Journal of Abnormal Child Psychology (2020) 48:679–691
https://doi.org/10.1007/s10802-020-00618-7

ADHD Parent and Teacher Symptom Ratings: Differential Item Functioning across Gender, Age, Race, and Ethnicity

George J. DuPaul[1] · Qiong Fu[1] · Arthur D. Anastopoulos[2] · Robert Reid[3] · Thomas J. Power[4]

Published online: 14 January 2020

Abstract

Parent and teacher ratings of the two attention-deficit/hyperactivity disorder (ADHD) symptom dimensions (i.e., inattention, hyperactivity-impulsivity) have been found to differ across child gender, age, race, and ethnicity. Group differences could be due to actual variation in symptomatic behaviors but also could be due to measurement items functioning differently based on child characteristics. This study extended prior investigations establishing measurement invariance at the symptom dimension and item levels, by examining possible measurement variance across child demographic characteristics at the item level (i.e., differential item functioning [DIF]) in two large national samples. Using the Rasch rating scale model (Andrich *Psychometrika*, 43, 561–73, 1978), we examined DIF of the 18 ADHD symptoms in samples of 2079 children ($n = 1037$ males) from 5 to 17 years old ($M = 10.7$; $SD = 3.8$) rated by parents and 1070 children ($n = 535$ males) aged from 5 to 17 years old ($M = 11.5$; $SD = 3.5$) rated by teachers. All but six ADHD symptom items showed DIF across child age, gender, race (Black vs. White), and ethnicity with more items showing DIF for age than for gender, race, or ethnicity. For child gender and age, more items showed DIF for parent than for teacher ratings. More items showed DIF across racial groups for teacher than for parent ratings. Only two parent- and teacher-rated items showed DIF for ethnicity. Implications of findings for practice, research, and future iterations of ADHD diagnostic criteria are discussed.

Keywords Attention-deficit/hyperactivity disorder · Parent ratings · Teacher ratings · Differential item functioning

Given the relatively high prevalence (Polanczyk, Willcutt, Salum, Kieling, & Rohde, 2014) and substantial costs (Chorozoglou et al., 2015) associated with attention-deficit/hyperactivity disorder (ADHD), it is important that developmentally appropriate and culturally sensitive assessments are available to identify youth who may require services for this disorder. Adult respondents (i.e., parents, teachers) typically are asked to report the frequency of ADHD symptoms in the context of broadband (e.g., Behavior Assessment System for Children-3rd edition; Reynolds & Kamphaus, 2015) or narrowband (e.g., Vanderbilt Rating Scale; Wolraich, Lambert, Doffing, Bickman, Simmons, & Worley, 2003) rating scales. Questionnaires that contain items linked to DSM-5 criteria for ADHD (American Psychiatric Association, 2013) are particularly valuable for diagnostic purposes. For example, the ADHD Rating Scale-5 (ARS-5; DuPaul, Power, Anastopoulos, & Reid, 2016a) includes 18 items that correspond to the nine inattention (IA) symptoms and nine hyperactivity-impulsivity (HI) symptoms described in the DSM-5.[1] Findings indicated that these dimensions were invariant across gender, age, informant, informant gender, and language (DuPaul, Reid, Anastopoulos, Lambert, Watkins, & Power, 2016b).

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10802-020-00618-7) contains supplementary material, which is available to authorized users.

✉ George J. DuPaul
gjd3@lehigh.edu

[1] Department of Education and Human Services, Lehigh University, 111 Research Drive, Bethlehem, PA 18015, USA

[2] University of North Carolina-Greensboro, Greensboro, NC, USA

[3] University of Nebraska-Lincoln, Lincoln, NE, USA

[4] Children's Hospital of Philadelphia, Perelman School of Medicine at University of Pennsylvania, Philadelphia, PA, USA

[1] Some studies have found ADHD symptom ratings to comprise three or more dimensions (e.g., Merrell & Tymms, 2003; Tymms & Merrell, 2011), and confirmatory factor analyses of the ARS-5 supported both a two-factor and three-factor structure. Because there was no substantive difference between these models, a two-factor solution consistent with the DSM-5 was applied.

Differences in ADHD Symptom Ratings across Child Demographic Characteristics

Although factor structure invariance has been established, mean IA and HI subscale (i.e., dimension) scores typically vary as a function of child age, gender, race, and ethnicity (Miller, Nigg, & Miller, 2009; Reid et al., 2000). For example, ARS-5 parent and teacher ratings were significantly higher for boys relative to girls and younger versus older children (DuPaul et al., 2016b). There also were significant differences across racial and ethnic minority groups for teacher ratings, with non-Hispanic African American children receiving higher symptom scores than non-Hispanic White, Asian, and Hispanic children. Equivocal findings have been obtained regarding differences between Hispanic and non-Hispanic children, with some studies finding higher ratings for non-Hispanic children (e.g., DuPaul et al., 2016b) and other investigations obtaining higher ratings for Hispanic children (e.g., de Ramirez & Shapiro, 2005). Some of the differences are likely due to actual variation in symptomatic behavior between boys and girls and younger versus older children; however, it is possible that symptom rating scales may function differently across child demographic groups. Prior studies have explicated demographic differences at a dimensional level (IA and HI) and, to a lesser extent, at the individual item (i.e., symptom) level. It is possible that mean parent or teacher ratings of individual symptoms could vary across child age, gender, race, or ethnicity even when the overall scale is functioning as intended. That is, differential item function (DIF) might be present when the probability of endorsement on an individual item differs across subgroups (e.g., gender, race, ethnicity) with equivalent levels of latent trait (Potenza & Dorans, 1995). If substantial DIF is found for ratings on an ADHD symptom for construct-irrelevant reasons (e.g., due to different teacher expectations or how a symptom is perceived for children from a particular racial or gender subgroup), this could mean that clinical interpretation of assessment results would need to account for the impact of child demographic characteristics on perception of symptomatic behaviors.

Item Response Theory and Rasch Modeling

One way to assess possible item-level measurement differences across demographic groups is to examine DIF using Rasch or item response theory (IRT) models. IRT (e.g., Lord & Novick, 1968; Rasch, 1960) is a measurement approach that relates the probability of endorsement or success on an item to respondents' latent trait and item characteristics. Compared with classical test theory, IRT has been increasingly applied in education, psychology, and health areas for its item-level information as well as its group- and test-independence (i.e., ability/trait and item parameter invariance regardless of the survey items and respondents, if the data-model fit is present). The Rasch model[2] (Rasch, 1960), mathematically identical to the one parameter-logistic (1-PL) IRT, was originally applied to binary data (e.g., *yes* vs. *no*). It predicts the probability of a specific response on an item as a joint function of a person's *ability* or level of an underlying *trait* (e.g., IA or HI dimensionality severity in the current study) and the item difficulty. The item difficulty is the item location on the *trait* scale for a 50% chance of endorsing a particular category for the presence of a symptom (item). For polytomous data (i.e., ordered data with more than two response options, such as Likert scale data), the rating scale model (RSM; Andrich, 1978) and the partial credit model (Masters, 1982) are commonly used.

