Journal of Psychopathology and Behavioral Assessment (2021) 43:778–792
https://doi.org/10.1007/s10862-021-09874-3

Do Parent and Teacher Ratings of ADHD Reflect the Same Constructs? A Measurement Invariance Analysis

Colleen M. Jungersen [1] · Christopher J. Lonigan [2]

Accepted: 8 February 2021 / Published online: 5 June 2021

Abstract

Discrepancies between parent and teacher ratings of problem behaviors related to Attention-Deficit/Hyperactivity Disorder (ADHD) are reported frequently. Previous studies have hypothesized that these discrepancies are the results of various informant biases and have evaluated whether the rating scales are measuring behaviors the same way across informants. The purpose of this study was to evaluate if two rating scales of ADHD behavior, the Strengths and Weaknesses of ADHD Symptoms and Normal Behavior Rating Scale (SWAN) and the Conners' Teacher Rating Scale-15 (CTRS-15), reflected the same underlying constructs across parent and teacher report. Measurement invariance analyses were conducted using parent and teacher report data from a sample of 1645 preschool to fifth-grade children (age range 46 to 169 months) that was comprised of roughly equal number of boys and girls and had racial/ethnic diversity similar to the community (i.e., 61% White, 22% Black/African American; 4% Hispanic/Latino). Although it was hypothesized that both rating scales would demonstrate measurement invariance across parent and teacher report, at least partial weak measurement invariance was only supported for the CTRS-15 across all grade groups. These results indicate that the meaning of any rating discrepancies on the SWAN are unknown because it is not reflective of the same underlying constructs across parents and teachers across all of the examined grade groups. In general, these results have potentially important implications regarding research on ADHD symptoms and related behaviors, and raise questions regarding the utility and measurement of ADHD symptoms.

Keywords ADHD · Rating scales · Parent-teacher agreement · Measurement invariance

Although parent and teacher ratings of behavior are commonly used in both clinical work and research as independent sources of information regarding children's problem behaviors, studies indicate consistent discrepancies between these ratings (e.g., Achenbach et al., 1987; Willcutt et al., 2012; Stranger & Lewis, 1993; De Los Reyes and Kazdin, 2005). These rating discrepancies can lead to issues concerning the diagnosis of certain psychiatric disorders, such as Attention-Deficit/Hyperactivity Disorder (ADHD) for which a diagnostic criterion is that symptoms must be present across at least two contexts of a child's life (Diagnostic and Statistical Manual of Mental Disorders, 5th Edition [DSM-5] American Psychiatric Association, 2013), typically school and home. A rating discrepancy for a child with ADHD could lead to a non-diagnosis of the disorder because the child would not meet the criterion of cross-context presentation of symptoms. This could result in a lack of necessary intervention. Previous studies have indicated that early detection and intervention can help children with ADHD and related problem behaviors decrease the chances of long-term negative effects (e.g. Rabiner et al., 2000; Massetti et al., 2008; Sonuga-Barke et al., 2011) that have been demonstrated in academic achievement (e.g., Lonigan et al., 1999; Sims and Lonigan, 2013; Walcott et al., 2010).

Hypotheses that have been proposed to explain rating discrepancies between parents and teachers have tended to focus on rater biases related to factors such as racial, cultural, or demographic differences between teachers and children (e.g., Javo et al., 2000; Treutler and Epkins, 2003), parental stress levels or mental health (e.g., Youngstrom et al., 2000; Richters & Pellegrini, 1989; Richters, 1992), or parents' lack of familiarity with age-appropriate behaviors (e.g., Antrop

✉ Colleen M. Jungersen
jungersen@psy.fsu.edu

✉ Christopher J. Lonigan
lonigan@psy.fsu.edu

[1] Department of Psychology, Florida State University, 1107 W. Call Street, Tallahassee, FL 32306-4301, USA

[2] Department of Psychology and Florida Center for Reading Research, Florida State University, 1107 W. Call Street, Tallahassee, FL 32306-4301, USA

et al., 2002). However, Narad et al. (2015) argued that before examining reporter biases as reasons for rater discrepancies, it is important to examine the measurement characteristics of the rating scales used. It is possible that discrepancies between parent and teacher ratings exist because the ratings scales used reflect different underlying constructs when used by different raters. Measurement invariance testing can be used to assess whether or not rating scales used for ADHD and related behaviors are measuring the same underlying constructs across parents and teachers by examining if the items of these rating scales have the same meaning across informants. The purpose of the current study was to examine measurement invariance of two rating scales used for both parent and teacher ratings of ADHD-related behaviors to determine if rating discrepancies could be the result of a lack of measurement invariance across informants.

Evidence of Rating Discrepancies

Parent and teacher ratings of behavior are often discrepant (e.g., Willcutt & Pennington, 2000; Willcutt et al., 2000; Murray et al., 2007; Willcutt et al., 2012, De Los Reyes et al., 2015). The agreement between ratings from informants who interact with children in different contexts, such as parents and teachers, tends to be low, with a correlation of about .28 (Achenbach et al., 1987; Touliatos & Lindholm, 1981). Previous studies have demonstrated that the most consistent ratings tended to be from informants who interact with children in similar contexts, such as parents and other caregivers or teachers and teachers' aides (Stanger and Lewis, 1993; De Los Reyes et al.,) and that agreement between raters from different contexts tended to be too low to allow for one informant to substitute for another. Touliatos and Lindholm (1981) reported that a third of children who were rated as free of problem behaviors by their parents were rated as demonstrating at least one problem behavior at school and that approximately half of the children who were rated by their parents as having significant problem behaviors at home were reported to be free of problem behaviors at school. Whereas differences in reports of problem behaviors may reflect actual differences in behaviors across settings, it is likely that at least some of these rating differences represent rating discrepancies for similar behavior across settings.

Discrepancies in ratings are common even when the same rating scale is used by parents and teachers (e.g., Achenbach et al., 1987; Willcutt et al., 2012; Stranger & Lewis, 1993; Lee et al., 2014; De Los Reyes et al., 2015). This has been demonstrated on various commonly used measures of ADHD. For instance, parent and teachers have demonstrated significantly discrepant responses when completing the Strengths and Weaknesses of ADHD Symptoms and Normal Behavior rating scale (SWAN; Gooch et al., 2017), the Conners' Teacher Rating Sale (CTRS; Tripp et al., 2006), Child Behavior Checklist (CBCL; De Los Reyes et al., 2015) and Behavior Assessment System for Children (BASC; Harvey et al., 2013).

Although rating discrepancies exist across many measures of ADHD-related behavior, many of these measures are structured in very different ways and require informants to respond to a variety of different items. For instance, the items of the SWAN are worded in an asymptomatic direction, but the items of the CTRS are worded in a symptomatic direction. The SWAN requires informants to report how often a particular behavior occurs in comparison to "an average same-aged child," but the CTRS simply requires informants to report the relative frequency with which a behavior occurs. Additionally, the SWAN has bidirectional rating that captures both symptomatic and asymptomatic levels of behaviors, but the CTRS captures behavior level only in the symptomatic direction.

These differences in structure and phrasing of the rating scales could contribute to some of the rating discrepancies seen between informants. For example, teachers are more likely than parents to be familiar with average behaviors of a child of a given age (Antrop et al., 2002). This could lead to a potential discrepancy for the SWAN but not the CTRS. A rating discrepancy on the SWAN could also result from the use of DSM phrasing. Some of the vocabulary of items such as "Engaging in tasks that require sustained mental effort," "Ignores extraneous stimuli," or "Modulates motor activity," may lead to different interpretations by different informants. In contrast, the items on the CTRS are not phrased like the DSM, and, therefore, items such as "Inattentive, easily distracted," "Short attention span," or "Restless, always up and on the go," may contain vocabulary that is more commonly used across all types of informants, which may result in less of a rating discrepancy.

Hypotheses Regarding Rating Discrepancies

Several hypotheses have been proposed to explain the discrepancy of parent and teacher ratings of problem and ADHD-related behaviors. For example, some researchers have hypothesized that demographics of the children, such as ethnicity and race, and socioeconomic status are associated with the differences between parent and teacher ratings of problem behaviors with children of certain racial and ethnic backgrounds tending to be more likely to be rated as having behavior problems by teachers of a different background (Harvey et al., 2013; Javo et al., 2000; Treutler and Epkins, 2003). Another hypothesis is that even if symptom behaviors remain the same across settings, teachers tend to be more familiar with developmental norms and therefore perceive some behaviors as developmentally appropriate or developmentally inappropriate, whereas parents often lack a frame of reference for their own child's behavior (Antrop et al., 2002;

Amador-Campos, et al., 2006). In addition, parental stress levels (Youngstrom et al., 2000) and parental mental health (Richters & Pellegrini, 1989; Richters, 1992) have been hypothesized to be related to how parents rate their children's behaviors. Despite the many hypotheses that have been proposed about rating discrepancies, they are not able to fully account for the rating discrepancies that exist across parent and teacher ratings (De Los Reyes and Kazdin, 2005; De Los Reyes et al., 2015). For example, although some studies have reported a negative relation between SES and rating discrepancies (Duhig et al., 2000), other findings have indicated that no such association exists when controlling for other characteristics such as parental psychopathology (Chi and Hinshaw, 2002).

