Advances in Neurodevelopmental Disorders
https://doi.org/10.1007/s41252-025-00452-2

REVIEW

A Scoping Review of Attention-Deficit/Hyperactivity Disorder Assessment and Diagnosis: Tools, Practices, and Sex Bias

Sasha L. Crocker[1] · Anja Roemer[2] · Sarah Strohmaier[3] · Grace Y. Wang[4] · Oleg N. Medvedev[1]

Accepted: 2 June 2025

Abstract

Objectives Accurately diagnosing attention-deficit/hyperactivity disorder (ADHD) is challenging due to the overlap of symptoms with other mental health conditions. This scoping review evaluated the dependability and accuracy of prevalent diagnostic scales and investigates potential obstacles to ADHD assessment diagnosis including potential sex bias.

Method Following the PRISMA-ScR guidelines, 11 widely used diagnostic scales were identified and included. All scales were evaluated based on their psychometric quality and alignment with DSM-5 diagnostic criteria for ADHD.

Results The Attention Deficit Disorders Evaluation Scale emerged as the most reliable among the 11 scales, with the Symptom Checklist-4 ranking as the least reliable. No single assessment tool was adequate for ADHD diagnosis; additional testing was required for accurate conclusions. The literature revealed sex and age biases in some of the assessments. It was discovered that girls were diagnosed with ADHD less often than boys, yet their likelihood of misdiagnosis was notably lower.

Conclusions This review emphasizes the necessity of comprehensive, multi-method assessment approaches for accurate ADHD diagnosis, as no single tool demonstrated sufficient diagnostic precision. Effective clinical assessment design must incorporate strong psychometric measures, address sex-based diagnostic disparities, and emphasize the importance of evaluating behavioural changes over time and their functional impact across settings.

Keywords ADHD · Assessment · Diagnosis · Sex differences · Reliability · Validity

Attention-deficit/hyperactivity disorder (ADHD) is a neurodevelopmental condition involving attention and organizational difficulties, increased impulsivity, and hyperactivity (American Psychiatric Association, 2013). Although ADHD could be diagnosed at any age, this disorder was typically identified in childhood and could negatively affect a person's life into adulthood. Thus, early diagnosis and intervention for children are essential, as it can positively influence the child's life trajectory (McGoey et al., 2002).

In the absence of diagnostically specific biomarkers, current diagnostic criteria primarily focus on behavioural symptoms (Feldman & Reiff, 2014). The extensive list of behavioural symptoms has been provided by the Diagnostic and Statistical Manual of Mental Disorders 5th Edition (DSM-5; American Psychiatric Association, 2013), such as a constant pattern of inattention and/or hyperactivity-impulsivity that disturbs functioning or development; fidgeting, tapping, excessive talking, and struggling to stay seated; and making careless mistakes with work, failing to follow instructions, being easily distracted, and struggling with self-organization (American Psychiatric Association, 2013, p. 59). However, many of these symptoms could be characteristics of typical development in children and adolescents. Thus, only symptoms that are severe, persistent, and out of proportion to expectations for the child's age or developmental level, and without appropriate alternative explanations count for the diagnosis of ADHD (Feldman & Reiff, 2014). Furthermore, there are no differential criteria for girls and boys, which assume a limited sex bias.

✉ Oleg N. Medvedev
oleg.medvedev@waikato.ac.nz

1 School of Psychological and Social Sciences, University of Waikato, Hamilton, New Zealand

2 School of Psychology, Massey University, Palmerston North, New Zealand

3 Department of Psychology, Institute for Health and Sport, Victoria University Melbourne, Melbourne, Australia

4 School of Psychology and Wellbeing, University of Southern Queensland, Ipswich, Australia

Published online: 07 August 2025

At present, there were indications of both underdiagnosis and overdiagnosis of ADHD in children (Hamed et al., 2015; Kazda et al., 2019; Paris et al., 2015). The causes of overdiagnosis may include growing awareness of mental disorders and associated reduction in stigmatization, changes in diagnostic thresholds, poor clinical judgement, and advertising by the pharmaceutical industry, while the causes of underdiagnosis could be related to the attitude, knowledge, and partnerships between schools, teachers, and children and diagnostic complexity caused by other comorbid psychiatric disorders (Quinn & Madhoo, 2014). Furthermore, a significant sex disparity in ADHD misdiagnosis was also found. Bruchmüller et al. (2012) examined sex differences in ADHD misdiagnoses and found that 157 out of 231 boys had been misdiagnosed compared to only 73 out of 231 girls.

Accurate diagnosis is vital, as misdiagnosis can result in inappropriate medication or treatment, leading to long-term mental health and educational outcomes (Nussbaum, 2012). Ford-Jones (2015) argued that ADHD misdiagnoses can negatively impact children's home life and education, potentially leading to fewer employment opportunities and a reduced social life in adulthood. Alternatively, a diagnosis has been shown to be helpful not only for the growing child themselves, but also the people in their lives, such as parents, siblings, teachers, and healthcare professionals to better understand the child's difficulties, and how best to help them (Hamed et al., 2015). Diagnosis in ADHD is generally via assessment tools; however, there has so far not been a comparison of different assessment tools for ADHD to determine whether they are similarly useful or differ significantly for diagnostic purposes. Specifically, there is a lack of consolidated evidence assessing how these tools perform in terms of psychometric robustness, practical application, and potential biases. This scoping review aimed to evaluate the reliability and validity of prevalent ADHD diagnostic scales while investigating potential obstacles to accurate assessment and diagnosis, including sex bias.

Method

This scoping review followed the PRISMA-ScR guidelines (Tricco et al., 2018) to evaluate established ADHD diagnostic tools with proven clinical and research utility. Our aim was to map diagnostic instruments for ADHD in children and adolescents, examining their psychometric properties, limitations, and applicability across diverse populations (Peters et al., 2015).