ADHD Symptom Ratings: IRT and Rasch Analyses

IRT has been used to demonstrate reliability, trait discriminability, and diagnostic utility of parent and teacher ADHD symptom ratings in community samples (e.g., Gomez, 2008a, 2008b; Li et al., 2016). Similarly, Rasch model analyses of parent ratings on the Strengths and Weaknesses of ADHD symptoms and Normal behavior scale (SWAN; Young, Levy, Martin, & Hay, 2009) supported diagnostic utility of inattention and, to a lesser extent, hyperactivity items. Gomez et al. (2011) examined DIF using IRT across English and Malay versions of parent and teacher ADHD symptom ratings and found invariant item functioning across languages. Similarly, Gomez (2007) found minimal DIF across child gender for parent ADHD symptom ratings.

More recently, Makransky and Bilenberg (2014) used the Rasch partial credit model to examine whether ADHD symptom items function similarly across gender and age for parent and teacher ratings of 566 Danish children between 6 to 16 years old. Two parent-rated IA items (*Fails to give close attention to details*, *Does not seem to listen when spoken to directly*) and two teacher-rated IA items (*Loses things necessary for tasks*, *Easily distracted*) displayed DIF by age with higher endorsement of *Fails to give close attention* and *Loses things* for older children, and higher endorsement of *Does not seem to listen* and *Easily distracted* for younger children. No IA items displayed DIF by gender. Two parent-rated HI items

[2] The 1-PL IRT and the dichotomous Rasch models are mathematically the same, with one free item parameter—item difficulty—for estimation. Both the 2- and 3-PL IRT models include an additional item parameter—item discrimination. Further, the 3-PL IRT estimates an extra pseudo-guessing parameter. In spite of their mathematical and measurement connection with each other, many researchers see the Rasch model as being fundamentally different from other IRT models (e.g., 1-, 2-, and 3-PL IRT). For instance, the Rasch model is more prescriptive than descriptive, requiring data to fit the model expectations, rather than the other way to explain as much variance in data as possible (Andrich, 2004; Boone, Staver, & Yale, 2014; Linacre, 2005).

(*Leaves seat in classroom*, *Talks excessively*) and three teacher-rated HI items (*Fidgets with hands or feet*, *Runs about or climbs excessively*, *Talks excessively*) displayed DIF by gender with higher endorsement of fidgets and runs about for boys, and talks excessively for girls. No HI items displayed DIF by age. Makransky and Bilenberg interpreted these findings to indicate that boys and girls with high levels of ADHD have different ways of expressing symptoms (e.g., boys more likely to fidget and leave seat, while girls more likely to talk excessively), which is consistent with previously reported concerns about the gender appropriateness of ADHD symptoms (Ohan & Johnston, 2005), and that parents and teachers have higher behavioral expectations for older students.

Gaps in Extant Literature

Assessment of ADHD symptoms should be conducted in a manner that is developmentally appropriate and culturally sensitive. Yet, the development of ADHD symptoms for the DSM (Lahey et al., 1994) and most studies of ratings that include items reflecting these symptoms (e.g., Wolraich et al., 2003) have emphasized traditional psychometric characteristics (i.e., reliability and validity) and ability to differentiate between diagnostic groups and typically developing children. Although there has been some emphasis on strategies to conduct assessments of ADHD in a developmentally appropriate way (e.g., separate norms for age), there has been little focus on examining the degree to which symptom reports are sensitive to child characteristics. Furthermore, the extant literature provides some information regarding DIF for parent and teacher ADHD symptom ratings across child age and gender, but fewer studies have examined DIF for race and ethnicity. In addition, we were unable to locate any prior studies using Rasch model analyses to examine DIF for ADHD symptom ratings across child demographic characteristics in a large, diverse, community sample. This is a critical gap in the literature because informant reports are important components of a comprehensive approach to diagnosing ADHD in youth, particularly by providing normative comparisons for evaluating the frequency and severity of IA and HI symptoms exhibited by an individual child. Given consistent evidence of mean differences in parent and teacher ADHD symptom dimension ratings across child age, gender, race and ethnic groups, it is necessary to explore measurement invariance at the item (i.e., symptom) level. It is possible that differences in symptom dimension ratings across gender, age, race, and ethnicity are primarily due to differences for a limited number of symptom items and not all or most items on the scale. Rasch rating scale model is used to help identify which items vary as a function of demographic characteristics and which are invariant and, thus, could inform directions for making assessment of ADHD more developmentally appropriate and culturally sensitive.

Purpose of Study

Prior studies with the ARS-5 data set focused on reliability and validity of ADHD IA and HI dimensions and functional impairment (DuPaul, Reid, Anastopoulos, Lambert, Watkins, & Power, 2016b; DuPaul, Reid, Anastopoulos, & Power, 2014; Power, Watkins, Anastopoulos, Reid, Lambert, & DuPaul, 2017). The primary purpose of the current study was to examine whether ARS-5 symptom items function similarly across child age, gender, race (Black vs. White[3]), and ethnic groups. Based on prior findings of mean symptom rating differences and symptom item DIF, we hypothesized that DIF would be found for most inattentive and hyperactive-impulsive symptoms as a function of child demographic characteristics and that greater measurement variance would be found for teacher than parent ratings. Prior to examining DIF, we conducted Rasch analyses to establish the degree to which parent and teacher ratings on the ARS-5 show acceptable levels of measurement reliability and/or differ at the item and scale or dimension level.

Method

Participants

Two separate samples were included in this study. One sample included 2074 parents who completed all 9 symptom items for each subscale (IA and HI). Parents and guardians were predominantly White (64.1%) and ranged in age from 20 to 77 years old ($M = 41.57$; $SD = 8.23$). Most parents were married (79.7%), had at least high school education or greater (89.9%), and were employed (72.3%). The parent sample was recruited from all regions of the US and included households from both metropolitan (86.4%) and non-metropolitan (13.6%) locations. English was spoken in most (89.4%) households. The children ($N = 2079$; 1037 males, 1042 females) rated by the parents ranged in age from 5 to 17 years old ($M = 10.68$; $SD = 3.75$). Children were from White (77.8%), Black (7.9%), Asian (4.3%), and other or multiracial (10.1%) backgrounds. Almost one-fourth of the children (23.2%) were Hispanic. Table 1 provides a breakdown of sample characteristics in terms of child gender, age, race, and ethnicity.

[3] DIF analyses for child race were restricted to children from Black or White backgrounds for two reasons. First, the largest and most consistent racial differences for ADHD symptom dimension ratings have been obtained for Black relative to White children. Second, the cell size for other racial groups (e.g., Asian) was relatively small.