Due to the inconsistent findings across studies regarding rater biases as the source of rating discrepancies, some authors have hypothesized that it is something about the rating forms and not reporter bias that results in rater discrepancies (e.g., Narad et al., 2015). For example, parents and teachers may interpret items on the rating scale differently, resulting in objectively similar behaviors being rated differently. This discrepancy would not be the result of bias from either informant but rather their interpretation of the phrasing of each specific item on the rating scales. Narad et al. examined this hypothesis with the Vanderbilt ADHD Rating Scales. Due to the large size of their sample, measurement invariance was evaluated using a method in which comparisons were based on differences in fit indices rather than using chi-square difference tests (Burns et al., 2006). Although they did not find evidence that the underlying constructs were being measured differently based on informant, Narad et al. also suggested that further research was needed that examined the measurement of constructs of ADHD-related behaviors using other commonly used rating scales and more diverse samples for more generalizable results.

Across the literature on rating discrepancies, the work by De Los Reyes and colleagues provides some insight into why these rating discrepancies exist. In a meta-analytic framework, De Los Reyes et al. (2015) examined rating discrepancies between informants regarding various types of psychopathology and problem behaviors. They suggested that rating discrepancies can occur for a variety of reasons but that it was important to consider that each informant brings their own "unique and valid perspective" to their ratings. To reconcile that informants provide their own unique perspectives but also examine issues behind rating discrepancies, De Los Reyes et al. (2013) developed the Operations Triad Model (OTM).

The OTM describes three ways that research conclusions can be drawn from informant reports: converging operations, diverging operations, and compensating operations. Converging operations refer to measurement conditions that allow for the interpretation of similar patterns and therefore similar conclusions regarding the behavior being rated by both informants. In this situation, the level of agreement between informant ratings would be high, and this consistency would likely demonstrate the presence or absence of the problem behaviors being assessed. Diverging operations refer to measurement conditions that result in inconsistent ratings between informants due to variations in the behaviors of the child. Such situations are explained well by De Los Reyes and Kazdin's (2005) Attribution Bias Context Model, which suggests that some rating discrepancies may be the result of different observable behaviors of the child across contexts. Finally, the compensating operations refer to measurement conditions that result in inconsistent ratings across informants that are the result of methodological issues such as error with the measures or biases of the informants. To test for measurement conditions that would allow for compensating operations to occur, the methodological tools, in this case the rating forms, should be examined for characteristics such as measurement invariance across informants. If measurement invariance was supported, it would indicate that the measures being used to assess certain problem behaviors are representative of the same constructs across informants, and it can be assumed that any rating discrepancies are the result of differing observable behaviors across contexts. If measurement invariance is not supported, the rating discrepancies could be the result of fundamental issues with the rating scales themselves which can contribute to diagnostic and research problems.

Measurement Invariance

One way to empirically determine if a measure operates similarly across contexts (e.g., informant, setting, age group, gender) is to evaluate measurement invariance using confirmatory factor analysis (CFA). The purpose of examining measurement invariance is to systematically evaluate the fit of the model across groups with each step involving more stringent rules to determine how the model fits the data across groups. The first step demonstrates that items are associated with the same underlying constructs across groups. The next step examines the degree to which each of the items contributes to the factor or construct. If weak or metric measurement invariance is supported in the second step, the next step would be to examine if the mean differences for the factor account for the variance of each of the items loading onto that factor across groups. Finally, the last step demonstrates that the variance that each item does not share with the factor is the same across groups.

Measurement invariance has been used in previous studies to examine whether underlying constructs were not being measured consistently across raters. For example, Burns et al. (2014) examined agreement between mothers and fathers, and teachers and teacher's aides on the Child and Adolescent Disruptive Behavior Inventory (CADBI).

Measurement invariance was first examined between parents and between teachers and aides. After measurement invariance was supported in the within-setting pairs, measurement invariance was examined, and supported, across the informants from home and school settings. The invariance across raters and settings demonstrated that the underlying constructs of the ADHD relevant scales of the CADBI were being measured consistently. Measurement invariance across informants was also supported in findings by Burns et al. (2013) and Narad et al. (2015).

Although measurement invariance across informants was supported in these three studies, non-invariance of ADHD-related behaviors and symptoms has been demonstrated as well. For example, Makransky and Bilenberg (2014) reported non-invariance between parent and teacher ratings using a modified/extended version of the ADHD Rating Scales (mADHD-RS). Their results suggested that this finding was likely the result of parents and teachers having a different frame of reference with which they evaluated behaviors. In addition, Vitoratou et al. (2019) examined measurement invariance of the specific ADHD symptoms and reported non-invariance for seven of the nine items representing inattentive behaviors and for six of the nine items representing hyperactive/impulsive behaviors. The authors argued that this was not due to bias, and that the results indicated that the different informants were rating different kinds of behaviors across settings. A third example of non-invariance of ADHD-related behaviors was reported by DuPaul et al. (2020). Measurement invariance was examined in this study in the context of child characteristics such as age, gender, and race. The results indicated that when taking these characteristics into consideration measurement invariance was not supported across informants for 12 of the 18 items examined.

Each of the aforementioned studies used different rating scales, which highlights the need to assess measurement invariance across informants on the various rating scales that have demonstrated rating discrepancies between parents and teachers before examining if there is a rating discrepancy due to rater biases. In addition, inconsistent results are demonstrated across many of these studies. Although some studies support measurement invariance between parent and teacher ratings (Burns et al., 2013; Burns et al., 2014; Narad et al., 2015), others report non-invariance between informants (Makransky and Bilenberg, 2014; Vitoratou et al., 2019; DuPaul et al., 2020). The range of findings speaks to the need for additional research to examine measurement invariance in the study of ADHD-related behaviors. The studies that did not support invariance examined the DSM symptoms of ADHD, and the studies with results that supported measurement invariance examined rating scales that only assessed ADHD-related behaviors. This pattern of results could call into question the ways that DSM symptoms and related behaviors are measured on various rating forms.

Current Study

Although there have been extensive demonstrations of discrepancies between parent and teacher report of ADHD symptoms, few studies have examined the degree to which the underlying symptom dimensions of ADHD—inattention and hyperactivity/impulsivity—are measured similarly by parent and teacher ratings. Consequently, the aim of the current study was to evaluate measurement invariance to assess whether parents' and teachers' ratings of problem behaviors were consistent on two measures of ADHD-related behaviors, the SWAN (Swanson et al., 2012; Lakes et al., 2012) and the CTRS-15 (Purpura and Lonigan, 2009).

Measurement invariance was evaluated in a relatively large sample of preschool and elementary-school age children; however, because of evidence of developmental changes in ADHD across this age-range and empirical evidence of differences in the dimensionality across ages (e.g., Allan and Lonigan, 2019; Biederman et al., 2000; Larsson et al., 2011), measurement invariance was examined in smaller grade groups. It was expected that across all grade groups both the SWAN and the CTRS-15 would demonstrate measurement invariance across parent and teacher ratings, indicating that both rating scales reflected the same underlying constructs across informants.

Method

Participants

Data for this study came from a larger study concerning the development of reading skills. The sample included 1645 children (329 children in preschool, 190 children in kindergarten, 283 children in first grade, 198 children in second grade, 254 children in third grade, 190 children in fourth grade, and 201 children in fifth grade). Children were recruited from 288 classrooms in 36 preschools and schools in north Florida. The children included in the sample for this study ranged in age from 46 months to 169 months, were comprised of approximately equal numbers of girls and boys, and were students for whom at least a teacher or a parent report form had been completed. The majority of the sample was White (60.8%) or Black/African American (22.4%). The remaining children in the sample were Asian (2%), Hawaiian/Pacific Islander (2%) Native American (< 1%), or multi-racial/not reported (10.2%). Four percent of the sample identified as Hispanic/Latino.