Search Strategy

A systematic search of PsycINFO, MEDLINE, ERIC, and Web of Science was conducted between January and March 2024, covering literature published from 1980 to 2024. Search terms included combinations of ("ADHD" OR "Attention deficit hyperactivity disorder" OR "ADD") AND ("assessment" OR "diagnosis" OR "evaluation" OR "screening") AND ("tools" OR "scales" OR "measures" OR "rating scales" OR "questionnaires").

Study Selection

Two reviewers (first and last authors) independently screened titles and abstracts against predetermined criteria. Inclusion criteria for scale selection consisted of (1) alignment with DSM-5 diagnostic criteria for ADHD; (2) assessment of ADHD symptoms across attention-deficit and hyperactivity-impulsivity domains; (3) publication in peer-reviewed journals; (4) demonstrated psychometric properties including Cronbach's alpha coefficients ≥ 0.80, and validation studies showing strong convergent validity; (5) documented usage in at least five peer-reviewed studies within the past decade; and (6) evidence of clinical implementation in practice settings. Measures were limited to English-language tools for assessing ADHD in individuals under 18. Emerging or highly specialized tools were excluded to maintain focus on established measures with broad applicability.

Data Extraction and Synthesis

Initial database searches extracted relevant articles describing and evaluating ADHD assessment tools. The selection process involved identifying key diagnostic scales that met all criteria, ensuring our review focused on tools with documented reliability, validity, and clinical utility. Data regarding psychometric properties, diagnostic accuracy, and potential biases were extracted and synthesized for each included measure.

Results

After screening and review according to our inclusion criteria, we identified a set of 11 ADHD diagnostic scales that align with DSM criteria and offer comprehensive data on ADHD symptoms, reflecting a targeted focus on tools with proven usage in both research and practice and omitting measures with limited validation or niche applications. Internal consistency as assessed through Cronbach's alpha as well as test–retest reliability coefficients were extracted to evaluate the scales' reliability. Identified scales alongside their characteristics and psychometric properties are presented in Table 1.

Table 1 Summary of ADHD diagnostic scales: structure, reliability, validity, and scholarly use

Scale	Reference	Items (subscales)	Cronbach's alpha (α)	Test–retest reliability	Google Scholar citations
ADHD-IV	DuPaul et al. (1997) DuPaul et al. (1998) McGoey et al. (2007)	P = 18 (2) T = 18 (2) T = 18 (2)	α = 0.86–0.92 α = 0.88–0.96 α = 0.78–0.90	r = 0.78–86 r = 0.88–0.90	437 486 148 (1071)
VADRS	Wolraich et al. (2003)	P = 55 (8) T = 43 (3)	α = 0.95 α = 0.90	r = 0.80 r = 0.80	651
CRS3	Conners (2008)	Short P = 45 Short T = 41 Short SR = 41 Long P = 99 Long SR = 99	α = 0.77–97	r = 0.71–98	207
ACTeRS	Ullmann et al. (1984)	P = 25 (4) T = 24 (4)	α = 78–0.96 α = 0.92–0.97	Not observed Not observed	195
CRS; CRS-R	Conners (1989) Conners (1997) Goyette et al. (1978)	Short P = 27 (7) Short T = 28 (6) Long P = 80 (7) Long T = 59 (6)	α = 0.72–0.94 α = 0.77–0.95 α = 0.87–0.94 α = 0.90–0.96	r = 0.47–0.85 r = 0.47–0.92 r = 0.47–0.85 r = 0.47–0.92	186
SC-4	Gadow and Sprafkin (1986)	P = 50 (4) T = 50 (4)	α = 0.93–0.95 α = 0.92–0.95	r = 0.75–0.82 r = 0.70–0.89	83
TTEF	Huang (2009)	3	α = 0.86	r = 0.85–0.94	2
BASC-M	Kamphaus and Reynolds (1998)	P = 46 (4) T = 47 (4)	α = 0.57–84 α = 0.77–0.93	r = 0.60–0.90 r = 0.72–0.93	36
TOVA	Leark et al. (2004)	Time based	Not observed	r = 0.70	39
ADDES	McCarney (1994)	P = 46 (2) T = 60 (2)	α = 0.96–0.98 α = 0.98–0.99	r = 0.88–0.91 r = 0.88–0.97	24
SNAP-IV	Swanson (1981)	P = 18 T = 18	α = 0.94 α = 0.97	Not observed Not observed	39

T teacher; *P* parent; *SR* self-report; *Short* short form; *L* long form; *ADHD-IV* ADHD Rating Scale-IV; *BASC-M* BASC Monitor for ADHD; *CRS* Conners' Rating Scale; *CRS3* Conners' Rating Scale 3rd Edition; *CSR-R* Conners' Rating Scale Revised; *SC–4* ADHD Symptom Checklist-4; *ADDES* Attention Deficit Disorders Evaluation Scale; *ACTeRS* ADD-H Comprehensive Teacher's Rating Scale; *TOVA* Test of Variables of Attention; *TTEF* Target Tests of Executive Functioning; *SNAP-IV* Swanson, Nolan and Pelham Rating Scale; *VADRS* Vanderbilt ADHD Rating Scale.

Characteristics of Commonly Used ADHD Rating Scales

Most rating scales require input from both parents and teachers, allowing for behaviour assessment across home and school settings (Narad et al., 2015). For example, the Attention Deficit Disorders Evaluation Scale (ADDES; Adesman, 1991) collects consistent reports from both parents and teachers to evaluate a child's behaviour relative to ADHD diagnostic criteria, demonstrating particularly strong reliability (α = 0.96–0.99, test–retest r = 0.88–0.97). This dual-reporter design is beneficial in identifying ADHD symptoms that manifest across different environments, enhancing diagnostic validity. Each scale aligns with DSM-5 criteria, translating behaviours from the inattentive and hyperactive-impulsive subtypes into test questions, thereby aiding psychologists in assessing ADHD likelihood based on DSM-5 standards. For a diagnosis, a score below the 93rd percentile cut-off typically predicts the inattentive subtype, while scores below the 90th percentile indicate the hyperactive-impulsive subtype. In research contexts, a more stringent 98th percentile cut-off is sometimes applied to ensure specificity.

