *not at all*, 1 (blue) = *just a little*, 2 (green) = *pretty much*, and 3 (red) = *very much*; Figure below shows item information curve.

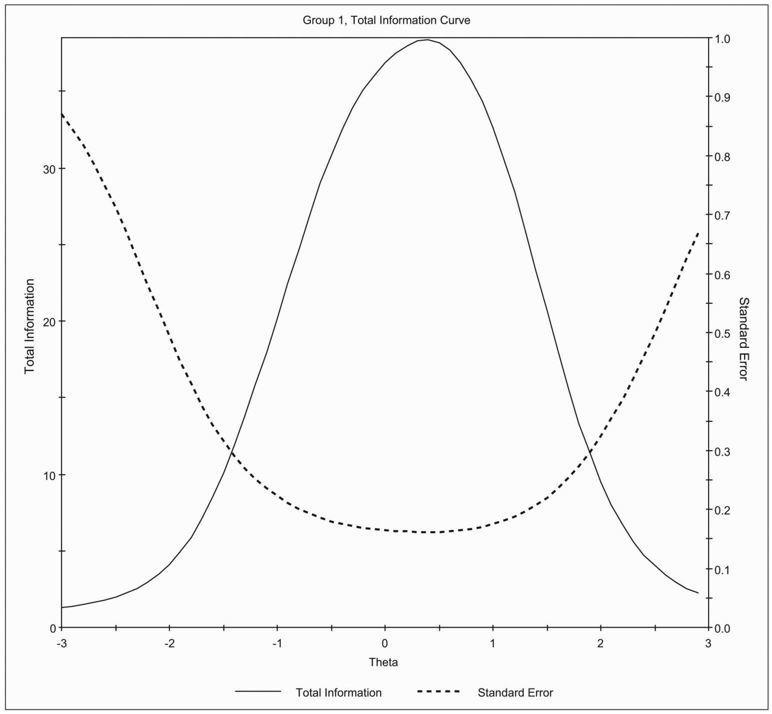

Figure 4.
Test information curve of teacher ratings.

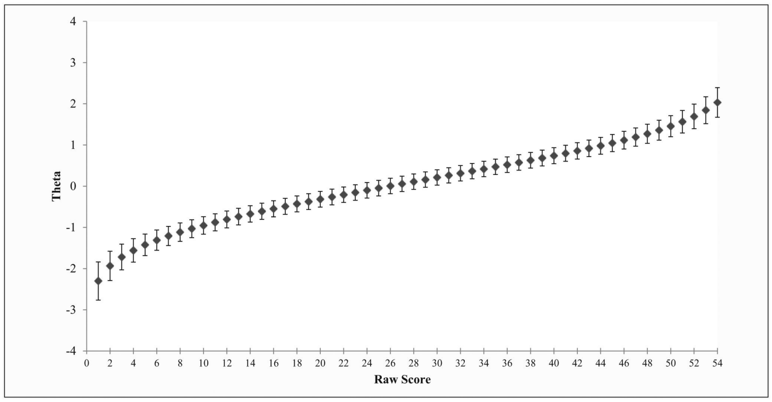

Figure 5.

Summed score to IRT scale scores of parent ratings.

Note. Standard deviation error bars are depicted. Marginal reliability of the scaled scores for summed scores = .95.

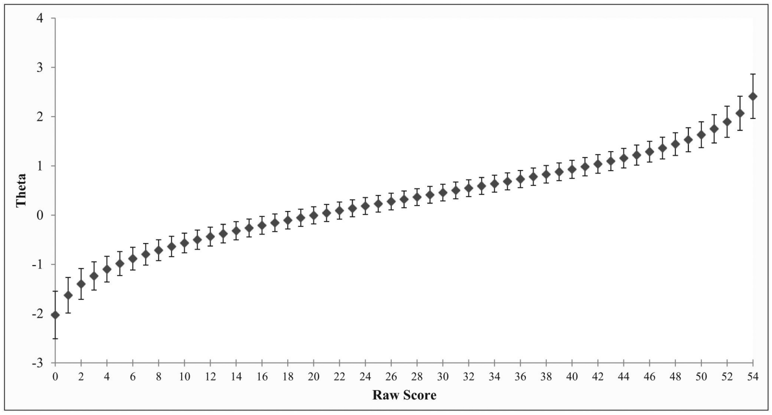

Figure 6.

Summed score to IRT scale scores of teacher ratings.

Note. Standard deviation error bars are depicted. Marginal reliability of the scaled scores for summed scores = .95.