Measures

Strengths and Weaknesses of ADHD Symptoms and Normal Behavior Rating Scale (SWAN) This measure is a parent- or

teacher-rating scale. The ratings on this measure are based on what parents or teachers believe to be average behaviors for the child's age. A score of −3 indicates "far below average" and a score of 3 indicates "far above average" with whole number intervals in between (Swanson et al., 2012; Lakes et al., 2012). The SWAN has three subscales: inattention, hyperactivity/impulsivity, and oppositional defiant disorder (ODD). Only the inattention and hyperactivity/impulsivity subscales were used in the analyses for the current study. The inattention subscale consists of nine questions related to inattentive behaviors that are present in the inattentive and combined presentations of ADHD. The hyperactivity/impulsivity subscale consists of nine questions that are related hyperactive/impulsive behaviors that are present in the hyperactive/impulsive and combined presentations of ADHD. A previous study indicated that the validity of the SWAN was large at ($r = .54$; Lakes et al., 2012). The SWAN had high internal consistency in this sample (inattention subscale $\alpha = .95$; hyperactivity/impulsivity subscale $\alpha = .95$).

Conners' Teacher Rating Scale-15 (CTRS-15) This measure is a parent- or teacher-rating scale. This measure was developed to be a brief adaptation of the CTRS (Conners, 1969) that mapped onto three dimensions of problem behaviors: inattention, hyperactivity/impulsivity, and oppositional behaviors. It consists of 15 questions that correspond to the three dimensions of problem behaviors and reflect the behavior scales of the CTRS (Purpura and Lonigan, 2009). The three dimensions of problem behaviors are split into three subscales: inattentive, hyperactive/impulsive, and ODD. The inattentive subscale consists of five items related to inattentive behaviors typically seen in ADHD. The hyperactivity/impulsivity subscale is comprised of five items which are related to hyperactive and impulsive behaviors associated with ADHD. The ODD subscale was not used in the analyses for the current study. A previous study indicated that the validity of the CTRS was large at ($r = .62$; Roberts et al., 1981). The CTRS-15 had good internal consistency in this sample (inattention subscale $\alpha = .90$; hyperactivity/impulsivity subscale $\alpha = .89$).

Procedure

After schools agreed to participate in the study, parental consent forms were sent home with students in all participating classrooms. As part of the larger study, children completed various standardized measures of reading and reading-related skills, which were not used in this study. Assessments were completed throughout the school year as schools, classrooms, and children were recruited. Coincident with the completion of the standardized assessments, the children's general-classroom teachers were asked to complete the SWAN and CTRS-15 on the consented children in their classrooms. Parents of consented children were sent a packet of questionnaires and rating scales that included the SWAN and CTRS-15.

Results

Preliminary Analyses and Descriptive Statistics

Across the full sample, the mean item scores and standard deviations on the SWAN inattention subscale for parent ($M = 0.28$, $SD = 0.78$) and teacher ($M = 0.17$, $SD = 1.27$) data and on the hyperactivity/impulsivity subscale for parent ($M = 0.30$, $SD = 0.81$) and teacher ($M = 0.23$, $SD = 1.20$) data were in the "average" rating range. Mean item scores and standard deviations for the full sample on the CTRS-15 inattention subscale for parent ($M = 0.36$, $SD = 0.61$) and teacher ($M = 0.59$, $SD = 0.77$) data and on the hyperactivity/impulsivity subscale for parent ($M = 0.33$, $SD = 0.58$) and teacher ($M = 0.42$, $SD = 0.68$) data were in the "seldom" to "occasionally" rating range. Item-level descriptive statistics for the inattention subscale and hyperactivity/impulsivity subscale of the SWAN by grade groups are shown in Table S1 and Table S2, respectively. Item-level descriptive statistics for both inattention and hyperactivity/impulsivity subscales of the CTRS-15 are described in Table S3.

A Little's MCAR test was conducted with the SWAN and CTRS-15 data from both parents and teachers. In consideration of grade group, the Little's MCAR test was not statistically significant ($p = .08$), indicating that data were missing at random, unrelated to grade or the level of behavior problems reported. All of the analyses in this study were conducted using the following subgroups of the data which, with the exception of the preschool group, contained data from two grades, which allowed for each group to be adequately powered: Preschool, Kindergarten and 1st Grade (K/1), 2nd and 3rd Grade (2/3), and 4th and 5th Grade (4/5). For the whole sample, 222 participants had only parent report, 717 had just teacher report, and 706 had both parent and teacher report. Table S4 shows the number of children in each grade group for whom there was parent report, teacher report, or both parent and teacher report.

Evaluation of Measurement Invariance

Testing of measurement invariance followed the recommended procedures outlined by Putnick and Bornstein (2016). Analyses of measurement invariance were conducted using Mplus 7.4 (Muthén and Muthén, 2014). Indicators for the SWAN were treated as continuous because each indicator has seven response options (Lubke and Muthén, 2004). Indicators for the CTRS-15 were treated as categorical because each indicator has four response options. Consequently, robust maximum likelihood estimation

(MLR) with full information maximum likelihood estimation was used to account for missing data and non-normality for analyses of the SWAN; in contrast, the weighted least square mean and variance (WLSMV) estimator was used for analyses of the CTRS-15. For the SWAN, model fit was evaluated using the Yuan-Bentler scaled chi-square (Y-Bχ^2), the comparative fit index (CFI), the Tucker Lewis Index (TLI), and the root mean square error of approximation (RMSEA). Non-significant chi-square values indicate that a model provides excellent fit to the data. For the CFI and the TLI, values greater than .90 indicate adequate model fit, and values greater than .95 indicate good model fit. RMSEA values below .05 indicate good model fit and values between .05 and .08 indicate adequate model fit (Hu and Bentler, 1999). Conventional benchmarks for fit indices are unlikely to be appropriate with the WLSMV estimator, with values typically indicating better fit than actually achieved (Xia & Yang, 2019). Therefore, only relative fit for the CTRS-15 was examined, using the DIFFTEST function in Mplus to derive the correct chi-square difference between nested models.

It has been argued that chi-square difference tests may result in an increase in Type II error (i.e., judging a measure to be non-invariant when invariance exists), leading to the suggestion that non-invariance be determined by changes in approximate fit indices (i.e., ΔCFI > .005, ΔRMSEA > .01, change in standardized root mean square residual [SRMR] > .025 for factor loadings or .005 for intercepts and residuals; Chen, 2007). Sass et al. (2014) reported that chi-square difference tests with ordered categorical data and maximum likelihood or MLR estimation functioned within expected Type 1 error rates. Changes in approximate fit indices functioned similarly to chi-square difference tests in correctly specified models and less well in misspecified models. Approximate fit indices were less sensitive when fewer indicators were non-invariant than when more indicators were non-invariant. Interestingly, the results of Sass et al. did not support the idea that chi-square difference tests were overpowered. Sass et al. results indicated that approximate fit indices performed poorly in all circumstances when using the WLSMV estimator. Sass et al. also reported that the criteria recommended by Chen (2007) did not perform well (i.e., accepting as invariant models that were non-invariant), and they suggested using criteria recommended by Meade et al. (2008; i.e., ΔCFI > .002). In this study, we focused primarily on chi-square difference tests to identify invariance, both because our samples within each age group were not excessively large and because it allowed the same criteria of invariance to be used for both the SWAN and the CTRS-15; however, where relevant, we note where changes in approximate fit indices yield different results.

Configural Invariance CFA was used to determine if the same model best described parent and teacher report. For the SWAN, a model in which all 18 items defined a single factor was compared to a two-factor model, with an Inattention factor that was comprised of the nine items representing inattentive behaviors and a Hyperactivity/Impulsivity factor that was comprised of the nine items representing the hyperactive/impulsive behaviors. Preliminary analysis of the data suggested adding several correlated residuals to the models that would enhance model fit. Correlated residuals that were consistently identified across parents, teachers, and grade groups were added for each rating scale to improve accuracy of model specification. The one-and two-factor models for the SWAN included correlated residuals for Items 10 and 11. For the CTRS-15, a model in which all 10 items defined a single factor was compared to a two-factor model in which the five inattention items defined an Inattention factor and the five hyperactivity/impulsivity items defined a Hyperactivity/Impulsivity factor. The one- and two-factor models for the CTRS-15 included correlated residuals for Items 8 and 9. Fit indices for each model for each grade group are shown in Table 1. Results indicated that, for both the SWAN and the CTRS-15, the two-factor model demonstrated significantly better fit than did the one-factor models for both parent and teacher report. CFI, TLI, and RMSEA values across all grade groups for both parent and teacher report indicated adequate model fit for the SWAN two-factor model.