Among the included scales, most met the acceptable cut-off of 0.70 for test–retest reliability, with notable variations. The ADDES, ADHD-IV, and VADRS demonstrated the strongest psychometric properties, with consistent reliability across settings. Other scales showed more variable results. The Conners' Rating Scale Revised (CRS-R; Conners, 1997) displayed a wide range of temporal stability (0.47 to 0.92), which suggests variability across settings or populations, potentially impacting temporal stability. Some scales, such as the Swanson, Nolan, and Pelham Rating Scale (SNAP-IV; Swanson et al., 2012) and the ADD-H Comprehensive Teacher's Rating Scale (ACTeRS; Ullmann et al., 1984), did not report temporal stability, which limits the ability to confirm their reliability in repeated applications.

Internal consistency, as measured by Cronbach's alpha, was generally satisfactory across scales, although two scales raised concerns. The BASC Monitor for ADHD reported alphas ranging from 0.57 to 0.84, indicating low to moderate internal consistency. The Test of Variables of Attention (TOVA), which is based on participant response times, did not report internal consistency due to its single-item nature, limiting its reliability assessment.

Limitations of the Rating Scales

A critical limitation observed in some scales is low reliability, which may compromise diagnostic accuracy. When test–retest reliability is inconsistent, as seen in the CRS-R and unreported in others like the SNAP-IV, the potential for differing results across repeated tests increases, potentially leading to misdiagnosis. Internal consistency issues, as with the BASC and TOVA, similarly impact confidence in the scale's ability to measure ADHD symptoms consistently.

The ADHD Symptom Checklist (SC-4; Gadow & Sprafkin, 1997), using a 4-point Likert scale, was noted for its limited response range (0 = not at all to 3 = very much). This scale lacks nuanced options for symptom intensity, which may force respondents to choose responses that do not fully represent symptom severity. For example, a symptom experienced moderately may be hard to distinguish between "pretty much" and "very much," potentially leading to over- or under-reporting and limiting measurement precision. Adding additional response options could improve this scale's reliability and validity.

In summary, while several scales display satisfactory reliability and alignment with DSM-5 criteria, issues of reliability and response range limit some scales' diagnostic utility. Improvements, particularly in scales like the SC-4, CRS-R, and BASC Monitor, could enhance diagnostic consistency and accuracy. Future refinement of these scales, with attention to comprehensive validity measures and response options, will improve their applicability in both clinical and research settings.

Summary of Scales Evaluation

Among the 11 scales evaluated, the Attention Deficit Disorders Evaluation Scale (ADDES) emerged as the most reliable diagnostic instrument, with exceptionally strong psychometric properties (internal consistency $\alpha = 0.96$–0.99, test–retest reliability $r = 0.88$–0.97). This scale demonstrated consistent reliability across both parent and teacher versions, making it particularly valuable for comprehensive assessment. In contrast, the ADHD Symptom Checklist-4 (SC-4) ranked as the least reliable among the evaluated scales, primarily due to its limited response range and inability to assess symptom duration, onset age, or functional impairment as required by DSM-5 criteria. Importantly, our analysis revealed that no single assessment tool was adequate for a definitive ADHD diagnosis. Each scale presented specific limitations in scope, reliability across settings, or alignment with comprehensive diagnostic criteria. This finding emphasizes the necessity of employing multiple assessment methods and supplementary testing for accurate diagnostic conclusions.

Sex and Age Bias

The literature also revealed notable sex and age biases in ADHD assessment. Several studies documented that girls were diagnosed with ADHD significantly less often than boys, despite similar symptom presentations. This disparity appears particularly pronounced in classroom settings, where boys' more externalized hyperactive symptoms received greater attention than girls' predominantly inattentive presentations. Interestingly, while girls were underdiagnosed, their likelihood of misdiagnosis was notably lower than boys, with Bruchmüller et al. (2012) finding that 157 out of 231 boys had been misdiagnosed compared to only 73 out of 231 girls. Age-related biases were also evident, with assessment tools often failing to account for developmental differences across childhood and adolescence, potentially contributing to misdiagnosis, particularly in younger children.

Discussion

The section aimed to synthesize the key findings of this scoping review by evaluating the psychometric properties, strengths, and limitations of commonly used ADHD assessment scales. Given the diverse approaches to ADHD diagnosis, this section is structured to first provide an overarching review of assessment reliability and validity, followed by a detailed evaluation of individual scales. The order of discussion was based on the overall reliability of the scales as identified in the "Results" section, with the most robustly supported tools discussed first, followed by those with greater limitations or concerns regarding validity and bias. This ordering facilitates a progressive critique, moving from stronger measures to those requiring caution in interpretation. The discussion then addresses broader issues such as sex bias in ADHD diagnosis before concluding with recommendations for future research and practice.

Highly Reliable and Widely Used Scales

ADHD Rating Scale-IV (ADHD-IV)

The ADHD-IV is a comprehensive questionnaire that assesses children's behaviour over the previous 6 months,

using DSM-5 criteria for both inattentive and hyperactive-impulsive subtypes. It effectively captures behavioural patterns across both school and home settings through parallel parent and teacher versions, allowing for identification of context-specific behaviours. For diagnostic screening, scores above the 93rd percentile suggest inattentive subtype, while scores above the 90th percentile indicate hyperactive-impulsive subtype; research studies often require the 98th percentile (Pappas, 2006). While the scale shows strong reliability in identifying potential ADHD cases, significant limitations include inadequate cultural adaptation and unclear socioeconomic representation in the normative sample. These validity concerns mean the ADHD-IV cannot stand alone for diagnosis but serves as an effective initial screening tool that must be supplemented with comprehensive psychological evaluation (McGoey et al., 2007; Pappas, 2006).

