Article

The Ability of Self-Report Methods to Accurately Diagnose Attention Deficit Hyperactivity Disorder: A Systematic Review

Journal of Attention Disorders
1–17

DOI: 10.1177/10870547231177470
journals.sagepub.com/home/jad

Allyson G. Harrison [1] and Melanie J. Edwards [1]

Abstract

Objective: To identify and analyze all studies validating rating scales or interview-based screeners commonly used to evaluate ADHD in adults. Method: A systematic literature search identified all studies providing diagnostic accuracy statistics, including sensitivity and specificity, supplemented by relevant articles or test manuals referenced in reviewed manuscripts. Results:

Corresponding Author:
Allyson G. Harrison, Regional Assessment & Resource Centre, Queens University, Mackintosh-Corry Hall, B100, 68 University Avenue, Kingston, Ontario K7L 3N6, Canada.
Email: harrisna@queensu.ca

functional impairment arising from the symptoms, establish chronicity, and rule out other possible causes prior to making this diagnosis. In childhood, this process is often fairly easy (Sibley, 2021). Indeed, a clinician can usually canvass parents and teachers to obtain confirmation of a sufficient number of inattentive and/or hyperactive symptoms in various environments and can typically obtain both educational and medical records to confirm both symptoms and impairment. Furthermore, the few conditions that can mimic the symptoms of ADHD in childhood (e.g., oppositional defiant disorder, substance use disorder, metabolic disorders, mood and anxiety disorders) are easily identified and ruled out (Sibley et al., 2018).

By contrast, diagnosis of ADHD in those over age 18 is more difficult and complex (Kolar et al., 2008; Sibley, 2021), especially in those seeking a first-time diagnosis (Ahmad et al., 2019; Sibley et al., 2018; Sibley, Rohde et al., 2018). Here, it is often more difficult to obtain childhood educational and medical records, ensure that collateral sources who know the person well provide input about both childhood and current symptoms, and rule out other common conditions that mimic the symptoms of ADHD (Ahmad et al., 2019; Caye et al., 2017; Sibley, 2021; Sibley et al., 2018; Sibley, Rohde et al., 2018; Weis et al., 2019). Adult retrospective recall of childhood symptoms is unreliable (Breda et al., 2020; Mannuzza et al., 2002; Miller et al., 2010), making it difficult to determine, with a high degree of confidence, whether an adult met the diagnostic criteria for ADHD in childhood based simply on self-report.

Unfortunately, it seems that many clinicians rely mainly on self-reported symptoms (expressed in semi-structured interviews or on self-report questionnaires) when diagnosing young adults with ADHD. For instance, research has found that the majority of diagnostic reports submitted by young adults seeking academic accommodations at postsecondary schools or on medical licensing exams failed to ensure that all five DSM diagnostic criteria were met before rendering the diagnosis (e.g., Joy et al., 2010; Nelson et al., 2019; Weis et al., 2019). The majority of these submitted reports conferred a diagnosis of ADHD based primarily or exclusively on current self-reported symptoms, with most failing to obtain collateral reports, confirm childhood onset, establish functional impairment, or rule out other potential causes for the reported symptoms.

These trends are worrisome. We know that young adults without ADHD often report experiencing symptoms of ADHD (Harrison, 2004; Harrison et al., 2013; J. A. Suhr & Johnson, 2022) especially when they experience high levels of stress, depression, and/or anxiety (Harrison et al., 2013; Lewandowski et al., 2008; J. A. Suhr & Johnson, 2022), meaning that symptom report alone is not sufficient to confirm this diagnosis. We also know that when clinicians rely on self-reported symptoms alone it increases the false positive rate of diagnosis (Faraone et al., 2003). Indeed, both Gathje et al. (2008) and Gordon et al. (2006) showed that the number of individuals diagnosed with ADHD is dramatically (40%–70%) higher when using symptom report alone relative to when other DSM criteria such as functional impairment are considered prior to diagnosis.

In recent years, there have been a number of contributory issue that might increase levels of stress, anxiety, and depression symptoms in the general population and lead to the experience of ADHD-like symptoms. For instance, the

understand that base rate of the disorder influences the interpretation of obtained scores. Screening tests are not designed to diagnose but rather to identify individuals whose symptoms require more careful evaluation. Because screening tests are often used to identify uncommon disorders (e.g., ones with a low base rate) the cut scores suggested for use on these tests are designed to err on the side of caution, overidentifying many more people than truly have the condition. By contrast, because these screening tests are overly sensitive they rarely miss those who are symptomatic (Gilbert et al., 2001). Similar to previous studies (e.g., Labarge et al., 2003; Morgan et al., 2021), most clinicians diagnosing ADHD in adults may not understand the actual probability of a true positive diagnosis based on a positive screening test score, leading to overdiagnosis.

A Brief Refresher on Sensitivity, Specificity, Positive, and Negative Predictive Values

Given studies showing that many clinicians fail to understand the predictive statistics that inform screening test results, a brief refresher seems in order. Interested readers may also consult any of the good review articles that provide a more comprehensive discussion of these terms (e.g., Gilbert et al., 2001; Lange & Lippa, 2017; Trevethan, 2017).