Table 1 Sample Size for Parent and Teacher Ratings by Child Demographic Groups

Characteristic	Parent Ratings		Teacher Ratings	
	Frequency	% in the sample	Frequency	% in the sample
Gender				
Male	1037	49.9	535	50.0
Female	1042	50.1	535	50.0
Age				
5–10	1041	50.1	456	42.6
11–17	1038	49.9	614	57.4
Race				
White	1617	77.8	677	63.3
Black	164	7.9	146	13.6
Asian (incl. PI)	89	4.3	64	6.0
Other (multi-racial)	209	10.1	183	17.1
Ethnicity				
Hispanic	482	23.2	264	24.7
Non-Hispanic	1596	76.8	806	75.3

Note. PI = Pacific Islanders

The second sample included 1070 teachers (766 female, 304 male) who completed ADHD symptom ratings; each teacher rated two randomly selected students (one male, one female) on their class rosters (see Procedures). Thus, these children were different from those rated by parents in the first sample. Teachers were predominantly White, non-Hispanic (87.3%) and reported a mean of 17.95 years of teaching experience (*SD* = 10.7). The teacher sample was recruited from all regions of the US and included general (83.3%) and special education (16.4%) teachers. To ensure independence of data, teacher ratings for only one of the two students were included in analyses for this study (see Procedures). For the selected teacher sample, the students (*N* = 1070; 535 males, 535 females) rated by teachers ranged in age from 5 to 17 years old (*M* = 11.53, *SD* = 3.54; including 42.6% 5–10 years old and 57.4% 11–17 years old) and attended Kindergarten through 12th grade (see Table 1). Most students attended general education classrooms (81.9%). Students were from White (63.3%), Black (13.6%), Asian (6.0%), other or multi-racial (17.1%) backgrounds, with Hispanic totaling 24.7%.

Measures: ADHD Symptom Ratings

Parents and teachers reported the frequency with which each child displayed the 18 symptomatic behaviors of ADHD using the *ADHD Rating Scale-5* Home and School versions (ARS-5; DuPaul et al., 2016a), respectively. The Home and School versions were identical in item wording and format. Parents and teachers indicated the frequency of each behavior on a 4-point Likert scale, including 0 (*never or rarely*), 1 (*sometimes*), 2 (*often*), and 3 (*very often*). For the Home Version, parents were asked to select the number that best described their child's behavior over the previous 6 months. For the School Version, teachers were asked to select the number that best described the student's behavior over the past 6 months or since the beginning of the school year. For adolescents ages 11 and older, additional wording (from the DSM-5) was provided for some items to make these developmentally relevant. The nine inattention (IA) items were listed separately from the nine hyperactivity-impulsivity (HI) items. The scale has adequate levels of internal consistency (coefficient alphas ranging .89 to .97), test-retest reliability (*r*s ranging .80 to .93), and criterion-related validity (moderate to large correlations with Conners Parent and Teacher Rating Scales) (DuPaul et al., 2016a).

Procedures

Parents were recruited through the GfK KnowledgePanel® to provide a sample of children and adolescents that was representative of the US population in terms of race, ethnicity, geographic region, and family income (for details, see DuPaul et al., 2016b). If more than one child between the ages of 5 to 17 was present in a given household, then parents were asked to provide ratings for one randomly selected child such that the number of cases was balanced across gender and age range.

Teacher data were collected via two national research firms: GfK Knowledge Panel® and e-Rewards®. Initially, 1509 teachers on the KnowledgePanel® were assigned to complete ratings. To obtain the desired sample size of 2000 students, additional teachers were recruited through e-Rewards Market Research®. To ensure equal gender representation, all teachers were asked to provide symptom ratings

for one randomly selected boy and one randomly selected girl on their class roster. Secondary school teachers were instructed to provide ratings for one randomly selected male and one randomly selected female in a randomly selected class. Each student selected was based on a randomly generated number provided in the instructions. Thus, for example, the teacher might be asked to select the 7th girl on the class roster. Further, the sample was recruited such that the number of cases was balanced across age and grade range and was representative of the US child population in terms of race and ethnicity, geographic region, age, and sex (for details, see DuPaul et al., 2016b).

Prior to completing online ratings, parents and teachers read information regarding the purpose of the study as well as possible risks and benefits associated with participation. Parents and teachers could opt out of the study at that point. To retain anonymity of ratings, parents and teachers were informed that their completion of the ARS-5 served as their consent to participate. Approval of ethical procedures was provided by insitutitonal review boards at Children's Hospital of Philadelphis, Lehigh University, University of Nebraska-Lincoln, and University of North Carolina-Greensboro.

To ensure independence of data for analyses (i.e., because child ratings were nested within teachers and nesting data structure is not possible incorporate in Rasch rating scale models), one of the two students rated by each teacher was randomly selected with a requirement that half the resultant sample was female and half was male. Thus, the school version sample included symptom ratings for 1070 children.

Data Analyses

We employed the Rasch rating scale model (RSM; Andrich, 1978) to address the research questions. Rasch analyses may yield more stable parameters than a 2- or 3-PL IRT model given the sample sizes for some of the subgroups in our study, particularly for race. Mathematically, the Rasch RSM is given by: $\ln(P_{nij}/P_{ni(j-1)}) = B_n - D_i - F_j$, where P_{nij} and $P_{ni(j-1)}$, respectively, is the probability of respondent n rating in category j and category j–1 of item i, B_n is the *ability* measure (i.e., level of the latent trait) of respondent n, D_i is the *difficulty* measure of item i (i.e., the point on the latent variable at which the highest and lowest categories of the item are equally probable for item endorsement), and F_j is the common threshold structure (also called *step measures*) shared by all items (the point where categories *j-1* and *j* are equally probable relative to the measure of the item). The WINSTEPS program version 4.4.5 (Linacre, 2019a) was used for the Rasch analyses. The key output that addressed our research questions included (a) for each subscale with confirmed unidimensionality, category probabilities, item fit statistics, reliability indices for both persons and items, and the variable maps (Research Question 2), (b) comparison of item and scale functioning across the parent and teacher samples (Research Question 2), and (c) differential item functioning (DIF) by gender, age, race, and ethnicity (Research Question 1).

Dimensionality of Data Rasch models have a fundamental assumption of data—unidimensionality, requiring that the items on a scale measure only one underlying trait, or a dominant one. However, using a unidimensional model for truly multidimensional data, in spite of its parsimony, would lose the diagnostic information from different dimensions of the construct. The dimensionality of data can be assessed as an inherent step of Rasch analysis in WINSTEPS, using a principal components analysis (PCA) of standardized residuals (Smith Jr., 2002; Smith & Miao, 1994). Our preliminary overall Rasch analysis for each sample suggested splitting entire scale into two subscales, each analyzed with a separate Rasch analysis.

Rating Scale Functioning The rating scale for each subscale is expected to meet the criteria in Linacre (2002): (1) There are at least 10 observations per category for stable estimation of the thresholds; (2) The mean square outfit statistic for each category is less than 2.0 to indicate less noise than information in the responses to a given category; (3) Observed average measures for categories are ordered so that each consecutively higher number on the rating scale corresponds to higher levels of trait; and (4) Thresholds are ordered so that as one moves up the continuum of the trait, each rating scale category in turn becomes the most probable response.

Rasch Fit Analysis Rasch fit statistics (in the forms of infit and outfit) for items in the WINSTEPS output are akin to chi-squares to indicate discrepancies between observed responses and Rasch model expectations. Because standardized item fit statistics may be over-powered for large N (Linacre, 2019b), we used unstandardized mean squares fit statistics. The expected value being 1.0, the range 0.5 to 1.5 is deemed acceptable to support accurate measurement. For this study, values above 1.5 indicate underfit (noise due to unusual or inappropriate response patterns), whereas values less than 0.5 are viewed as overfit (too little variation in the response pattern). Items of underfit are of concern but presence of overfit is not (Smith Jr., 2005).

Rasch Reliability Indices Rasch reliability estimates for both persons and items, ranging from 0.0 to 1.0, are indices of how well the persons or items are spread out along the continuum and how reproducible their ordering is. Conceptually, Rasch person reliability is analogous to Cronbach's alpha in true score theory. Rasch item reliability is an important aspect for construct validation because a spread of items is required to form a well-defined variable (Smith Jr., 2001).