Vanderbilt ADHD Diagnostic Rating Scale (VADRS)

The VADRS (National Institute for Children's Health Quality, 2002) employs separate parent (55 items) and teacher (43 items) versions to assess children aged 6–12 across multiple domains, including ADHD symptoms, academic performance, and relationships. Both scales use a 0–3 Likert scale for symptoms (Never to Very Often) and 1–5 for performance ratings (Excellent to Problematic). The 6-month assessment period helps capture persistent behaviours rather than daily fluctuations. With strong psychometric properties (temporal reliability $r = 0.80$; internal consistency $\alpha = 0.95$ parent, $\alpha = 0.90$ teacher), the VADRS remains valid despite using DSM-IV criteria, as DSM-5 made no significant changes (National Institute for Children's Health Quality, 2002; American Academy of Pediatrics, 2014). However, the scale's complex scoring system requires meeting specific thresholds across different subscales—for example, scoring 2–3 on at least 6 of 9 items for ADHD subtypes or 4 of 8 items for oppositional defiant disorder. This structural complexity may deter referrals for comprehensive assessment, as practitioners must navigate multiple scoring rules for accurate interpretation.

Attention Deficit Disorders Evaluation Scale (ADDES)

The ADDES (McCarney, 1994) offers separate parent (46 items) and teacher (60 items) versions for broad age ranges—teachers assess children 4–19 years while parents evaluate ages 3–19. Using a 0–4 rating scale (from "does not engage" to "several times an hour"), it captures both frequency and duration of behaviours. Raw scores convert to subscale standard scores and percentiles, closely aligning with DSM-5 ADHD criteria for both inattentive and hyperactive-impulsive symptoms (Demaray et al., 2003). While higher scores suggest greater ADHD likelihood, the absence of specific cut-off scores creates diagnostic ambiguity. The computerized scoring system generates treatment recommendations, but clinicians must interpret what constitutes a "high" score without clear thresholds, limiting the scale's practical application in making diagnostic decisions.

Moderately Reliable Scales with Some Limitations

Conners' Rating Scales–Revised (CRS-R)

The CRS-R (Conners, 1997) evaluates problematic behaviours through separate parent and teacher reports for children aged 3–17, offering both long forms for diagnostic assessment and short forms for screening or repeated use. The Teacher version includes six subscales assessing cognitive problems, oppositional behaviour, hyperactivity-impulsivity, inattention, social difficulties, anxiety/shyness, and perfectionism (Purpura & Lonigan, 2009). The Parent version contains fewer items but more subscales, adding psychosomatic symptoms while including home-specific behaviours absent from the teacher form, such as mealtime behaviour and social exclusion. These dual perspectives provide valuable context for psychologists to identify setting-specific behaviours, though final interpretation requires professional evaluation (Zelnik et al., 2012). Despite acceptable internal consistency above 0.70, test–retest reliability varies substantially from 0.47 to 0.92 (Table 1), indicating temporal instability.

For the CRS-R, cut-off scores vary by both age and sex, creating some ambiguity in interpretation. The Parent Rating Scale uses age-based cut-offs: a score of 50 for children aged 3–9 years and 43 for those aged 10–17 years. The Teacher Rating Scale is more complex, with both age-based and sex-based criteria. For age, the cut-offs are 48 for children aged 3–9 years and 38 for those aged 10–17 years. However, the Teacher scale also specifies sex-based cut-offs that differ from these age standards: 38 for males and 47 for females. This dual system presents a challenge, as Deb et al. (2008) note, since it remains unclear whether clinicians should prioritize age or sex criteria when these cut-offs lead to different diagnostic conclusions.

The CRS-R effectively assesses ADHD symptoms through behavioural questions that align well with DSM-5 criteria. Parents and teachers are ideal raters because they observe children over time, capturing patterns like losing personal items or having few friends—behaviours that cannot be evaluated in a single session. While the scale demonstrates good psychometric properties, its limitations include potential rater bias and, critically, confusing cut-off scores that differ not only between parent and teacher versions but also by age and sex. This ambiguity in scoring criteria complicates diagnosis, as clinicians must navigate

conflicting cut-off standards without clear guidance on which to prioritize.

Conners' 3 Rating Scale

The Conners' 3 (Conners, 2008) rating scale updated normative data from previous versions and introduced a self-report measure. While parent and teacher forms assess children aged 6–18, the self-report is limited to ages 8–18 (Conners et al., 2011). This revision aligned with DSM-V-TR criteria, though minimal changes were made from the CRS-R. Responses use a four-point Likert scale and are converted to standardized T-scores, which have a mean of 50 and standard deviation of 10. This standardization allows comparison across age groups and sex. T-scores of 65–69 are considered elevated, while scores ≥ 70 indicate clinically significant symptoms (Morales-Hidalgo et al., 2017). These standardized scores provide clearer interpretation than raw scores, as they show how a child's symptoms compare to age and sex norms. All three versions assess daily functioning and behavioural patterns, with final T-scores determining whether further evaluation is warranted. The scales demonstrate acceptable internal consistency and test–retest reliability (Table 1).

Behaviour Assessment System for Children Monitor for ADHD (BASC-M)

The BASC-M differentiates between four subtypes: attention problems, hyperactivity, internalizing problems, and adaptive skills (Kamphaus & Reynolds, 1998). This test is used to screen children aged 4 to 18 years old (Angello et al., 2003). The items in this scale are based on the behaviours expected to be seen in a child with ADHD. The scale is based on DSM-IV, as this was the current DSM available at the time of the scale development. This scale has been updated to the BASC-3, which is aligned with DSM-5; however, there have been no significant changes made to the behaviours required for an ADHD diagnosis in the DSM-5, making this scale still relevant (American Psychiatric Association, 2013). The cut-off score for the BASC-M is 59.9, and any child that scores above this is suspected to have ADHD, while a score below this cut-off does not suggest ADHD (Ostrander et al., 1998). The BASC-M demonstrates acceptable internal consistency and test–retest reliability, though the wide range of coefficients (Table 1) suggests inconsistent reliability across subscales. Documentation for this 27-year-old measure is scarce, particularly regarding its scoring system, as research has shifted to the current BASC-3 version. This limited availability of information and the test's age significantly constrain its utility in contemporary clinical practice (Reynolds et al., 2011).