Weak/Metric Measurement Invariance Weak (metric) measurement invariance was evaluated by comparing a model in which factor loadings were constrained to be equal across parent and teacher data to a model in which the factor loadings were freely estimated. Results of evaluation of weak measurement invariance for the SWAN are shown in Table 2. Weak measurement invariance was not supported in any of the four grade groups for the SWAN. Modification indices in Mplus were examined to identify sequentially the constraint resulting in the largest model misspecification; this constraint was released, and the model was again compared to the unconstrained model, using the Benjamini-Hochberg correction to account for inflated Type I error with multiple non-independent comparisons (Benjamini and Hochberg, 1995). This process was continued until invariance was achieved or no additional relevant constraints were identified by the modification indices. For the preschool group, releasing five constraints (i.e., items 8, 9, 13, 6, 2) resulted in a model that fit as well as the fully unconstrained model ($p < .06$). For the 2/3 grade group, releasing three constraints (i.e., items 8, 2, 12) resulted in a model that fit as well as the fully unconstrained model ($p < .14$). For both the K/1 and 4/5 grade groups, releasing items identified by the modification indices did not result in models that fit as well as the fully unconstrained models ($ps < .001$). Consequently, partial weak measurement invariance was supported for the preschool and 2/3 grade

Table 1 Parent and Teacher Configural Fit for SWAN and CTRS One- and Two-Factor Models Across All Grade Groups

Measure Respondent Grade Group	One-Factor Model					Two-Factor Model					
	Y-Bχ^2	*df*	CFI	TLI	RMSEA [90% CI]	Y-Bχ^2	*df*	CFI	TLI	RMSEA [90% CI]	$\Delta\chi^2$
SWAN											
Parent											
Pre-K	356.6*	134	.86	.84	.11 [.09–.12]	239.84*	133	.93	.92	.08[.06–.09]	156.52*
K & 1st	478.42*	134	.84	.82	.10 [.09–.11]	298.79*	133	.92	.91	.07[.06–.08]	307.15*
2nd & 3rd	608.64*	134	.84	.82	.12 [.11–.13]	289.37*	133	.95	.94	.07[.06–.08]	462.64*
4th & 5th	669.52*	134	.73	.70	.13 [.12–.14]	381.09*	133	.88	.86	.09[.08–.10]	432.54*
Teacher											
Pre-K	1018.99*	134	.80	.77	.15 [.14–.16]	434.06*	133	0.93	0.92	09[.08–.10]	821.92*
K & 1st	1649.14*	134	.73	.69	.17 [.16–.17]	500.25*	133	.94	.93	.08[.08–.09]	1774.29*
2nd & 3rd	1851.64*	134	.72	.68	.18 [.18–.19]	513.07*	133	.94	.93	.09[.08–.09]	1994.47*
4th & 5th	1354.70*	134	.75	.71	.17 [.16–.18]	453.58*	133	.93	.92	.09[.08–.10]	1464.47*
CTRS-15											
Parent											
Pre-K	90.23*	33	.99	.98	.10 [.08–.13]	72.47*	32	.99	.98	.09[.06–.011]	12.59*
K & 1st	234.87*	33	.96	.94	.15 [.13–.17]	116.17*	32	.98	.98	.10[.08–.12]	46.38*
2nd & 3rd	202.00*	33	.97	.96	.14 [.12–.16]	76.26*	32	.99	.99	.07[.05–.10]	38.49*
4th & 5th	143.10*	33	.96	.94	.12 [.10–.14]	91.38*	32	.98	.97	.09[.07–.011]	24.92*
Teacher											
Pre-K	228.73*	33	.98	.97	.14 [.12–.16]	162.42*	32	.99	.98	.12[.10–.013]	28.40*
K & 1st	438.94*	33	.98	.97	.17 [.16–.19]	258.76*	32	.99	.98	.13[.12–.15]	68.11*
2nd & 3rd	496.38*	33	.96	.95	.19 [.18–.21]	246.67*	32	.98	.98	.13[.12–.15]	79.20*
4th & 5th	395.37*	33	.96	.94	.19 [.17–.20]	141.79*	32	.99	.98	.10[.09–.12]	67.71*

Notes. Y-Bχ^2 = Yuan-Bentler χ^2 ; CFI = comparative fit index; TLI = Tucker-Lewis fit index; RMSEA = root mean square-error of approximation; CI = Confidence intervals; $\Delta\chi^2$ corrected for robust maximization for analyses with the SWAN and derived from the DIFFTEST function in Mplus for analyses with the CTRS-15; * $p < .001$

Table 2 Invariance Testing Across Grade Groups for the SWAN

Grade Group	Y-Bχ^2	Model *df*	CFI	TLI	RMSEA [90% CI]	$\Delta\chi^2$	*df*
Preschool							
Baseline	683.13***	266	.932	.922	.083 [.08–.09]	–	–
Weak Invariance	733.71***	284	.927	.921	.084 [.08–.09]	51.80***	18
Partial Weak Invariance	706.84***	279	.930	.924	.082 [.08–.09]	22.00	13
Strong Invariance	746.95***	297	.927	.924	.082 [.08–.09]	37.16***	18
Partial Strong Invariance	734.78***	293	.928	.925	.082 [.08–.09]	24.75***	14
Kindergarten and 1st Grade							
Baseline	793.24***	266	.932	.922	.077 [.07–.08]	–	–
Weak Invariance	861.17***	284	.926	.920	.077 [.07–.08]	73.64***	18
Partial Weak Invariance	835.02***	282	.929	.923	.076 [.07–.08]	38.96***	16
2nd and 3rd Grade							
Baseline	807.32***	266	.941	.932	.080 [.07–.09]	–	–
Weak Invariance	864.11***	284	.937	.932	.080 [.09–.07]	57.76***	18
Partial Weak Invariance	836.17***	280	.939	.934	.079 [.07–.09]	19.81	14
Strong Invariance	921.36***	298	.932	.930	.081 [.07–.09]	100.91***	18
Partial Strong Invariance	861.75***	292	.940	.940	.080 [.07–.09]	20.51	12
Strict Invariance	919.39***	310	.930	.940	.080 [.07–.08]	56.50***	18
Partial Strict Invariance	878.40***	308	.940	.940	.080 [.07–.08]	25.07	16
4th and 5th Grade							
Baseline	837.55***	266	.919	.907	.080 [.08–.10]		
Weak Invariance	905.10***	284	.912	.905	.080 [.08–.10]	72.29***	18
Partial Weak Invariance	883.85***	283	.915	.908	.080 [.08–.10]	42.88***	17

Notes. Y-Bχ^2 = Yuan-Bentler χ^2; CFI = comparative fit index; TLI = Tucker-Lewis fit index; RMSEA = root mean square-error of approximation; CI = Confidence intervals; $^{+}p < .10$; $^{**}p < .05$; $\Delta\chi^2$ corrected for robust maximization; $^{***}p < .001$

groups but not supported for the K/1 and 4/5 grade groups for the SWAN.

Results of evaluation of weak measurement invariance for the CTRS-15 are shown in Table 3 for all grade groups. Weak measurement invariance was not supported in any of the four grade groups. Again, modification indices were used to identify items resulting in model misspecification; constraints for these items were released and resulting models compared to the unconstrained model. For the preschool group, releasing four constraints (i.e., items 10, 6, 1, 2) resulted in a model that fit as well as the fully unconstrained model ($p < .14$). For the K/1 grade group, releasing four constraints (i.e., items 1, 6, 2, 4) resulted in a model that fit as well as the fully unconstrained model ($p < .15$). For the 2/3 grade group, releasing two constraints (i.e., items 2, 1) resulted in a model that fit as well as the fully unconstrained model ($p < .16$). For the 4/5 grade group, releasing three constraints (i.e., items 2, 6, 5) resulted in a model that fit as well as the fully unconstrained model ($p < .18$). Consequently, partial weak measurement invariance was supported for all four grade groups.