Item-Person Variable Map The variable map (also called Wright Map) from each Rasch analysis provides a hierarchy of the item endorsement difficulty, which adds further validity evidence as it is checked against (a) what would be theoretically and conceptually expected and (b) the person distribution along the same logit scale of measurement. The item-person mapping can visually reveal mistargeting of items at persons, or ceiling or floor effects. The issues found may justify future scale optimization by removing redundant existing items of same likelihood for endorsement or by adding new items for gaps in the item hierarchy.

Differential Item Functioning Assuming an equal amount of latent trait for two subgroups (e.g., Rasch HI dimension severity for boys and girls), no substantial contrast is expected to occur between their average Rasch item difficulty measures. To address our primary research question, DIF contrast sizes (rather than *statistical* significance) across gender (male vs. female), age (Young—5-10 vs. Old—11-17 years old), race (White vs. Black) and ethnicity (Hispanic vs. non-Hispanic) were examined. We utilized the standard effect size (ES) or DIF classification scheme that was recommended by Longford, Holland, and Thayer (1993) and transferred to the context of Rasch models by Paek (2002): (A) Negligible difference: DIF contrast < |±0.426| logits; (B) Intermediate difference: DIF contrast between |0.426| and |±0.638| logits; and (C) Large difference: DIF contrast over |±0.638| logits. We flagged items displaying intermediate or large DIF contrasts (0.426 logits or more), regardless of directionality for the pairwise contrasts between two sub-groups of each background characteristic variable.

Results

Frequency for response categories of each item are shown in Tables 1A and 2A in online supplementary material (*online* hereafter). Rasch analyses results are presented separately for parent and teacher rating scale data. For each respondent sample, we conducted two Rasch analyses, one for each item subscale—IA and HI. The unidimensionality assumption was checked and supported for each subscale; the first contrast eigenvalue, ranging from 1.46 to 1.96 (Table 3A online), indicates that the strength of a potential secondary dimension for each subscale was less than two items, and thus negligible.

Scale Functioning and Reliabilities for the IA and HI Subscales of both Samples

Rating Categories The rating categories for IA and HI subscales for both parent and teacher ratings satisfactorily met the criteria in Linacre (2002). Namely, there were minimally 10 observations per category; both average category measures and step measures were ordered (see Table 4A online). Consistently, the plots (see Figure 1 online) for the category probabilities indicate no disordering; each category is the more probable one to endorse as the assessed children have higher severity of the trait IA or HI.

Item* Difficulty Measures and *fit statistics For the IA subscale, the item difficulty (i.e., likelihood for endorsement) measures ranged from −0.90 to 0.77 logits for parent ratings, and − 0.88 to 1.07 logits for teacher ratings *(see* Table 5A online*)*. For the HI subscale, the item difficulty measures ranged from −0.96 to 0.94 logits for parents, and from −0.95 to 1.54 logits for teachers (see Table 6A online). All the items had infit and outfit mean square statistics within the acceptable range (i.e., < 1.50; larger than 1.50 fit statistics indicate underfit).

Person and Item Mean Measures and Variance For all subscales of the two samples, the average person measure was lower than the corresponding average difficulty *(see* Table 7A online*)*. By default, the average item difficulty in each Rasch analysis is fixed at 0.0 logit in WINSTEPS. The Rasch person measure variabilities on the four subscales (*SD* ranging from 1.96 to 3.28 logits for entire samples) were larger than the item difficulty variabilities (*SD* ranging 0.52 to 0.78 logits), possibly due to the large person sample size and characteristics as well as the limited number of items on each subscale.

Person and Item Reliabilities All the Rasch reliability information is presented *in* Table 7A (online). For parent ratings, the person reliability for the IA subscale was .82 for the full sample, .86 for non-extreme respondents only (i.e., omitting maximum and minimum possible ratings across all items on the subscale or dimension). The person reliability was lower for the HI subscale, including .61 for full sample and .73 for non-extreme respondents only.

Compared with the parent ratings, children rated by teachers were better separated by items, and thus had higher person reliabilities on each corresponding subscale. For teacher ratings, the person reliability for the IA subscale was .88 for full sample, and .91 for non-extreme respondents only. Similar to the results for parent ratings, the HI person reliabilities were lower than those on the IA subscale (.72 for full sample and .85 for non-extreme respondents only).

The large sample size possibly helped spread items well along the construct scale, and thus the almost perfect reliability (.99) for items on both subscales for both the parent and teacher samples. Alternatively, persons were not so reliably separated by items (i.e., only 9 items, some having similar likelihood for endorsement as indicated by their small *SDs*).

Item-Person Variable Maps In the variable maps (see Figs. 2 and 3 online), children with most severe IA or HI symptoms as

rated by parents or teachers and the items that are least likely to be endorsed are placed at the top of the maps. Consistent with previous findings for person and item mean measures and variability, the maps show a consistent pattern across parent and teacher ratings in terms of the large range and variability for the person distribution on the left side of the logit scale relative to the small range and variability for the item hierarchy on the right side.

Although the children rated by parents and teachers were different from each other, the rank ordering of IA or HI items based on endorsement difficulty, showed some consistency between parents and teachers. For example, as shown in the online Fig. 2 and Table 5A, six IA items (e.g., IA #1 [*Fails to give close attention*], IA #7 [*Loses things necessary*]), had similar rank ordering and differed by less than 0.50 logits in item difficulties between two samples. Alternatively, three items (IA #2 [*Has difficulty sustaining attention*], IA #3 [*Does not seem to listen*], IA #9 [*Forgetful*]) differed substantially in endorsement difficulty (by 0.90 or more logits) and varied in rank ordering between the two sample ratings. Similarly, as shown in the online Fig. 3 and Table 6A, with the exception of two symptoms (HI#3 [*Runs about*], HI#5 [*On the go*]), HI items have largely similar ranking order in endorsement difficulties between parents and teachers.

Differential Item Functioning (DIF) for the IA and HI Subscales of both Samples

Gender DIF Based on ARS-5 parent and teacher ratings, assuming equal IA symptom dimension severity across gender, IA items displayed maximum gender DIF contrasts at 0.36 logits and thus were all negligible effect size (ES $< |\pm 0.426|$ logits; see Table 2). By contrast, four out of the nine HI items rated by parents are worth noting, including: (1) two items (#1—*Fidgets* and #3—*Run about*) had higher probabilities for parents to endorse for boys than for girls; and (2) two items (#6—*Talks excessively* and #9—*Interrupts*) had higher probabilities for parents to endorse for girls than for boys. Two HI items rated by teachers displayed similar DIF, including (1) HI #1 (*Fidgets*) had higher probabilities of endorsement for boys than for girls; (2) HI #6 (*Talks excessively*) had higher probabilities of endorsement for girls than for boys. The HI item #6 displayed DIF with a large ES ($> |\pm 0.638|$ logits) for parent ratings, while the other three items (#1, #3, and #9) had intermediate ES (between $|\pm 0.426|$ and $|\pm 0.638|$ logits).