ADHD Symptom Checklist-4 (SC-4)

The SC-4 (Gadow & Sprafkin, 1997) was designed to assess ADHD along with Operant Defiant Disorder (ODD). There are 50 items total in this scale, and all items are relevant to the DSM-IV, as this was the current DSM at the time of the scale release, although, as mentioned above, there were no changes to the behaviour criteria between the DSM-IV and the DSM-5. The SC-4 uses two scoring methods, "symptom count" and "symptom severity," across four subscales: ADHD symptoms, ODD symptoms, Peer Conflict Scale, and Stimulant Side Effects Checklist. Symptoms marked as "often" or "very often" are clinically relevant (scored 1), while "never" or "sometimes" score 0. If the symptom count meets or exceeds DSM criteria, diagnosis may be warranted; if not, no diagnosis is made. The cut-off is binary (Yes/No) rather than numeric. However, the SC-4 has limitations: it does not assess symptom duration, age of onset, or functional impairment as required by the DSM-5. While the scale shows excellent internal consistency and acceptable test–retest reliability (Table 1), clinicians must separately evaluate these additional DSM criteria, including whether symptoms significantly impact daily functioning.

Swanson, Nolan, and Pelham (SNAP) Rating Scale

The SNAP rating scale (Swanson et al., 2012) consists of two subscales, inattentive and hyperactive-impulsive, and adheres to the DSM-IV-R, which is aimed at children who are currently in school. The objective of this scale is to aid in identifying children with ADHD by noting their behaviours at school. This scale is presented on a Likert scale ranging from 0 to 3 (0 = Not at all, 1 = Just a little, 2 = Pretty much, 3 = Very much). Information such as age, ethnicity, school year, type of class, and class size are all obtained in this scale. There is no singular cut-off score; the scores depend on age and gender. For example, if boys over 8 years old and girls of all ages marked the answer 2 (Pretty much) eight times or more, then this would highly suggest that the child has ADHD. For boys under the age of 8 years, the cut-off was 2.5, meaning that their answers must consist of multiple 2 and 3 answers (Swanson et al., 2012). Both parents and teachers fill out the SNAP-IV form, and the test considers both perspectives for the final scores. This test has very generalized questions that both the parent and teacher can apply to both settings; however, this test is not based over a period. The evaluation is set in the present moment, so it would be hard for a teacher who does not know the student well to answer the questions accurately (Bussing et al., 2008). The scale has excellent internal consistency; however, test–retest reliabilities have not been reported (Table 1), so it is not clear whether assessments would be consistent over time. Furthermore, the evaluation does not state symptoms from

the DSM-5, so it would be hard to accurately diagnose using this evaluation, but it can be used as a predictor of ADHD based on the child's behaviour.

Scales with Significant Limitations or Bias Concerns

Test of Variables of Attention (TOVA)

The TOVA (Leark et al., 2004) assesses attention and impulse control through a computerized test, but its validation study raises methodological concerns. The study required 31 participants to complete the test four times: twice in one day with a 90-min interval and twice more a week later. This repeated administration design introduces potential confounds, as factors like fatigue, mood fluctuations, or practice effects could influence performance across sessions. For example, a child's changed emotional state between testing weeks might affect scores independently of actual attention abilities. Leark et al. (2004) did not address these limitations, which undermines confidence in the TOVA's reliability for clinical assessment.

ADD-H Comprehensive Teacher's Rating Scale (ACTeRS)

The ACTeRS (Angello et al., 2003) assesses attention disorders in children aged 5 to 12 using 24 items across four subscales: hyperactivity, attention, oppositional behaviour, and social skills. While both parents and teachers complete the scale, only teacher scores determine outcomes, with T-scores > 61 triggering further testing. The scale demonstrates good internal consistency, though parent version reliability varies more than teacher version (Table 1); test–retest reliability remains unreported. Several limitations undermine the ACTeRS' validity: separate but unexplained sex-specific scales introduce potential bias, a 5-year-old receives the same test as a 12-year-old despite vast developmental differences, and teacher ratings alone determine outcomes without parental input, risking subjective bias. The scale's is not openly available, which prevents verification against DSM-5 criteria, further questioning its diagnostic utility (Carlini & Parks, 1993).

The Target Tests of Executive Functioning (TTEF) The TTEF (Huang, 2009), part of the Pediatric Attention Disorders Diagnostic Screener, assesses working memory and executive functioning through three computer-based tasks: target recognition, target sequencing, and target tracking. The first task tests attention to detail and emotional modulation by showing five coloured squares that must be matched after disappearing in 1.5 s, repeated 153 times. Target sequencing evaluates distractibility and organization by requiring children to remember the sequence of coloured circles matched with appearing squares. The final tracking task measures instruction recall and focus by having children replicate shape movements between rows. While the test lacks specific symptom metrics, it captures observable behaviours during administration that can be evaluated against DSM-5 criteria. Strong internal consistency and test–retest reliability make the TTEF a reliable assessment tool for both hyperactive-impulsive and inattentive symptoms (Huang, 2009). However, like other scales reviewed, no single assessment suffices for ADHD diagnosis; multiple measures are necessary due to varying reliability and potential biases across instruments.

Difficulties of Screening for ADHD in the Context of Research Studies

ADHD screening faces several key challenges. First, the DSM-5 criteria encompass symptoms that often overlap with other diagnoses, such as depression (Newson et al., 2021). Second, diagnosis heavily relies on parent and teacher perceptions, which can be inconsistent (Zelnik et al., 2012). Research indicates that objective, performance-based tests like TOVA, while promising, show low specificity—in one study identifying 78.4% false positives in a sample of 179 children (Zelnik et al., 2012). Furthermore, temporal factors significantly impact diagnosis, as indicated by the study conducted by Morrow et al. (2012), which included 938,000 Canadian children and showed that those born later in the academic year were more likely to receive ADHD diagnoses and medication. The study found lower IQ scores (averaging 86) among 366 diagnosed children, though this finding may reflect sampling bias. Environmental and contextual factors also influence symptom presentation, with children potentially modifying behaviour in response to rewards, complicating consistent assessment (Morrow et al., 2012; Whitely, 2015).