Table 1

Descriptive Information by Study.

Variable	Study 1	Study 2	Study 3	Study 4
n	228	70	98	130
Residential area	Metropolitan	Suburban and metropolitan	Suburban and metropolitan	Suburban and rural
% ADHD	57	100	100	49
Age, *M* (*SD*)	7.35 (1.10)	8.06 (1.16)	8.76 (2.06)	8.22 (1.22)
Sex (% male)	71	71	67	67
Race/ethnicity (%)				
White	46	36	49	85
Black	5	41	31	5
Hispanic	11	4	0	1
Mixed	26	7	0	7
Other, refused	12	12	20	2
Comorbidity (%)				
ODD	27	48	41	15
CD	1	20	26	0

Note. *N* = 526; Study 1 = Li & Lee, 2012a; Study 2 and 3 = Chronis-Tuscano et al., 2008; Study 4 = Mikami, Jack, Emeh, & Stephens, 2010; ADHD = attention-deficit/hyperactivity disorder; ODD = oppositional defiant disorder; CD = conduct disorder.

Table 2

Fit Statistics of Confirmatory Factor Models for Parent and Teacher Ratings of ADHD.

	χ^2	*df*	CFI	TLI	RMSEA
Parent ratings					
1 Factor	630.05	36	.92	.97	.18
2 Factor	392.51	58	.96	.99	.10
2 Factor bifactor	261.91	66	.97	.99	.08
Teacher ratings					
1 Factor	510.60	21	.91	.96	.24
2 Factor	381.48	38	.94	.98	.15
2 Factor bifactor	220.12	44	.98	.99	.08

Note. ADHD = attention-deficit/hyperactivity disorder; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root-mean-square error of approximation.

Table 3

Results of Traditional Confirmatory Factor Model Versus Bifactor Model on Parent-Rated Attention-Deficit/ Hyperactivity Disorder Symptoms.

	Traditional confirmatory factor model		Bifactor model			
	1 Factor		2 Factor		2 Factor	
Symptoms	**1**	**1**	**2**	***g***	**F1**	**F2**
Careless	.79	.82		.72	.39	
Attention	.87	.89		.84	.30	
Listen	.83	.82		.86	.11	
Finish	.88	.92		.80	.43	
Disorganized	.87	.90		.74	.54	
Avoids	.80	.84		.70	.46	
Loses	.81	.84		.73	.42	
Distracted	.85	.85		.88	.13	
Forgetful	.86	.89		.73	.54	
Fidgets	.83		.85	.82		.25
Sitting	.80		.83	.79		.26
Climbs	.81		.86	.74		.42
Loud	.77		.80	.74		.30
Motor	.81		.87	.73		.50
Talkative	.74		.78	.67		.43
Blurts	.78		.83	.71		.42
Waiting	.84		.86	.82		.30
Interrupts	.80		.85	.75		.42

Note. Careless = *often does not give close attention to details or makes careless mistakes in schoolwork, work, or other activities*; Attention = *often has trouble keeping attention on tasks or play activities*; Listen = *often does not seem to listen when spoken to directly*; Finish = *often does not follow through on instructions and fails to finish schoolwork, chores, or duties in the workplace (not due to oppositional behavior or failure to understand directions)*; Disorganized = *often has trouble organizing activities*; Avoid = *often avoids, dislikes, or doesn't want to do things that take a lot of mental effort for a long period of time;* Loses = *often loses things needed for tasks and activities (e.g., toys, assignments, pencils, books, or tools)*; Distracted = *is often easily distracted;* Forgetful = *is often forgetful in daily activities;* Fidgets = *often fidgets with hands or feet or squirms in seat when sitting still is expected*; Sitting = *often gets up from seat when remaining in seat is expected*; Climbs = *often excessively runs about or climbs when and where it is not appropriate*; Loud = *often has trouble playing or doing leisure activities quietly*; Motor = *is often "on the go" or often acts as if "driven by a motor*"; Talkative = *often talks excessively*; Blurts = *often blurts out answers before questions have been finished*; Waiting = *often has trouble waiting one's turn*; Interrupts = *often interrupts or intrudes on others (e.g., butts into conversations or games).*

Table 4

Results of Traditional Confirmatory Factor Model Versus Bifactor Model on Teacher-Rated Attention-Deficit/ Hyperactivity Disorder Symptoms.