Age DIF Based on parent ratings, assuming equal IA symptom dimension severity across age groups, two IA items had

Table 2 Rasch DIF analysis across gender for the ARS-5 Inattention (IA) and Hyperactivity-Impulsivity (HI) subscales.

Items	How often does your child display this behavior?	Parents ($N = 2074$)		Teachers ($N = 1070$)	
		(M-F) Contrast	Joint S.E.	(M-F) Contrast	Joint S.E.
IA1	[Fails to give close attention …]	−0.08	0.09	0.14	0.13
IA2	[Has difficulty sustaining attention …]	−0.36	0.10	−0.14	0.13
IA3	[Does not seem to listen when spoken to directly]	0.24	0.10	0.13	0.15
IA4	[Does not follow through …]	0.22	0.10	0.00	0.14
IA5	[Has difficulty organizing tasks and activities]	0.00	0.10	−0.08	0.14
IA6	[Avoids, dislikes, or is reluctant to engage …]	−0.28	0.10	0.12	0.14
IA7	[Loses things necessary for tasks or activities …]	0.00	0.11	−0.18	0.15
IA8	[Easily distracted]	−0.07	0.09	−0.21	0.13
IA9	[Forgetful in daily activities…]	0.29	0.10	0.20	0.14
HI1	[Fidgets with or taps hands or feet or squirms in seat]	−0.44	0.09	−0.55	0.14
HI2	[Leaves seat in situations …]	−0.25	0.10	−0.11	0.14
HI3	[Runs about or climbs in situations … inappropriate]	−0.49	0.12	−0.28	0.19
HI4	[Unable to play or engage … quietly]	−0.11	0.12	0.10	0.16
HI5	[On the go, acts as if driven by a motor]	−0.23	0.10	−0.36	0.16
HI6	[Talks excessively]	0.64	0.09	0.61	0.13
HI7	[Blurts out an answer …]	0.20	0.10	0.15	0.14
HI8	[Has difficulty waiting his or her turn …]	−0.18	0.11	−0.06	0.15
HI9	[Interrupts or intrudes on others]	0.43	0.09	0.27	0.14

Note. DIF contrast is the difference in Rasch difficulty of the item between the two groups. M = Male, F = Female

higher probabilities of endorsement for younger children (5–10 years old) than for adolescents (11–17 years old; *see Table 3*): IA #2—*Has difficulty sustaining attention* (intermediate ES) and IA #8—*Easily distracted* (large ES). Two other IA items had higher probabilities of endorsement for younger children than for adolescents, both with intermediate ES: IA #4—*Does not follow through* and IA #6—*Avoids, dislikes*. Based on teacher ratings, the IA item #4 had higher probabilities of endorsement for younger children than for adolescents (intermediate ES). IA #8 was also more likely for teachers to endorse for younger children than for adolescents, but the DIF contrast (0.42 logits) was still negligible.

The Rasch analysis of parent HI ratings revealed that two HI items had higher probabilities of endorsement for younger children than for adolescents, including HI #2—*Leaves seat* (intermediate ES) and HI #3—*Run about* (large ES) (see Table 3). Alternatively, HI #1—*Fidgets* was less likely to be endorsed for younger children than for adolescents (large ES). For teacher ratings, all HI items displayed age DIF with contrasts below 0.33 logits, and thus negligible ES.

DIF for Ethnicity *Assuming equal IA and HI symptom dimension severity across ethnic groups,* as shown in *Table 4,* only IA #3 (*Does not seem to listen*) and HI #3 (*Run about*) showed DIF above 0.40 logits, barely reaching the threshold for an intermediate ES. HI #3 was more likely for parents to endorse for Hispanic than for non-Hispanic children. Teachers were more likely to endorse IA #3 for Hispanic than for non-Hispanic children. All the other IA or HI items had negligible DIF contrasts.

DIF for Race Assuming equal IA and HI symptom dimension severity between White and Black groups, two IA items showed DIF: (a) IA #7 (*Loses things necessary*) for parent ratings (lower probabilities of endorsement for White than for Black children) and (b) IA #1 (*Fails to give close attention*) for teacher ratings (higher probabilities of endorsement for White than for Black children), both with intermediate ES (*see Table 5*). For HI items, teachers were less likely to endorse HI #3 (*Runs about*), but more likely to endorse HI #6 (*Talks excessively*), for White than for Black children, both with intermediate ES *(Table 5*). All of the parent-rated HI items displayed DIF with negligible ES between White and Black children.

Discussion

Most prior studies of DSM ADHD symptom reports by parents and teachers have focused on dimensional differences across child gender, age, race, and ethnicity (e.g., DuPaul et al., 2016b; Miller et al., 2009; Reid et al., 2000). Although the original field trials of DSM criteria for ADHD examined symptom (item) performance in differentiating between diagnostic groups and typically developing controls (Lahey et al., 1994), minimal attention has been given to how individual symptoms are interpreted by parents and teachers as a function of child gender,

Table 3 Rasch DIF analysis across age for the ARS-5 IA and HI subscales.

Items	How often does your child display this behavior?	Parents ($N = 2074$)		Teachers ($N = 1070$)	
		(Y-O) Contrast	Joint S.E.	(Y-O) Contrast	Joint S.E.
IA1	[Fails to give close attention …]	0.17	0.09	0.07	0.13
IA2	[Has difficulty sustaining attention …]	−0.49	0.10	0.22	0.13
IA3	[Does not seem to listen when spoken to directly]	−0.09	0.10	−0.39	0.15
IA4	[Does not follow through …]	0.44	0.10	0.45	0.14
IA5	[Has difficulty organizing tasks and activities]	0.27	0.10	0.05	0.14
IA6	[Avoids, dislikes, or is reluctant to engage …]	0.52	0.10	0.00	0.14
IA7	[Loses things necessary for tasks or activities …]	0.07	0.11	0.16	0.15
IA8	[Easily distracted]	−0.80	0.09	−0.42	0.13
IA9	[Forgetful in daily activities…]	0.00	0.10	−0.22	0.14
HI1	[Fidgets with or taps hands or feet or squirms in seat]	0.70	0.09	−0.10	0.13
HI2	[Leaves seat in situations …]	−0.44	0.11	−0.11	0.14
HI3	[Runs about or climbs in situations … inappropriate]	−0.67	0.13	−0.19	0.18
HI4	[Unable to play or engage … quietly]	−0.21	0.12	−0.13	0.16
HI5	[On the go, acts as if driven by a motor]	−0.26	0.10	0.30	0.15
HI6	[Talks excessively]	0.02	0.09	0.33	0.13
HI7	[Blurts out an answer …]	0.28	0.10	0.13	0.14
HI8	[Has difficulty waiting his or her turn …]	−0.18	0.11	−0.22	0.14
HI9	[Interrupts or intrudes on others]	0.13	0.09	−0.12	0.14

Note. Y = Young (5–10 years old), O = Old (11–17 years old)

Table 4 Rasch DIF analysis across ethnicity (Hispanic minus Non-Hispanic) for the ARS-5 IA and HI subscales.