Potential Sex Bias in the Assessment of ADHD

Sex-based diagnostic disparities represent a significant concern in ADHD assessment. Research consistently shows higher diagnosis rates in boys, attributed largely to their more visible hyperactive symptoms compared to girls' predominantly inattentive presentations (Berry & Brunet, 2021; Einarsson & Granström, 2002). Bruchmüller et al. (2012) examined this disparity, finding 157 out of 231 boys had been misdiagnosed, compared to only 73 out of 231 girls. Their study of 473 psychotherapists revealed that clinicians tended to diagnose boys more frequently even when presenting identical symptoms.

Clinical bias compounds these issues. Girls often present with inattentive symptoms that may be mistaken for daydreaming or quiet behaviour (Hill, 2021; Steer & Bilbow, 2021). As Ivens (2021) documented, seemingly compliant

behaviours like quietly drawing or appearing to pay attention while not listening can mask ADHD symptoms. Quinn and Madhoo (2014) noted that depressed mothers were more likely to over-report problematic behaviours, potentially contributing to diagnostic inconsistencies.

The impact of non-binary gender identities on ADHD assessment remains understudied, with most research focusing on binary gender differences. Recent literature suggests the need for more inclusive diagnostic approaches that consider diverse gender expressions and their influence on symptom presentation (Clay et al., 2024; Johansson et al., 2022).

Limitations of ADHD Assessment and Suggestions for Improvements

The reviewed assessment tools reveal several significant limitations. Most notably, many scales lack sufficient validity and reliability data, which are essential for accurate diagnosis. All diagnostic tools should demonstrate strong psychometric properties to ensure acceptable diagnostic outcomes. Another key limitation is that research validation often uses only children with existing ADHD diagnoses, excluding undiagnosed children from the statistical analysis. This sampling bias potentially skews results and limits the generalizability to broader populations.

Among the evaluated scales, the ADDES appears most promising for reliable diagnosis due to its comprehensive structure and dual parent-teacher approach. Its response scale, ranging from "multiple times an hour" to "multiple times a month," effectively captures behavioural frequency over time without requiring repeated administration. While the scale collects demographic data (age and sex) that does not influence scoring—raising questions about its necessity—the ADDES demonstrates robust psychometric properties, test–retest reliability ($r = 0.88$–0.97) and internal consistency ($r = 0.96$–0.99), confirming its reliability and validity for ADHD assessment.

To conclude, of the presented scales, there is a definite need for improvement in terms of adhering closer to the DSM-5 diagnostic criteria and making sure to account for changes in behaviour over time. Additionally, none of these tests can be done as the sole diagnostic test for ADHD, making the process cumbersome for the client, their family, and the psychologist. Finally, due to sex bias in diagnosis, there is a need to review each of these tests further in depth to understand how they might specifically be contributing to sex bias in diagnosis.

Limitations of the Current Review

This review has several methodological limitations that should be considered. First, our search strategy, while comprehensive, was limited to English-language measures, potentially missing relevant assessment tools from other languages and cultures. Second, comparing scales developed across different time periods presented challenges, particularly regarding alignment with evolving DSM criteria. While we focused on current DSM-5 relevance, some older scales required careful interpretation of their diagnostic frameworks. Additionally, access to comprehensive psychometric data varied across scales, with some having limited published reliability or validity information. These constraints highlight the need for ongoing validation studies of ADHD assessment tools.

Directions for Future Research

When reviewing the limitations, ideas for further studies were elucidated. Further research should aim to discover why there is such a big bias around boys and girls when it comes to ADHD and what factors contribute to girls being underdiagnosed compared to boys, even when presenting the same symptoms. This is a big factor when diagnosing, as displayed in Bruchmüller et al. (2012) study, which demonstrated that boys were more likely to be misdiagnosed compared to girls. Further studies could also aim to perhaps create a single test battery with both a 95% reliability and validity rate. This would make it easier to diagnose ADHD and remove bias and human error. Finally, there should be more resources dedicated to educating psychologists on sex bias in ADHD diagnosis and how to overcome this, including for non-binary genders. ADHD is a complex disorder to diagnose, and further research on biases and better testing can help to improve the rates of correct and accurate diagnoses.

Conclusion

This review highlights several critical aspects of ADHD assessment. While multiple diagnostic tools exist, their utility varies significantly, with the ADDES emerging as the most reliable among current options. However, no single tool provides comprehensive diagnostic certainty, emphasizing the need for multiple assessment methods. Sex-based diagnostic disparities and screening difficulties remain significant challenges, suggesting the need for more inclusive and objective assessment approaches. Future development of ADHD assessment tools should focus on addressing these limitations while maintaining strong psychometric properties and clinical utility. Additionally, greater attention to cultural sensitivity and gender diversity in diagnostic criteria could improve assessment accuracy across diverse populations.

Acknowledgements The authors acknowledge the founding editor of the *Advances in Neurodevelopmental Disorders*, Professor Nirbay Singh, who inspired the authors to conduct this scoping review.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Declarations

Ethical Approval Not required.

Conflict of Interest None.

References

Adesman, A. R. (1991). The Attention Deficit Disorders Evaluation Scale. *Journal of Developmental and Behavioral Pediatrics, 12*(1), 65–66. https://doi.org/10.1097/00004703-199102000-00012

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. https://doi.org/10.1176/appi.books.9780890425596

Angello, L. M., Volpe, R. J., DiPerna, J. C., Gureasko-Moore, S. P., Gureasko-Moore, D. P., Nebrig, M. R., & Ota, K. (2003). Assessment of attention-deficit/hyperactivity disorder: An evaluation of six published rating scales. *School Psychology Review, 32*(2), 241–262. https://doi.org/10.1080/02796015.2003.12086196

Berry, C., & Brunet, J. (2021). Assessment of ADHD in girls: Unlocking hidden superpowers through understanding ADHD. In J. Steer & A. Bilbow (Eds.), *Understanding ADHD in girls and women* (pp. 35–72). Jessica Kingsley Publishers.