	Traditional confirmatory factor model			Bifactor model		
	1 Factor	2 Factor		2 Factor		
Symptoms	**1**	**1**	**2**	**g**	**F1**	**F2**
Careless	.77	.80		.67	.46	
Attention	.90	.93		.83	.40	
Listen	.80	.84		.80	.23	
Finish	.86	.89		.74	.51	
Disorganized	.89	.91		.70	.65	
Avoids	.79	.83		.68	.48	
Loses	.83	.87		.70	.53	
Distracted	.91	.95		.88	.33	
Forgetful	.81	.83		.62	.63	
Fidgets	.87		.94	.96		.01
Sitting	.86		.90	.85		.30
Climbs	.84		.87	.81		.33
Loud	.83		.87	.80		.32
Motor	.88		.91	.83		.39
Talkative	.81		.84	.69		.54
Blurts	.81		.83	.64		.62
Waiting	.87		.89	.74		.55
Interrupts	.85		.88	.72		.56

Table 5

Graded Response Model Parameter Estimates: Parent Ratings.

Symptom	*a*	*SE*	b_1	*SE*	b_2	*SE*	b_3	*SE*
Careless	2.19	.16	−1.41	.10	−0.14	.06	0.80	.07
Attention	3.21	.23	−0.81	.07	0.09	.05	0.88	.06
Listen	2.77	.20	−1.11	.08	−0.06	.06	0.86	.07
Finish	3.08	.23	−0.94	.07	0.01	.05	0.75	.06
Disorganized	2.62	.19	−0.81	.07	0.08	.06	0.86	.07
Avoids	2.16	.16	−0.85	.08	0.01	.06	0.89	.08
Loses	2.34	.17	−0.82	.08	0.28	.06	1.09	.08
Distracted	2.84	.21	−1.32	.08	−0.42	.06	0.38	.06
Forgetful	2.56	.19	−0.71	.07	0.22	.06	1.01	.07
Fidgets	2.38	.18	−0.90	.08	−0.15	.06	0.58	.06
Sitting	2.36	.18	−0.57	.07	0.42	.06	1.15	.08
Climbs	2.20	.17	−0.32	.07	0.56	.07	1.33	.10
Loud	2.08	.17	−0.23	.07	0.67	.07	1.42	.11
Motor	2.06	.16	−0.40	.07	0.29	.06	0.95	.08
Talkative	1.65	.13	−0.88	.09	0.12	.07	0.92	.09
Blurts	1.89	.15	−0.63	.08	0.35	.07	1.29	.10
Waiting	2.60	.20	−0.51	.07	0.38	.06	1.04	.08
Interrupts	2.07	.16	−1.33	.10	−0.08	.06	0.85	.08

Note. *a* = item discriminations; b_1 to b_3 = item locations; *SE* = standard error.

Table 6

Graded Response Model Parameter Estimates: Teacher Ratings.

Symptom	*a*	*SE*	b_1	*SE*	b_2	*SE*	b_3	*SE*
Careless	1.94	.18	−1.13	.13	0.13	.08	0.92	.09
Attention	3.26	.29	−0.72	.10	0.09	.07	0.82	.06
Listen	2.57	.23	−0.42	.09	0.55	.07	1.20	.09
Finish	2.57	.23	−0.56	.10	0.24	.07	0.80	.07
Disorganized	2.45	.22	−0.68	.10	0.17	.07	0.79	.07
Avoids	2.16	.20	−0.39	.09	0.46	.07	1.12	.09
Loses	2.27	.21	−0.28	.09	0.65	.07	1.30	.10
Distracted	3.48	.32	−1.21	.12	−0.31	.08	0.32	.06
Forgetful	1.99	.19	−0.50	.10	0.44	.07	1.34	.11
Fidgets	3.47	.32	−0.56	.09	0.02	.07	0.52	.06
Sitting	3.33	.32	−0.22	.08	0.42	.06	0.97	.07
Climbs	3.00	.34	0.42	.06	0.95	.07	1.38	.09
Loud	2.76	.28	−0.08	.08	0.62	.06	1.24	.09
Motor	3.19	.34	0.02	.07	0.54	.06	0.99	.07
Talkative	2.01	.20	−0.57	.11	0.33	.07	1.09	.09
Blurts	1.91	.20	−0.49	.11	0.61	.08	1.29	.11
Waiting	2.48	.25	−0.29	.09	0.50	.06	1.14	.08
Interrupts	2.35	.23	−0.65	.10	0.37	.07	1.05	.08

Note. *a* = item discriminations; b_1 to b_3 = item locations; *SE* = standard error.