Items	How often does your child display this behavior?	Parents ($N = 2074$)		Teachers ($N = 1070$)	
		(H-NH) Contrast	Joint S.E.	(H-NH) Contrast	Joint S.E.
IA1	[Fails to give close attention …]	−0.04	0.11	0.26	0.15
IA2	[Has difficulty sustaining attention …]	0.08	0.13	0.00	0.15
IA3	[Does not seem to listen when spoken to directly]	−0.28	0.12	−0.43	0.16
IA4	[Does not follow through …]	0.21	0.12	−0.17	0.15
IA5	[Has difficulty organizing tasks and activities]	0.04	0.12	0.11	0.15
IA6	[Avoids, dislikes, or is reluctant to engage …]	0.08	0.12	0.00	0.15
IA7	[Loses things necessary for tasks or activities …]	−0.08	0.13	0.00	0.16
IA8	[Easily distracted]	−0.12	0.11	0.26	0.15
IA9	[Forgetful in daily activities…]	0.08	0.12	−0.14	0.16
HI1	[Fidgets with or taps hands or feet or squirms in seat]	0.39	0.11	0.00	0.15
HI2	[Leaves seat in situations …]	−0.22	0.12	−0.25	0.16
HI3	[Runs about or climbs in situations … inappropriate]	−0.42	0.14	−0.16	0.2
HI4	[Unable to play or engage … quietly]	0.00	0.15	0.02	0.17
HI5	[On the go, acts as if driven by a motor]	0.21	0.12	−0.12	0.17
HI6	[Talks excessively]	−0.05	0.10	0.00	0.15
HI7	[Blurts out an answer …]	−0.19	0.12	0.16	0.16
HI8	[Has difficulty waiting his or her turn …]	−0.16	0.13	0.17	0.16
HI9	[Interrupts or intrudes on others]	0.24	0.11	0.06	0.16

age, race, and ethnicity. It is possible that previously obtained group differences at the dimension level could be due to actual variation in symptomatic behaviors but also due to measurement items functioning differently based on child characteristics. Specifically, child characteristics could affect adult perceptions of the frequency of ADHD symptoms such that adults are more likely to report certain symptoms dependent on child age, gender, race, or ethnicity. The current study went beyond prior investigations that have established measurement invariance at the symptom dimension or factor level (e.g., DuPaul et al., 2016b; Leopold et al., 2018) and item level (e.g., Gomez, 2007; Makransky & Bilenberg, 2014), by examining possible measurement invariance across multiple child demographic characteristics (most notably race and ethnicity) at the symptom item level in two large, national samples.

Measurement Invariance across Child Demographic Characteristics

In support of our hypotheses, most (i.e., 12 of 18) ADHD symptom items showed differential functioning across child age, gender, race, and ethnicity with more items showing DIF for age than for gender, ethnicity, or race. Contrary to our hypothesis, more items showed DIF for parent than for teacher ratings in the context of child gender and age.

Gender differences were found for four HI symptoms, but not for any IA symptoms. Not surprisingly and consistent with prior findings of higher mean HI symptom ratings for boys than girls (e.g., Anastopoulos et al., 2018; Burns, Walsh, Gomez, & Hafetz, 2006; Leopold et al., 2018), with symptom dimension severity assumed equal across gender, those HI behaviors involving overt motor activity (*Fidgets, Runs about*) were more likely to be endorsed for boys than girls; while symptoms involving verbal social activity (*Talks excessively, Interrupts*) had higher probabilities of endorsement for girls than for boys. These HI symptom items are similar to those found to show DIF in the Makransky and Bilenberg (2014) study using child participants from Denmark. As has been found previously (e.g., Leopold et al., 2018; Makransky & Bilenberg), parent and teacher ratings of IA symptoms demonstrated measurement invariance across youth gender, indicating that adults perceive IA symptoms in a similar manner for boys and girls.

Child age was the child characteristic that had the greatest impact on adult (particularly parent) symptom frequency perception, thus calling to question the developmental appropriateness of ADHD symptom wording across children between

Table 5 Rasch DIF analysis across race (White minus Black) for the ARS-5 IA and HI subscales.

Items	How often does your child display this behavior?	Parents ($N = 2074$)		Teachers ($N = 1070$)	
		(W-B) Contrast	Joint S.E.	(W-B) Contrast	Joint S.E.
IA1	[Fails to give close attention …]	−0.05	0.18	−0.46	0.19
IA2	[Has difficulty sustaining attention …]	0.28	0.19	−0.04	0.19
IA3	[Does not seem to listen when spoken to directly]	−0.36	0.19	−0.13	0.21
IA4	[Does not follow through …]	−0.14	0.18	−0.03	0.19
IA5	[Has difficulty organizing tasks and activities]	0.05	0.19	0.33	0.19
IA6	[Avoids, dislikes, or is reluctant to engage …]	−0.37	0.19	−0.07	0.19
IA7	[Loses things necessary for tasks or activities …]	0.57	0.19	0.14	0.20
IA8	[Easily distracted]	0.24	0.17	0.25	0.19
IA9	[Forgetful in daily activities…]	−0.19	0.18	0.00	0.20
HI1	[Fidgets with or taps hands or feet or squirms in seat]	−0.23	0.17	−0.32	0.18
HI2	[Leaves seat in situations …]	0.10	0.19	0.27	0.19
HI3	[Runs about or climbs in situations … inappropriate]	−0.19	0.22	0.43	0.23
HI4	[Unable to play or engage … quietly]	0.08	0.21	−0.26	0.21
HI5	[On the go, acts as if driven by a motor]	0.21	0.18	−0.10	0.20
HI6	[Talks excessively]	0.25	0.16	−0.43	0.18
HI7	[Blurts out an answer …]	−0.01	0.18	0.27	0.18
HI8	[Has difficulty waiting his or her turn …]	0.05	0.19	0.22	0.19
HI9	[Interrupts or intrudes on others]	−0.28	0.17	0.05	0.18

ages 5 and 17. Assuming equal ADHD symptom dimension severity across age groups, parents had higher probabilities to endorse HI symptoms involving gross motor activity (*Leaves seat, Runs about*) for younger children than for adolescents. Conversely, parents had higher probabilities to endorse an HI symptom that involves more subtle motor activity (i.e., *Fidgets*) for adolescents than for younger children. This finding could indicate that adults focus on different aspects of physical activity displayed by youth as a function of developmental expectations. Interestingly, Makransky and Bilenberg (2014) did not find DIF for HI symptoms between age groups. Discrepant DIF findings across studies could be due to cross-country cultural differences in parental standards for what behaviors are considered problematic and the threshold for behavioral frequency or severity that must be crossed in order for that behavior to be viewed as impairing.

A similar pattern was found for IA symptoms with parents more likely to endorse some forms of inattention (i.e., *Difficulty sustaining attention, Easily distracted*) for younger children, while manifestations of inattention that involve more independent responsibility (i.e., *Does not follow through, Avoids tasks*) had higher probabilities to be endorsed for adolescents. Again, these DIF findings could reflect parental response to developmental context when considering IA symptom manifestation or DIF could indicate measurement issues for these specific items. Surprisingly, DIF for age for teacher ratings was found for only one IA item (*Does not follow through*, which had higher probabilities of endorsement for adolescents than younger children) and for none of the HI items. Less DIF for teacher ratings was possibly because they have more experience than parents observing children at a given age, specifically under structured, high demand classroom conditions and thus their symptom reports may be less subject to bias for age.