Bruchmüller, K., Margraf, J., & Schneider, S. (2012). Is ADHD diagnosed in accord with diagnostic criteria? Overdiagnosis and influence of client gender on diagnosis. *Journal of Consulting and Clinical Psychology, 80*(1), 128–138. https://doi.org/10.1037/a0026582

Bussing, R., Fernandez, M., Harwood, M., Hou, W., Garvan, C. W., Eyberg, S. M., & Swanson, J. M. (2008). Parent and teacher SNAP-IV ratings of attention deficit hyperactivity disorder symptoms: Psychometric properties and normative ratings from a school district sample. *Assessment, 15*(3), 317–328. https://doi.org/10.1177/1073191107313888

Clay, T., Callen, E. F., Alai, J., Goodman, D. W., Adler, L. A., & Faraone, S. V. (2024). Measuring quality care for adult ADHD patients: How much does gender and gender identity matter? *Journal of Attention Disorders, 28*(3), 364–376. https://doi.org/10.1177/10870547231218449

Conners, C. (1997). *Conners' rating scales--Revised*. Multi-Health Systems.

Conners, C. K. (1989). *Conners' Rating Scales manual.* Multi Health Systems.

Conners, C. K. (2008). *Conners third edition (Conners 3)*. Western Psychological Services.

Conners C. K., Pitkanen J., Rzepa S.R. (2011) Conners 3rd Edition (Conners 3; Conners 2008). In: Kreutzer J.S., DeLuca J., Caplan B. (eds) *Encyclopedia of Clinical Neuropsychology*. Springer. https://doi.org/10.1007/978-0-387-79948-3_1534

Carlini, R. J., & Parks, T. W. (1993). ADD-H comprehensive teacher's rating scale. *Journal of Psychoeducational Assessment, 11*(1), 95–97. https://doi.org/10.1177/073428299301100114

Deb, S., Dhaliwal, A. J., & Roy, M. (2008). The usefulness of Conners' Rating Scales-Revised in screening for attention deficit hyperactivity disorder in children with intellectual disabilities and borderline intelligence. *Journal of Intellectual Disability Research, 52*(11), 950–965. https://doi.org/10.1111/j.1365-2788.2007.01035.x

DuPaul, G. J., Anastopoulos, A. D., Power, T. J., Reid, R., Ikeda, M. J., & McGoey, K. E. (1998). Parent ratings of attention-deficit/hyperactivity disorder symptoms: Factor structure and normative data. *Journal of Psychopathology and Behavioral Assessment, 20*, 83–102.

DuPaul, G. J., Power, T. J., Anastopoulos, A. D., Reid, R., McGoey, K. E., & Ikeda, M. J. (1997). Teacher ratings of attention deficit hyperactivity disorder symptoms: Factor structure and normative data. *Psychological Assessment, 9*(4), 436.

Demaray, M. K., Elting, J., & Schaefer, K. (2003). Assessment of attention-deficit/hyperactivity disorder (ADHD): A comparative evaluation of five, commonly used, published rating scales. *Psychology in the Schools, 40*(4), 341–361. https://doi.org/10.1002/pits.10112

Einarsson, C., & Granström, K. (2002). Gender-biased Interaction in the Classroom: The influence of gender and age in the relationship between teacher and pupil. *Scandinavian Journal of Educational Research, 46*(2), 117–127. https://doi.org/10.1080/00313830220142155

Feldman, H. M., & Reiff, M. I. (2014). Attention deficit–hyperactivity disorder in children and adolescents. *New England Journal of Medicine, 370*(9), 838–846. https://doi.org/10.1056/NEJMcp1307215

Ford-Jones, P. C. (2015). Misdiagnosis of attention deficit hyperactivity disorder: "Normal behaviour" and relative maturity. *Paediatrics & Child Health, 20*(4), 200–202. https://doi.org/10.1093/pch/20.4.200

Gadow, K. D., & Sprafkin, J. (1986). *Stony Brook child psychiatric checklist-3*. State University of New York at Stony Brook.

Gadow, K. D., & Sprafkin, J. (1997). *Child Symptom Inventory-4 norms manual*. Checkmate Plus.

Goyette, C. H., Conners, C. K., & Ulrich, R. F. (1978). Normative data on revised Conners parent and teacher rating scales. *Journal of Abnormal Child Psychology, 6*, 221–236.

Hamed, A. M., Kauer, A. J., & Stevens, H. E. (2015). Why the diagnosis of attention deficit hyperactivity disorder matters. *Frontiers in Psychiatry, 6*, 168.

Hill, P. (2021). Treatment of ADHD in Girls. In J. Steer., & A. Bilbow (Eds.), *Understanding ADHD in girls and women*. (p. 73–106). Jessica Kingsley Publishers.

Huang, L. V. (2009). Test and product review: Pediatric attention disorders diagnostic screener. *Journal of Attention Disorders, 13*(3), 310–314. https://doi.org/10.1177/1087054709346681

Ivens, V. (2021). Coaching girls with ADHD. In J. Steer., & A. Bilbow (Eds.), *Understanding ADHD in girls and women*. (p. 173–210). Jessica Kingsley Publishers.