Strong/Scalar Measurement Invariance When partial weak measurement invariance was supported, strong (scalar) measurement invariance was evaluated by comparing a model with factor loadings and intercepts (SWAN) or thresholds (CTRS-15) constrained to be equal across parents and teachers to a model for which the intercepts or thresholds were freely estimated. Results for the SWAN for the preschool and 2/3 grade groups are shown in Table 2 (strong invariance on the SWAN for the K/1 and 4/5 grade groups was not evaluated because partial weak invariance was not obtained). Strong invariance was not supported for either the preschool or 2/3 grade groups. For the preschool group, modification indices identified four constraints (i.e., items 11, 4, 18, 17); however, the model with these four constraints released fit significantly worse than the model with all intercepts freely estimated ($p < .04$). For the 2/3 grade group, modification indices identified six constraints (i.e., items 2, 11, 17, 1, 6, 4), and the model with these six constraints released resulted in a model that fit as well as the model with intercepts freely estimated ($p < .06$). Consequently, partial strong measurement invariance was supported for the 2/3 grade group but not for the preschool group. As seen in Table 2, the difference in CFI between the partial strong model and the

Table 3 Model Results for Invariance Testing Across Preschool and Kindergarten and 1st Grade Groups for the CTRS-15

Grade Group	Y-Bχ^2	Model *df*	CFI	TLI	RMSEA [90% CI]	$\Delta\chi^2$	*df*
Preschool							
Baseline	235.19***	64	.986	.980	.11 [.09–.12]	–	–
Weak Invariance	270.39***	74	.990	.990	.11 [.09–.12]	60.69***	10
Partial Weak Invariance	185.25***	70	.990	.990	.08 [.07–.10]	9.78	6
Strong Invariance	228.90***	100	.989	.990	.07 [.06–.09]	69.27***	30
Partial Strong Invariance	186.92***	98	.992	.992	.06 [.05–.08]	37.95+	18
Kindergarten and 1st Grade							
Baseline	383.76***	64	.987	.982	.12 [.11–.13]	–	–
Weak Invariance	452.76***	74	.990	.990	.12 [.11–.13]	100.78***	10
Partial Weak Invariance	317.46***	70	.990	.988	.10 [.09–.11]	9.45	6
Strong Invariance	309.42***	100	.991	.991	.08 [.07–.09]	69.06***	30
Partial Strong Invariance	284.84***	91	.992	.991	.08 [.07–.09]	32.10+	21
2nd and 3rd Grade							
Baseline	340.19***	64	.984	.977	.12 [.10–.13]	–	–
Weak Invariance	264.96***	74	.989	.987	.09 [.08–.10]	27.42***	10
Partial Weak Invariance	249.21***	72	.990	.991	.09 [.08–.10]	11.85	8
Strong Invariance	236.72***	102	.992	.993	.06 [.05–.08]	43.30+	30
4th and 5th Grade							
Baseline	233.93***	64	.986	.980	.10 [.09–.11]	–	–
Weak Invariance	307.76***	74	.989	.987	.11 [.10–.12]	72.42***	10
Partial Weak Invariance	205.33***	70	.991	.992	.09 [.07–.10]	8.88	6
Strong Invariance	217.38***	100	.991	.991	.07 [.05–.08]	51.70***	30
Partial Strong Invariance	201.35***	97	.991	.991	.06 [.05–.08]	36.82+	27

Notes. Y-Bχ^2 = Yuan-Bentler χ^2; CFI = comparative fit index; TLI = Tucker-Lewis fit index; RMSEA = root mean square-error of approximation; CI = Confidence Intervals; $\Delta\chi^2$ derived from the DIFFTEST function in Mplus; ** $p < .05$; *** $p < .001$

partial weak model for the preschool group was .002, indicating that partial strong invariance was just supported based on approximate fit indices.

Results for the CTRS-15 are shown in Table 3. Strong measurement invariance was supported for the 2/3 group ($p < .06$) but not for the other three grade groups ($ps < .008$). For the preschool group, modification indices identified one constraint (i.e., item 2), and the model with this constraint released fit as well as the model with unconstrained thresholds ($p = .10$). For the K/1 grade group, modification indices identified three constraints (i.e., items 2, 1, 10), and the model with these constraints released fit as well as the model with unconstrained thresholds ($p < .06$). For the 4/5 grade group, modification indices identified one constraint (i.e., item 2), and the model with this constraint released fit as well as the model with unconstrained thresholds ($p = .10$). Consequently, partial strong measurement invariance was supported for the preschool, K/1, and 4/5 grade groups.

Strict/Residual Measurement Invariance Partial strong invariance for the SWAN was supported only for the 2/3 grade group. For this grade group, strict (residual) measurement invariance was examined by comparing a model with factor loadings, intercepts, and residuals constrained to be equal across parents and teachers to a model for which the residuals were freely estimated. As seen in Table 2, strict measurement invariance was not supported. Modification indices identified two constraints (i.e., items 12, 2), and the model with these constraints released fit as well as the model with unconstrained residuals ($p < .07$). Consequently, partial strict measurement invariance was achieved for the SWAN in the 2/3 grade group. As seen in Table 2, the difference in CFI between the strict model and the partial strong model for the preschool group was .001, indicating that strict invariance was supported based on approximate fit indices.

Discussion

The purpose of this study was to determine whether the items on two ratings scales used to assess ADHD-related behaviors in elementary-school-age children (i.e. the SWAN and the CTRS-15) had the same meaning, and therefore measured the same constructs, across parent and teacher ratings. It was hypothesized that the SWAN and CTRS-15 would both demonstrate measurement invariance across informants. Although both the SWAN and the CTRS-15 yielded the expected two-factor structure (i.e., Inattention and Hyperactivity/

Impulsivity factors) for both parents' and teachers' ratings, partial weak measurement invariance (i.e., metric invariance) was only supported in all grade groups for the CTRS-15. Partial strong measurement invariance (i.e., scalar invariance) also was supported for the CTRS-15 across all grade groups. For the SWAN, although partial weak measurement invariance was supported for the preschool grade group and partial strict measurement invariance was supported for the 2/3 grade group, invariance was not supported at all in the K/1 and 4/5 grade groups. It is unclear as to why measurement invariance was supported in the preschool and 2/3 grade groups and not for the K/1 and 4/5 grade groups. These results, however, may explain, in part, the common finding of substantial discrepancies in parent and teacher reports of ADHD symptoms. That is, depending on the measure used to obtain parent and teacher ratings and the age group of the children being rated, the items may not have the same meaning across informants.

Based on the Operations Triad Model (De Los Reyes et al., 2013), if measurement invariance analyses indicate that items on a rating scale may not have the same meaning across parents and teachers this may be the result of measurement conditions that result in compensating operations. Given the lack of invariance that was demonstrated between parents and teachers on the SWAN, any rating discrepancy may be the result of compensating operations, which are measurement conditions in which discrepant ratings by informants are due to methodological errors rather than different observable behaviors across contexts. Although children sometimes behave differently in different settings (Timmons et al., 2016), measurement invariance analyses are important to be able to understand if it is the observable behaviors of the child, the interpretation of the measures by the informants, or rater biases due to factors such as demographics or parental stress that are related to the rating discrepancies. De Los Reyes et al. claim that different observable behaviors would result in diverging operations, which are measurement conditions that result in discrepant ratings because the child is behaving differently across contexts. Because the results supported measurement invariance on the CTRS-15, rating discrepancies with that rating scale could be assumed to be the result of diverging operations, which would indicate that the children may have exhibited different observable behaviors across contexts. The findings of this study have implications for common methods of measuring ADHD-related behavior and the development of rating scales used to assess such problem behaviors.

Parent Versus Teacher Report: Same or Different Constructs?

The lack of weak measurement invariance for all grade groups for the SWAN indicated that what parents and teachers were rating when using the measure may have been representative of different meanings of the items and therefore potentially different underlying constructs. That is, the inattention and hyperactivity/impulsivity rated by parents differed in meaningful ways from the inattention and hyperactivity/impulsivity rated by teachers in the K/1 and 4/5 grade groups. In contrast, support for at least partial weak measurement invariance for the CTRS-15 indicated that what parents and teachers were rating when using the measure represented largely the same underlying constructs across all grade groups. Moreover, support for partial strong invariance for the CTRS-15 in all grade groups indicated that differences in levels of the underlying constructs affected many items in largely the same way. The significance of these findings is that comparisons between parent and teacher report on the SWAN will not always yield a meaningful outcome because the items are interpreted differently; however, comparisons between parent and teacher report on the CTRS-15 have the potential to yield meaningful outcomes because the constructs being measured are the same and, therefore, resultant scores on the items reflect the same underlying level of the constructs across parent and teacher data.

For both the SWAN and the CTRS-15, it is important to note the number of non-invariant items for which constraints were released leading to partial measurement invariance. There were approximately 40% of non-invariant items across grade groups and rating scales, which could have significant implications. Comparisons across raters may be meaningless if a significant portion of the parameters demonstrate non-invariance because this could result in the meaning of the underlying constructs to differ across informants. Previous research has highlighted some of the problems with partial invariance, although also supporting the use and utility of it. For example, research has demonstrated that comparisons between groups can potentially be greatly affected by non-invariant intercepts (i.e., strong/scalar invariance) but that only marginal effects were seen with non-invariant factor loadings (i.e., weak/metric invariance; Shi et al., 2019). Shi et al., also stressed the importance of using partial invariance, if done correctly. They suggested that if the non-invariant parameters are freely estimated, as they were in the current study, the models tend to result in accurate and consistent estimates. Shi et al. also indicated that ignoring partial invariance, failing to detect non-invariant parameters, and not allowing them to be freely estimated can lead to a greater risk of Type II errors. In conclusion, it is important to recognize both the issues with and importance of partial measurement invariance in such analyses.