Because prior studies have consistently shown higher mean ADHD symptom ratings for Black relative to White youth (DuPaul et al., 2014; Miller et al., 2009), we expected parents and teachers to show differential perceptions of symptom frequency as a function of child race. However, only 2 (*Loses things necessary for tasks or activities* [rated by teachers], *Runs about* [rated by parents]) of 18 ADHD symptom items had higher probabilities of endorsement by parents or teachers for Black children. In contrast, teachers were more likely to endorse one IA (*Fails to give close attention*) and one HI (*Talks excessively*) symptom for White relative to Black students. The present results suggest that ADHD items generally are functioning similarly for parent ratings of Black and White children, and that only 3 of 18 teacher-rated ADHD items show measurement differences across these racial groups. Thus, prior findings of mean symptom dimension rating differences across racial groups may not be due to measurement variance at the symptom item level.

Prior studies have been equivocal regarding IA and HI dimension rating differences across ethnic groups (e.g., de Ramirez & Shapiro, 2005; DuPaul et al., 2016b). In similar fashion to race, only two symptom items met DIF effect size criteria for child ethnicity. Assuming equal IA symptom dimension severity across ethnic groups, teachers were more likely to endorse *Does not seem to listen* for students from Hispanic relative to non-Hispanic backgrounds. Also, assuming equal HI symptom dimension severity across ethnic groups, parents were more likely to endorse *Runs about* for youth from Hispanic relative to non-Hispanic backgrounds. Thus, for the most part, ADHD symptom ratings demonstrated measurement invariance for ethnicity (i.e., respondents generally endorse ADHD symptom items in a similar manner for children of Hispanic and non-Hispanic background). Of course, this conclusion is tempered by the fact that the Latinx population is heterogeneous with respect to country of origin and cultural practices. Thus, generalization of these findings to the broad Latinx population should be done with caution.

Limitations

Conclusions based on the present findings are limited by several factors. First, we only examined ADHD symptom ratings. Given that symptom-related impairment in academic and social functioning is a critical diagnostic indicator (APA, 2013), it would be important to conduct IRT and Rasch analyses for parent and teacher ratings of child impairment. Second, although children rated by teachers were similar in terms of gender, race, ethnicity, and SES to the US population, White children were overrepresented in the parent rating sample. We only examined DIF for two racial groups (Black vs. White). Future investigations should assess racial DIF with a wider range of racial groups. Third, we did not examine the degree to which informant demographic characteristics impacted DIF findings across child characteristics. Given that female respondents typically provide higher ADHD symptom ratings for male children than do male respondents (Anastopoulos et al., 2018), the potential influence of informant gender, among other respondent characteristics, on measurement invariance at the symptom item level should be explored. Fourth, we examined symptoms separately for IA and HI dimensions as opposed to a multidimensional approach despite the strong corrleations between IA and HI. Finally, we did not examine the degree to which interactions of child demographic characteristics (e.g., gender by age) may impact symptom item ratings. Future studies could examine the degree to which measurement invariance is evident across subgroups of demographic categories (e.g., young Black vs. older Black children).

Implications and Conclusions

Parent and teacher ADHD symptom ratings generally provide reliable indicators of IA and HI and should be used routinely in assessing this disorder. However, it appears that some symptoms may have more or less probabilities for parents and teachers to endorse as a function of child demographic characteristics, particularly age (i.e., child vs. adolescent). Although ADHD symptom rating scales typically provide separate norms based on child age and gender, norms based on IA or HI dimension scores do not account for differential symptom frequency report at the symptom item level.

The findings of this study have implications for the revision of diagnostic systems (e.g., DSM) for determining the presence of ADHD as well as clinical assessment strategies using current symptom criteria. There is evidence that some ADHD symptom items (e.g., *Fidgets, Runs about, Talks excessively*) do not operate as effectively (i.e., do not assess ADHD behaviors equally well) for children and adolescents as a function of their gender, race, and especially age. To address developmental differences in the expression of ADHD,, modification of symptom descriptions beyond those recommended in the DSM-5 are needed (e.g., the item for children, *Fidgets with or taps hands or feet or squirms in seat,* could be modified for adolescents, *Becomes restless when expected to sit for an extended time*). Given current symptom descriptions, clinicians could account for the impact of DIF by not only asking informants to report symptom frequency but also asking them to report whether they consider a specific symptomatic behavior to be a problem and the degree to which the symptom impairs a child's academic or social functioning. Stated differently, the impact of DIF on adult report of symptom frequency might be mitigated, in part, by making decisions about presence or absence of symptoms based on the degree to which a behavior represents a problem and is associated with functional impairment.

The six ADHD symptom items that did not show DIF for any child characteristic (i.e., *Difficulty organizing, Forgetful, Unable to play quietly, On the go, Blurts out answers,* and *Difficulty awaiting turn*) may provide guidance for symptom wording revision. In addition to wording modification to account for developmental differences, symptoms should be described as specifically as possible and should not include multiple behavioral descriptions. For example, the symptom *runs about or climbs on things* should refer to only one behavior, such as *physically moves around when not appropriate.* Furthermore, among items demonstrating meaningful gender DIF findings based on parent ratings, it is important to understand how parents interpret variations in symptom behavior for girls versus boys, and the threshold they apply in determining whether a behavior is a problem and, if so, how severe of a problem. For example, qualitative data could be collected through interviews with separate samples of parents of girls

and boys with ADHD to ascertain what behaviors may describe overactive and impulsive behavior among girls versus boys. Findings could inform how symptom items can be revised to account for gender DIF by identifying behavior descriptions that are commonly used regardless of child sex (i.e., rather than behavior descriptions that are unique for girls or boys).

Issues about ADHD symptom measurement generally were more concerning for parent than for teacher ratings, particularly as a function of child developmental level. Thus, it is important for clinicians to use both parent and teacher ratings, to supplement symptom counts with norm-referenced measures that account for child gender and age, and to corroborate rating scale data with parent (and preferably also teacher) interview information. Given that symptom frequency reports in clinical interviews could be subject to DIF as found for behavior ratings, clinicians should inquire not only about frequency of symptomatic behaviors, but also whether the symptom is viewed as a problem and the degree to which the behavior impairs child academic and social functioning.

Our findings from two large national samples of parents and teachers indicate that ADHD symptom items show differential functioning based on child age and gender, but generally not for child race and ethnicity. Thus, although ADHD symptoms generally show cross-cultural invariance, parent and teacher perceptions of symptom frequency are affected by child gender and developmental level. These findings call into question the common practice of attributing reduction in ADHD symptom frequency from childhood to adolescence solely due to child maturation. Also, clinicians should be cautious in interpreting rating scale data when evaluating youth for ADHD, especially when using parent ratings. Parent symptom reports are particularly affected by child age so adjunctive measures (e.g., diagnostic interview, teacher ratings, direct observations of classroom behavior) and exploration of parental expectations for child behavior based on age are needed. Further examination is needed regarding the developmental and cultural appropriateness of ADHD diagnostic symptoms as well as adult perceptions of symptom-related impairment in academic and social functioning that may be affected by child demographic characteristics.

Compliance with Ethical Standards

Ethical Approval All study procedures were in accordance with ethical standards of the institution and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Informed consent was obtained from all individual participants included in the study.

Conflict of interest Drs. Anastopoulos, DuPaul, Power, and Reid receive royalties for the *ADHD Rating Scale-5*.