Johansson, C., Kullgren, C., Bador, K., & Kerekes, N. (2022). Gender non-binary adolescents' somatic and mental health throughout

2020. *Frontiers in Psychology, 13*, Article 993568. https://doi.org/10.3389/fpsyg.2022.993568

Kazda, L., Bell, K., Thomas, R., McGeechan, K., & Barratt, A. (2019). Evidence of potential overdiagnosis and overtreatment of attention deficit hyperactivity disorder (ADHD) in children and adolescents: Protocol for a scoping review. *British Medical Journal Open, 9*(11), Article e032327. https://doi.org/10.1136/bmjopen-2019-032327

Leark, R. A., Wallace, D. R., & Fitzgerald, R. (2004). Test-retest reliability and standard error of measurement for the Test of Variables of Attention (T.O.V.A.) with healthy school-age children. *Assessment, 11*(4), 285–289. https://doi.org/10.1177/1073191104269186

Kamphaus, R., & Reynolds, C. (1998). *BASC monitor for ADHD*. American Guidance Service.

McCarney, S. (1994). *Attention deficit disorders intervention manual (2nd ed.)*. Hawthorne Educational Services.

McGoey, K. E., Eckert, T. L., & Dupaul, G. J. (2002). Early intervention for preschool-age children with ADHD: A literature review. *Journal of Emotional and Behavioral Disorders, 10*(1), 14–28. https://doi.org/10.1177/106342660201000103

McGoey, K. E., DuPaul, G. J., Haley, E., & Shelton, T. L. (2007). Parent and teacher ratings of attention-deficit/hyperactivity disorder in preschool: The ADHD rating scale-IV preschool version. *Journal of Psychopathology and Behavioral Assessment, 29*(4), 269–276.

National Institute for Children's Health Quality. (2002). *NICHQ Vanderbilt Assessment Scales*. Retrieved April 19, 2025, from https://nichq.org/downloadable/nichq-vanderbilt-assessment-scales/

Morales-Hidalgo, P., Hernández-Martínez, C., Vera, M., Voltas, N., & Canals, J. (2017). Psychometric properties of the Conners-3 and Conners Early Childhood Indexes in a Spanish school population. *International Journal of Clinical and Health Psychology, 17*(1), 85–96. https://doi.org/10.1016/j.ijchp.2016.07.003

Morrow, R. L., Garland, E. J., Wright, J. M., Maclure, M., Taylor, S., & Dormuth, C. R. (2012). Influence of relative age on diagnosis and treatment of attention-deficit/hyperactivity disorder in children. *Canadian Medical Association Journal, 184*(7), 755–762. https://doi.org/10.1503/cmaj.111619

Narad, M. E., Garner, A. A., Peugh, J. 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

Newson, J. J., Pastukh, V., & Thiagarajan, T. C. (2021). Poor separation of clinical symptom profiles by DSM-5 disorder criteria. *Frontiers in Psychiatry, 12*, Article 775762. https://doi.org/10.3389/fpsyt.2021.775762

Nussbaum, N. L. (2012). ADHD and female specific concerns: A review of the literature and clinical implications. *Journal of Attention Disorders, 16*(2), 87–100. https://doi.org/10.1177/1087054711416909

Ostrander, R., Weinfurt, K. P., Yarnold, P. R., & August, G. J. (1998). Diagnosing attention deficit disorders with the Behavioral Assessment System for Children and the Child Behavior Checklist: Test and construct validity analyses using optimal discriminant classification trees. *Journal of Consulting and Clinical Psychology, 66*(4), 660–672. https://doi.org/10.1037/0022-006X.66.4.660

Pappas, D. (2006). ADHD Rating Scale-IV: Checklists, norms, and clinical interpretation. *Journal of Psychoeducational Assessment, 24*(2), 172–178. https://doi.org/10.1177/0734282905285792

Paris, J., Bhat, V., & Thombs, B. (2015). Is adult attention-deficit hyperactivity disorder being overdiagnosed? *The Canadian Journal of Psychiatry, 60*(7), 324–328. https://doi.org/10.1177/070674371506000705

Peters, M. D. J., Godfrey, C. M., Khalil, H., McInerney, P., Parker, D., & Soares, C. B. (2015). Guidance for conducting non-systematic scoping reviews. *International Journal of Evidence-Based Healthcare, 13*(3), 141–146. https://doi.org/10.1097/XEB.0000000000000050

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

Reynolds, C. R., Kamphaus, R. W., Vannest, K. J. (2011). Behavior Assessment System for Children (BASC). In: Kreutzer, J.S., DeLuca, J., Caplan, B. (eds) *Encyclopedia of Clinical Neuropsychology*. Springer. https://doi.org/10.1007/978-0-387-79948-3_1524

Quinn, P. O., & Madhoo, M. (2014). A review of attention-deficit/hyperactivity disorder in women and girls: Uncovering this hidden diagnosis. *Primary Care Companion for CNS Disorders, 16*(3), PCC.13r01596. https://doi.org/10.4088/PCC.13r01596

Steer, J., & Bilbow, A. (2021). *Understanding ADHD in girls and women*. Jessica Kingsley Publishers.

Swanson, J. M. (1981). *The SNAP Rating Scale for the diagnosis of the attention deficit disorder*. ERIC Document Reproduction Service. https://eric.ed.gov/?id=ED217047

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.

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., … Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine, 169*(7), 467–473. https://doi.org/10.7326/M18-0850

Ullmann, R. K., Sleator, E. K., & Sprague, R. L. (1984). *ADD-H Comprehensive Teacher Rating Scale (ACTeRS)* [Database record]. *APA PsycTests*. https://doi.org/10.1037/t08014-000

Whitely, M. (2015). Attention deficit hyperactive disorder diagnosis continues to fail the reliability and validity tests. *Australian and New Zealand Journal of Psychiatry, 49*(6), 497–498. https://doi.org/10.1177/0004867415579921

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), 559–568. https://doi.org/10.1093/jpepsy/jsg046

Zelnik, N., Bennett-Back, O., Miari, W., Goez, H. R., & Fattal-Valevski, A. (2012). Is the test of variables of attention reliable for the diagnosis of attention-deficit hyperactivity disorder (ADHD)? *Journal of Child Neurology, 27*(6), 703–707. https://doi.org/10.1177/0883073811423821