Both rating scales had items that could be considered problematic because factor loadings differed significantly between parents and teachers across all grade groups. Although multiple items on the SWAN were identified by the modification indices to have constraints removed, Item 8, "Ignores extraneous stimuli," was the only item to be identified across all grade groups. A potential reason that Item 8 was problematic

could be the result of the vocabulary (i.e. "extraneous"), leading to the intended meaning not being as accessible to some raters as it was for others. There were similarly problematic items on the CTRS-15. Items 1 and 2 were identified as problematic by the modification indices across most of the grade groups. Item 1 on the CTRS-15, "Restless in the 'squirmy' sense," may have been problematic because some raters may interpret the word "squirmy" differently than others. Item 2 on the CTRS-15, "Fails to finish things s/he starts," may have been problematic because there are considerably different responsibilities at school and at home. It may be more likely for children to fail to finish tasks observed by teachers (i.e., multi-step assignments) compared to tasks observed by parents (i.e., single-step chores).

Differences between the SWAN and CTRS-15

It is possible that differences in the design and structure of the SWAN and CTRS-15 account for the differences in measurement invariance of the two rating scales across parent and teacher data. There are at least four substantial differences in how the SWAN and the CTRS-15 were designed and developed to assess ADHD symptoms. First, the SWAN was constructed to mirror the DSM-IV symptoms of ADHD (and consequently maps onto the DSM-5 symptoms of ADHD), whereas items on the CTRS-15 describe behaviors that are consistent with DSM ADHD symptoms but are not the symptoms themselves. For example, the corresponding item on the SWAN for the DSM-5 symptom, "Often has difficulty sustaining attention in tasks or play activities" (American Psychiatric Association, 2013, p. 59) is "Sustains attention on tasks or play activities," and the most similar item on the CTRS-15 is "Short attention span." In addition to the differences in item-to-symptom mapping, the SWAN includes all 18 DSM ADHD symptoms, whereas the CTRS-15 includes only 10 items related to inattention and hyperactivity/impulsivity. The mirroring of DSM symptoms of ADHD on the SWAN but not the CTRS-15 suggests that the lack of invariance across all grade groups may not be the result of the rating scale, but of the actual DSM symptoms for ADHD. This is consistent with the findings of previous literature when measurement invariance was not supported in studies that examined DSM symptoms of ADHD (DuPaul et al. 2020; Vitoratou et al., 2019). Willcutt et al. (2012) reported that in consideration of the strength of the relations between symptoms and underlying constructs and a variety of other factors that consideration should be given to eliminating, replacing, or rewording some DSM symptoms of ADHD. These findings may support those claims.

Second, as exemplified by the sample items above, the SWAN is worded in an asymptomatic direction. The items on the SWAN are phrased such that informants are asked to rate the opposite of the problem behavior, and symptomatic behavior is captured by a rating scale that indicates that the behavior is exhibited less often than other children of the same age. In contrast, the items on the CTRS-15 are phrased in a symptomatic direction, and asymptomatic behavior is not measured beyond a rating of "never/seldom." Consequently, whereas the CTRS-15 directs raters' attention to the problem behaviors associated with ADHD, the SWAN directs raters' attention to positive, asymptomatic behaviors. The valance of the items could have impacted these results because the items in the SWAN require informants to report how often a symptomatic behavior is not happening rather than simply reporting the frequency with which a symptomatic behavior occurs, if at all. Consequently, the need to report how often the asymptomatic behavior is occurring has the potential to prevent the informant from accurately reporting how often the symptomatic behavior is occurring. This might result in a discrepancy between parent and teacher report. With a large number of students and likely a better understanding of developmentally appropriate behaviors than parents, teachers may be more likely to notice symptomatic behaviors which may result in an under-reporting of asymptomatic behaviors. Parents, on the other hand, may have a better understanding of the breadth of their child's normal and problem behaviors, allowing them to be better able to report asymptomatic and symptomatic behaviors.

A third difference between the SWAN and the CTRS-15 concerns the number and nature of the rating-scale anchors used. SWAN items require the rater to compare the child's behavior to the average behavior of a child of similar age using a 7-point rating scale that ranges from "far below average" to "far above average." CTRS-15 items require the rater to report how often a particular behavior occurs using a 4-point rating scale that ranges from "Never, Seldom," to "Very Often, Very Frequently." Because teachers have more familiarity with children of a particular age than do parents, they are likely better able to use "average," which has a subjective frame of reference based on experience with children, as an anchor on a rating scale than are parents whose frame of reference is much smaller, perhaps even limited to their child's behavior, the behavior of their child's siblings, or a small group of their child's friends. (Antrop et al., 2002). Consequently, it may be more difficult for parents than for teachers to report differences in behavior consistent with a common underlying construct. Although anchors like "often," "quite a bit," or "occasionally," as used on the CTRS-15, also have a subjective frame of reference, the frame of reference (i.e., frequency of occurrence) includes more things than children's behaviors.

Finally, the methods used to construct the SWAN and the CTRS-15 were considerably different. The SWAN mirrors the 18 symptoms of ADHD described by both the DSM-IV and the DSM-5, albeit worded in the asymptomatic direction. In contrast, the CTRS-15 was empirically constructed by

Purpura and Lonigan (2009) using item-response-theory analyses to optimize information concerning the two dimensions of ADHD symptoms based on items from a hybrid version of multiple forms of the CTRS. That is, items included on the CTRS-15 were selected from a large pool of CTRS items based on high discrimination (i.e., strongly related to the underlying construct) and the degree to which the items provided information across the continuum of difficulty/impairment associated with underlying behavior problems related to inattention or hyperactivity/impulsivity (i.e., items provided the least redundant information representing higher, average, and lower levels of inattention or hyperactivity/impulsivity in the developmental sample).

Whereas factor analytic studies consistently support the distinction between inattentive and hyperactive/impulsive dimensions for symptoms of ADHD and related behaviors as well as strong links between these ADHD-related behaviors and the underlying constructs they represent (Willcutt et al., 2012), few studies have examined the degree to which symptoms and behaviors provide non-redundant information across the range of inattention and hyperactivity/impulsivity. From a measurement perspective, multiple items that provide information about the same level of impairment may be reliable but provide little discrimination of individuals across the full range of the underlying construct. Consequently, measures constructed as the CTRS-15 was may better anchor the underlying construct across a broader range of impairment, leading to more consistent association between the items and the underlying constructs.

Implications and Future Directions

One major implication of these findings is that that the lack of partial weak measurement invariance for the K/1 and 4/5 grade groups for the SWAN results in a lack of clarity concerning observed discrepancies in ratings between parents and teachers. Because the underlying constructs being assessed by the SWAN differ between parent and teacher data in the K/1 and 4/5 grade groups, discrepancies between parent and teacher ratings may not reflect differences concerning the same things. A second implication concerns the connection between the items on these rating scales and the DSM-5 diagnostic criteria for ADHD. The results suggest that parents and teachers were interpreting the SWAN items in different ways for half of the grade groups. Consequently, the relation between the SWAN items and the DSM criteria could indicate that the lack of measurement invariance across all grade groups on the SWAN is actually due to an issue with the ADHD symptoms themselves. A final implication relates to consideration of scale development for research purposes. The purpose of such scales is to be accurate in terms of measuring the behaviors they are intended to assess. The failure to do so could lead to results from future empirical studies being biased due to the informant.

Because of the importance of measurement invariance to multiple aspects of research and clinical work, including the DSM requirement of the ADHD symptom threshold being met in two or more contexts, there are several important future research directions that should be pursued. First, it is important for future studies to replicate the results of this study using different samples, different age groups, and different measures. As noted above, lack of measurement invariance has substantial implications for accurate interpretation of a host of research findings, including how aspects of ADHD relate to other important developmental outcomes and potential causal influences on ADHD-related behaviors. Therefore, identification of the specific measures and contexts in which measurement invariance is obtained represents an important goal. Moreover, because it is frequently used in clinical and research contexts, there is a strong imperative to determine if the finding of this study concerning the lack of measurement invariance for the SWAN is a consistent finding. Because of the frequent use of the SWAN, however, such replications should be easy to produce using extant data.