References

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

Anastopoulos, A. D., Beal, K. K., Reid, R. J., Reid, R., Power, T. J., & DuPaul, G. J. (2018). Impact of child and informant gender on parent and teacher ratings of attention-deficit/hyperactivity disorder. *Psychological Assessment.* Advance online publication. https://doi.org/10.1037/pas0000627

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–573.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 143–166). Maple Grove: JAM Press.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Rasch analysis in the human sciences. Dordrecht: Springer.

Burns, G. L., Walsh, J. A., Gomez, R., & Hafetz, N. (2006). Measurement and structural invariance of parent ratings of ADHD and ODD symptoms across gender for American and Malaysian children. *Psychological Assessment, 18*, 452–457.

Chorozoglou, M., Smith, E., Koerting, J., Thompson, M. J., Sayal, K., & Sonuga-Barke, E. J. S. (2015). Preschool hyperactivity is associated with long-term economic burden: Evidence from a longitudinal health economic analysis of costs incurred across childhood, adolescence and young adulthood. *Journal of Child Psychology and Psychiatry, 56*, 966–975.

de Ramirez, R. D., & Shapiro, E. S. (2005). Effects of student ethnicity on judgments of ADHD symptoms among Hispanic and white teachers. *School Psychology Quarterly, 20*, 268–287.

DuPaul, G. J., Power, T. J., Anastopoulos, A. D., & Reid, R. (2016a). *ADHD rating Scale-5 for children and adolescents: Checklists, norms, and clinical interpretation*. New York: Guilford.

DuPaul, G. J., Reid, R., Anastopoulos, A. D., Lambert, M. C., Watkins, M. W., & Power, T. J. (2016b). Parent and teacher ratings of attention-deficit/hyperactivity disorder symptoms: Factor structure and normative data. *Psychological Assessment, 28*, 214–225.

DuPaul, G. J., Reid, R., Anastopoulos, A. D., & Power, T. J. (2014). Assessing ADHD symptomatic behaviors and functional impairment in school settings: Impact of student and teacher characteristics. *School Psychology Quarterly, 29*, 409–421.

Gomez, R. (2007). Testing gender differential item functioning for ordinal and binary scored parent rated ADHD symptoms. *Personality and Individual Differences, 42*, 733–742.

Gomez, R. (2008a). 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.

Gomez, R. (2008b). Parent ratings of the ADHD items of the disruptive behavior rating scale: Analyses of their IRT properties based on the generalized partial credit model. *Personality and Individual Differences, 45*, 181–186.

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.

Lahey, B. B., Applegate, B., McBurnett, K., Biederman, J., Greenhill, L., Hynd, G. W., et al. (1994). DSM-IV field trials for attention deficit hyperactivity disorder in children and adolescents. *American Journal of Psychiatry, 151*, 1673–1685.

Leopold, D. R., Christopher, M. E., Olson, R. K., Petrill, S. A., & Willcutt, E. G. (2018). Invariance of ADHD symptoms across sex and age: A latent analysis of ADHD and impairment ratings from early childhood into adolescence. *Journal of Abnormal Child Psychology. Advance online publication.* https://doi.org/10.1007/s10802-018-0434-6.

Li, J. J., Reise, S. P., Chronis-Tuscano, A., Mikami, A. Y., & Lee, S. S. (2016). Item response theory analysis of ADHD symptoms in children with and without ADHD. *Assessment, 23*, 655–671.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*, 85–106.

Linacre, J. M. (2005). Rasch dichotomous model vs. one-parameter logistic model. *Rasch Measurement Transactions, 19*(3), 1032.

Linacre, J. M. (2019a). *Winsteps® (Version 4.4.5)* [computer software]. Beaverton, Oregon: Winsteps.com.

Linacre, J. M. (2019b). Winsteps® Rasch measurement computer program User's guide. Beaverton, Oregon: Winsteps.com.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171–196). Hillsdale: Lawrence Erlbaum Associates, Inc..

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Oxford: Addison-Wesley.

Makransky, G., & Bilenberg, N. (2014). Psychometric properties of the parent and teacher ADHD rating scale (ADHD-RS): Measurement invariance across gender, age, and informant. *Assessment, 21*, 694–705.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrica, 47*, 149–174.

Merrell, C. & Tymms, P. (2003, April). Rasch analysis of inattentive, hyperactive, and impulsive behaviour in young children and the link with academic achievement. Paper presented at the American Educational Research Association annual meeting, Chicago IL.

Miller, T. W., Nigg, J. T., & Miller, R. L. (2009). Attention deficit hyperactivity disorder in African American children: What can be concluded from the past ten years? *Clinical Psychology Review, 29*, 77–86.

Ohan, J. L., & Johnston, C. (2005). Gender appropriateness of symptom criteria for attention-deficit/hyperactivity disorder, oppositional defiant disorder, and conduct disorder. *Child Psychiatry and Human Development, 35*, 359–381.

Paek, I. (2002). *Investigation of differential item functioning: Comparisons among approaches, and extension to a multidimensional context (unpublished doctoral dissertation)*. Berkeley: University of California.

Polanczyk, G. V., Willcutt, E. G., Salum, G. A., Kieling, C., & Rohde, L. A. (2014). ADHD prevalence estimates across three decades : An updated systematic review and meta-regression analysis. *International Journal of Epidemiology, 43*2), 434–442. https://doi.org/10.1093/ije/dyt261.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23–37.

Power, T. J., Watkins, M. W., Anastopoulos, A. D., Reid, R., Lambert, M. C., & DuPaul, G. J. (2017). Multi-informant assessment of ADHD symptom-related impairments among children and adolescents. *Journal of Clinical Child & Adolescent Psychology, 46*, 661–674.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (expanded edition, 1980. Chicago: University of Chicago Press.)

Reid, R., Riccio, C. A., Kessler, R. H., DuPaul, G. J., Power, T. J., Anastopoulos, A. D., Rogers-Adkinson, D., & Noll, M. B. (2000). Gender and ethnic differences in ADHD as assessed by behavior ratings. *Journal of Emotional and Behavioral Disorders, 8*, 38–48.

Reynolds, R. C., & Kamphaus, W. R. (2015). *Behavior Assessment System for Children-3rd ed. (BASC-3).* Bloomington, MN: Pearson.

Smith Jr., E. V. (2001). Reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2*, 281–311.

Smith Jr., E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*, 205–231.

Smith Jr., E. V. (2005). Effect of item redundancy on Rasch item and person estimates. *Journal of Applied Measurement, 6*, 147–163.

Smith, R. M., & Miao, C. Y. (1994). Assessing unidimensionality for Rasch measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2, pp. 316–327). Norwood: Ablex.

Tymms, P., & Merrell, C. (2011). ADHD and academic attainment: Is there an advantage in impulsivity? *Learning and Individual Differences, 21*, 753–758.

Wolraich, M. L., Lambert, W., Doffing, M. A., Bickman, L., Simmons, T., & Worley, K. (2003). Psychometric properties of the Vanderbilt ADHD diagnostic parent rating scale in a referred population. *Journal of Pediatric Psychology, 28*(8), 559–567.

Young, D. J., Levy, F., Martin, N. C., & Hay, D. A. (2009). Attention deficit hyperactivity disorder: A Rasch analysis of the SWAN rating scale. *Child Psychiatry and Human Development, 40*, 543–559.