More broadly, identification of the characteristics of rating scales that are mostly likely to result in measurement invariance would provide high utility for development of future measures as well as methods of obtaining information about symptoms used to make diagnostic decisions (Willcutt et al., 2012). It is possible that the way information about symptoms is collected could also increase the utility of symptoms of ADHD. Manipulation of rating scale items could be used to identify optimal ways to characterize symptoms so that ratings collected from different informants (i.e., teachers, parents, self) reflect the same intended underlying constructs. For instance, comparing current wording of the SWAN to a version in which all items were worded in the symptomatic direction could identify whether it is the valance, or the actual symptoms described that result in a lack of measurement invariance. In addition, measurement invariance could be evaluated using the current version of the SWAN and a version in which the anchor for the items was frequency rather than average.

Limitations

Although this study had a number of strengths, including a relatively large sample across multiple grades, there were several limitations. One possible limitation is that the sample was not sufficiently large to evaluate measurement invariance in individual grades, other than in the Preschool group. However, the fact that the pattern of results was similar across grade groups likely indicates that the impact of the sample as a limitation was not large. A second limitation is that this study was conducted using a community sample. Although some children had reported symptoms in or near the clinical range,

clinical samples generally report more severity in the level of reported symptoms. Although it is unknown how the current results may have been changed in a sample with a greater number of cases with behavior in the clinical range for ADHD, previous research has reported that there is often higher levels of agreement between parent and teacher ratings when rating more severe behaviors (Antrop et al. 2002). If this higher level of agreement between informants also resulted in more measurement invariance, then it could suggest that the non-invariance demonstrated in this study indicates that the SWAN only exhibits a problem in measuring sub-clinical levels of behaviors. Future research is needed to replicate this study with a clinical population. A third limitation was the use of the CTRS-15 rather than a more commonly used version of the CTRS. Because the CTRS-15 is an abbreviated version of the CTRS that was constructed by optimizing information, it is possible that the same lack of weak measurement invariance demonstrated by the SWAN would also be demonstrated by a more commonly used version of the CTRS.

Summary and Conclusions

The results of this study revealed that, contrary to expectations and despite configural invariance, weak or partial weak measurement invariance was not supported for the SWAN for the K/1 and 4/5 grade groups, but partial weak measurement invariance was supported in the preschool and 2/3 grade groups. Partial strong measurement invariance was supported for the CTRS-15 across all of these grade groups. These results indicate that the items on the SWAN have different meanings depending on whether children are rated by parents or teachers for half of the grade groups examined in this study. Therefore, the meaning of direct comparisons of parent and teacher ratings on the SWAN is unknown because different constructs are potentially being compared in half of the grade groups. Additional research is necessary both to replicate the results of this study and to better understand factors influencing parent versus teacher ratings of problem behaviors. Such results will identify the need for and methods of adaptations to commonly used measures that may enhance the accurate assessment and diagnosis of children with ADHD.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10862-021-09874-3.

Funding Portions of this research were supported by grants from the Institute of Educational Science (R305F100027) and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (HD052120). The views expressed herein are those of the authors and have not been reviewed or approved by the granting agencies.

Declarations

Conflict of Interest Colleen M. Jungersen and Christopher J. Lonigan declare that they have no conflict of interest.

Experiment Participants The previous study, during which the data used in this study was collected, was reviewed and approved by the Florida State University Institutional Review Board with HSC Number: 2015.15149. The approval allowed for de-identified data to be used for future studies, such as the current one.

Informed Consent Prior to any data collection, district, school, and classroom agreement to participate in the study was obtained, and informed consent/permission was obtained from the children's parents. Assent was obtained from the children who participated in this study.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10862-021-09874-3.

References

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*(2), 213–232. https://doi.org/10.1037/0033-2909.101.2.213.

Allan, D. M., & Lonigan, C. J. (2019). Examination of the structure and measurement of inattentive, hyperactive, and impulsive behaviors from preschool to grade 4. *Journal of Abnormal Child Psychology, 47*(6), 975–987. https://doi.org/10.1007/s10802-018-0491-x.

Amador-Campos, J. A., Forns-Santacana, M., Guàrdia-Olmos, J., & Peró-Cebollero, M. (2006). DSM-IV attention deficit hyperactivity disorder symptoms: Agreement between informants in prevalence and factor structure at different ages. *Journal of Psychopathology and Behavioral Assessment, 28*(1), 23–32. https://doi.org/10.1007/s10862-006-4538-x.

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

Antrop, I., Roeyers, H., Oosterlaan, J., & Oost, P. V. (2002). Agreement between parent and teacher ratings of disruptive behavior disorders in children with clinically diagnosed ADHD. *Journal of Psychopathology and Behavioral Assessment, 24*(1), 67–73. https://doi.org/10.1023/A:1014057325752.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Biederman, J., Mick, E., & Faraone, S. V. (2000). Age-dependent decline of symptoms of attention deficit hyperactivity disorder: Impact of remission definition and symptom type. *American Journal of Psychiatry, 157*(5), 816–818. https://doi.org/10.1176/appi.ajp.157.5.816.

Burns, G. L., Servera, M., Bernard, M., Carrillo, J. M., & Geiser, C. (2014). Ratings of ADHD symptoms and academic impairment by mothers, fathers, teachers and aides: Construct validity within and across settings as well as occasions. *Psychological Assessment, 26*(4), 1247–1258. https://doi.org/10.1037/pas0000008.

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*(4), 452–457. https://doi.org/10.1037/1040-3590.18.4.452.

Burns, G. L., Walsh, J. A., Servera, M., Lorenzo-Seva, U., Cardo, E., & Rodriguez-Fornells, A. (2013). Construct validity of ADHD/ODD rating scales: Recommendations for the evaluation of forthcoming *DSM–V* ADHD/ODD scales. *Journal of Abnormal Child Psychology, 41*(*1*), 15–26. https://doi.org/10.1007/s10802-012-9660-5.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834.

Chi, T. C., & Hinshaw, S. P. (2002). Mother–child relationships of children with ADHD: The role of maternal depressive symptoms and depression-related distortions. *Journal of Abnormal Child Psychology, 30*, 387–400.

Conners, C. K. (1969). A teacher rating scale for use in drug studies with children. *American Journal of Psychiatry, 126*(6), 884–888. https://doi.org/10.1176/ajp.126.6.884.

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin, 141*(4), 858–900. https://doi.org/10.1037/a0038498.

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 131*(4), 483–509. https://doi.org/10.1037/0033-2909.131.4.483.

De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kundey, S. M. (2013). Principles underlying the use of multiple informants' reports. *Annual Review of Clinical Psychology, 9*, 123–149. https://doi.org/10.1146/annurev-clinpsy-050212-185617.

Duhig, A. M., Renk, K., Epstein, M. K., & Phares, V. (2000). Interparental agreement on internalizing, externalizing, and total behavior problems: A meta-analysis. *Clinical Psychology: Science and Practice, 7*, 435–453.

DuPaul, G. J., Fu, Q., Anastopoulos, A. D., Reid, R., & Power, T. J. (2020). ADHD parent and teacher symptom ratings: Differential item functioning across gender, age, race, and ethnicity. *Journal of Abnormal Child Psychology, 48*, 679–691. https://doi.org/10.1007/s10802-020-00618-7.

Gooch, D., Maydew, H., Sears, C., & Norbury, C. F. (2017). Does a child's language ability affect the correspondence between parent and teacher ratings of ADHD symptoms? *BMC Psychiatry, 17*(1), 129. https://doi.org/10.1186/s12888-017-1300-8.

Harvey, E. A., Fischer, C., Weieneth, J. L., Hurwitz, S. D., & Sayer, A. G. (2013). Predictors of discrepancies between informants' ratings of preschool-aged children's behavior: An examination of ethnicity, child characteristics, and family functioning. *Early Childhood Research Quarterly, 28*(4), 668–682. https://doi.org/10.1016/j.ecresq.2013.05.002.

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

Javo, C., Heyerdahl, S., & Rønning, J. A. (2000). Parent reports of child behavior problems in young Sami children: A cross-cultural comparison. *European Child & Adolescent Psychiatry, 9*(3), 202–211.

Lakes, K. D., Swanson, J. M., & Riggs, M. (2012). The reliability and validity of the English and Spanish strengths and weaknesses of ADHD and normal behavior rating scales in a preschool sample: Continuum measures of hyperactivity and inattention. *Journal of Attention Disorders, 16*(6), 510–516. https://doi.org/10.1177/1087054711413550.

Larsson, H., Dilshad, R., Lichtenstein, P., & Barker, E. D. (2011). Developmental trajectories of DSM-IV symptoms of attention-deficit/hyperactivity disorder: Genetic effects, family risk and associated psychopathology. *Journal of Child Psychology and Psychiatry, 52*(9), 954–963. https://doi.org/10.1111/j.1469-7610.2011.02379.x.

Lee, S., Burns, G. L., Snell, J., & McBurnett, K. (2014). Validity of the sluggish cognitive tempo symptom dimension in children: Sluggish cognitive tempo and ADHD inattention as distinct symptom symptoms. *Journal of Abnormal Child Psychology, 42*, 7–19. https://doi.org/10.1007/s10802-013-9714-3.

Lonigan, C. J., Bloomfield, B. G., Anthony, J. L., Bacon, K. D., Phillips, B. M., & Samwel, C. S. (1999). Relations among emergent literacy skills, behavior problems, and social competence in preschool children from low-and middle-income backgrounds. *Topics in Early Childhood Special Education, 19*(1), 40–53. https://doi.org/10.1177/027112149901900104.

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling, 11*(4), 514–534. https://doi.org/10.1207/s15328007sem1104_2.

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*(6), 694–705. https://doi.org/10.1177/1073191114535242.

Massetti, G. M., Lahey, B. B., Pelham, W. E., Loney, J., Ehrhardt, A., Lee, S. S., & Kipp, H. (2008). Academic achievement over 8 years among children who met modified criteria for attention-deficit/hyperactivity disorder at 4–6 years of age. *Journal of Abnormal Child Psychology, 36*(3), 399–410. https://doi.org/10.1007/s10802-007-9186-4.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568–592. https://doi.org/10.1037/0021-9010.93.3.568.

Murray, D. W., Kollins, S. H., Hardy, K. K., Abikoff, H. B., Swanson, J. M., Cunningham, C., & Chuang, S. Z. (2007). Parent versus teacher ratings of attention-deficit/hyperactivity disorder symptoms in the Preschoolers with Attention-Deficit/Hyperactivity Disorder Treatment Study (PATS). *Journal of Child and Adolescent Psychopharmacology, 17*(5), 605–619. https://doi.org/10.1089/cap.2007.0060.

Muthén, L. K., & Muthén, B. O. (2014). *Mplus 7.4 [software]*. Los Angeles: Muthén and Muthén.

Narad, M. E., Garner, A. A., Peugh, J. L., Tamm, L., Antonini, T. N., Kingery, K. M., Simon, J. O., & Epstein, J. N. (2015). Parent–teacher agreement on ADHD symptoms across development. *Psychological Assessment, 27*(1), 239–248. https://doi.org/10.1037/a0037864.

Purpura, D. J., & Lonigan, C. J. (2009). Conners' teacher rating scale for preschool children: A revised, brief, age-specific measure. *Journal of Clinical Child & Adolescent Psychology, 38*(2), 263–272. https://doi.org/10.1080/15374410802698446.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004.

Rabiner, D., Coie, J. D., & Conduct Problems Prevention Research Group. (2000). Early attention problems and children's reading achievement: A longitudinal investigation. *Journal of the American Academy of Child & Adolescent Psychiatry, 39*(7), 859–867. https://doi.org/10.1097/00004583-200007000-00014.

Richters, J. E. (1992). Depressed mothers as informants about their children: a critical review of the evidence for distortion. *Psychological bulletin, 112*(3), 485.

Richters, J., & Pellegrini, D. (1989). Depressed mothers' judgments about their children: an examination of the depression-distortion hypothesis. *Child Development, 60,* 1068–1075. https://doi.org/10.2307/1130780.

Roberts, M. A., Milich, R., Loney, J., & Caputo, J. (1981). A multitrait-multimethod analysis of variance of teachers' ratings of aggression, hyperactivity, and inattention. *Journal of Abnormal Child Psychology, 9*(3), 371–380. https://doi.org/10.1007/BF00916841.

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(2), 167–180. https://doi.org/10.1080/10705511.2014.882658.

Shi, D., Song, H., & Lewis, M. D. (2019). The impact of partial factorial invariance on cross-group comparisons. *Assessment, 26*(7), 1217–1233. https://doi.org/10.1177/1073191117711020.

Sims, D. M., & Lonigan, C. J. (2013). Inattention, hyperactivity, and emergent literacy: Different facets of inattention relate uniquely to preschoolers' reading-related skills. *Journal of Clinical Child & Adolescent Psychology, 42*(2), 208–219. https://doi.org/10.1080/15374416.2012.738453.

Sonuga-Barke, E. J., Koerting, J., Smith, E., McCann, D. C., & Thompson, M. (2011). Early detection and intervention for attention-deficit/hyperactivity disorder. *Expert Review of Neurotherapeutics, 11*(4), 557–563. https://doi.org/10.1586/ern.11.39.

Stanger, C., & Lewis, M. (1993). Agreement among parents, teachers, and children on internalizing and externalizing behavior problems. *Journal of Clinical Child Psychology, 22*(1), 107–116. https://doi.org/10.1207/s15374424jccp2201_11.

Swanson, J. M., Schuck, S., Porter, M. M., Carlson, C., Hartman, C. A., Sergeant, J. A., Clevenger, W., Wasdell, M., McCleary, R., Lakes, K., & Wigal, T. (2012). Categorical and dimensional definitions and evaluations of symptoms of ADHD: History of the SNAP and the SWAN rating scales. *The International Journal of Educational and Psychological Assessment, 10*(1), 51–70.

Timmons, K., Pelletier, J., & Corter, C. (2016). Understanding children's self-regulation within different classroom contexts. *Early Child Development and Care, 186*(2), 249–267. https://doi.org/10.1080/03004430.2015.1027699.

Touliatos, J., & Lindholm, B. W. (1981). Congruence of parents' and teachers' ratings of children's behavior problems. *Journal of Abnormal Child Psychology, 9*(3), 347–354. https://doi.org/10.1007/BF00916839.

Treutler, C. M., & Epkins, C. C. (2003). Are discrepancies among child, mother, and father reports on children's behavior related to parents' psychological symptoms and aspects of parent–child relationships? *Journal of Abnormal Child Psychology, 31*(1), 13–27. https://doi.org/10.1023/A:1021765114434.

Tripp, G., Schaughency, E. A., & Clarke, B. (2006). Parent and teacher rating scales in the evaluation of attention-deficit hyperactivity disorder: Contribution to diagnosis and differential diagnosis in clinically referred children. *Journal of Developmental & Behavioral Pediatrics, 27*(3), 209–218.

Vitoratou, S., Garcia-Rosales, A., Banaschewski, T., Sonuga-Barke, E., Buitelaar, J., Oades, R. D., Rothenberger, A., Steinhausen, H. C., Taylor, E., Faraone, S. V., & Chen, W. (2019). Is the endorsement of the attention deficit hyperactivity disorder symptom criteria ratings influenced by informant assessment, gender, age, and co-occurring disorders? A measurement invariance study. *International Journal of Methods in Psychiatric Research, 28*(4), e1794. https://doi.org/10.1002/mpr.1794.

Walcott, C. M., Scheemaker, A., & Bielski, K. (2010). A longitudinal investigation of inattention and preliteracy development. *Journal of Attention Disorders, 14*(1), 79–85. https://doi.org/10.1177/1087054709333330.

Willcutt, E. G., Nigg, J. T., Pennington, B. F., Solanto, M. V., Rohde, L. A., Tannock, R., Loo, S. K., Carlson, C. L., McBurnett, K., & Lahey, B. B. (2012). Validity of DSM-IV attention deficit/hyperactivity disorder symptom dimensions and subtypes. *Journal of Abnormal Psychology, 121*(4), 991–1010. https://doi.org/10.1037/a0027347.

Willcutt, E. G., & Pennington, B. F. (2000). Comorbidity of reading disability and attention-deficit/hyperactivity disorder: Differences by gender and subtype. *Journal of Learning Disabilities, 33*(2), 179–191. https://doi.org/10.1177/002221940003300206.

Willcutt, E. G., Pennington, B. F., & DeFries, J. C. (2000). Etiology of inattention and hyperactivity/impulsivity in a community sample of twins with learning difficulties. *Journal of Abnormal Child Psychology, 28*(2), 149–159. https://doi.org/10.1023/A:1005170730653.

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods, 51*(1), 409–428. https://doi.org/10.3758/s13428-018-1055-2.

Youngstrom, E., Loeber, R., & Stouthamer-Loeber, M. (2000). Patterns and correlates of agreement between parent, teacher, and male adolescent ratings of externalizing and internalizing problems. *Journal of Consulting and Clinical Psychology, 68*(6), 1038–1050. https://doi.org/10.1037/0022-006X.68.6.1038.