All tests function on probabilities; a screening test provides the user with a score that is felt to maximize the probability that a true positive case will not be missed while ensuring that very few individuals with a negative score are really symptomatic. Sensitivity is the actual percentage of true positives; how many known positive cases the test detects. In essence, it answers the question, "I already know that my client has the illness in question. What is the chance that this test will show that my client has it?" Specificity, by contrast, is the actual percentage of true negatives; how many known negative cases are correctly classified as such using this test. In essence, it answers the question, "I already know that my client does not have the illness in question. What is the chance that this test shows my client does not have it?"

While these are useful metrics to know about a test, they are usually employed to determine whether a new test works as well as the gold standard method of diagnosis (Lange & Lippa, 2017; Trevethan, 2017). Because sensitivity and specificity are determined by comparing known diagnoses with obtained test scores, they are not influenced by the base rate of the condition.

However, knowing the sensitivity and specificity of a given test does not help a clinician interpret data from a screening test given to an individual client. When evaluating a client in one's office, the clinician does not already know what the true answer is (e.g., they don't know for certain whether the client has the illness or not), and so they rely on the test scores to help decide whether a client's symptoms are consistent with a particular diagnosis. To obtain this type of clinical information, one must instead know the positive predictive value (PPV) and negative predictive value (NPV) of a given test; these predictive values are influenced heavily by the base rate of the disorder or illness within a specified population (Labarge et al., 2003). The PPV answers the question, "my client just tested positive on this test. What is the chance that my client truly has this illness?" The NPV, by contrast, answers the question, "my client just tested negative on this test. What is the chance that my client does not have this illness?" As one can see, these are clinically relevant questions asked by most evaluators completing diagnostic evaluations. To understand how base rate affects PPV and NPV it may be instructive to use a clinical example.

Assume that you have 60 adults whom you know have ADHD (based on gold standard diagnostic procedures). You administer a new ADHD self-report measure to these adults as well as to 60 adults whom you know do not have ADHD. The new test performs as shown in Table 1. As may be seen, the new test correctly identifies 90% of your ADHD sample as having ADHD and 72% of your non-ADHD group as not having ADHD. Hence, sensitivity is 90% and specificity is 72%. Note, too, that these scores would not change depending on how common ADHD is in your sample, because these metrics simply say how often the test correctly identifies persons whose status is already known.

However, it is easy for a test to identify people correctly when half of them have the condition in question. In this example, when half of the people in the sample have ADHD then the PPV is 76.3% and the NPV is 87.8%. In reality, however, ADHD occurs in only about 5% of the adult population (e.g., Kessler et al., 2006). In order to evaluate how the new test functions clinically (when the true diagnosis is not known), we would need to evaluate how the new test performs in a population in which only 5% of people have the condition (rather than 50% as was the case in Table 1). Using the 90% specificity and 72% sensitivity values obtained when testing against the gold standard, we can calculate the PPV and NPV of this new test when the base rate of ADHD is 5%. Table 2 presents the resulting identification rates that would occur if we used this test to determine who did or did not have ADHD in a population of 1,000 people, where only 5% actually have the condition of interest.

Here, out of 1,000 people only 50 truly have ADHD (e.g., 5%) and 950 do not. However, the clinician does not know who has the condition and who does not, and so we use our new test to make this determination. Table 2 shows how our new test performs in this scenario. With a known sensitivity of 90% (e.g., I already know you do have ADHD, and 90/100 times the test gets it right) the new ADHD test

Table 1. Performance of New ADHD Self-Report Test Compared With Gold Standard.

		Results of new ADHD self-report measure		
		Test Says Not ADHD	Tests Says ADHD	Total
Actual diagnosis/reality	Not ADHD	43	17	60
	ADHD	6	54	60
	Total	49	71	120

Table 2. Ability of New Test to Correctly Identify ADHD When Base Rate is 5%.

		Results of new ADHD self-report measure		
		Test Says Not ADHD	Test Says ADHD	Total
Actual diagnosis/reality	Not ADHD	684	266	950
	ADHD	5	45	50
	Total	689	311	1,000

will correctly identify 45/50 individuals as having ADHD. However, applying specificity of 72% to these data (e.g., I already know that you don't have ADHD, and for the 950 people without ADHD the test gets it right 72% of the time), we can see that the new test also falsely identifies 266 of the normal (not ADHD) adults as having ADHD. In other words, for every 311 people the test identifies as ADHD, it is wrong 266 times. Hence, when the base rate of a condi-

Table 3. Inclusion and Exclusion Criteria.

Inclusion criteria	Exclusion criteria
• In peer reviewed journals or published test manual • Participants aged 18 or older • Group study investigating interviews, behavior rating scales, and/or neuropsychological tests for screening or identification of ADHD • ADHD rating scales commercially available or in public domain • Comparison groups: adults diagnosed with ADHD vs. control participants and/or participants with psychiatric disorders or clinical complaints • Results provide diagnostic accuracy statistics, at minimum sensitivity and specificity	• Publication not in English • Sensitivity and specificity scores not provided or calculable • Scales not specific to ADHD symptoms

information regarding diagnostic sensitivity or specificity). A similar method of diagnosis (e.g., use of a semi-structured interview) was used in the Brown (1996), Erhardt et al. (1999), Pettersson et al. (2018), Ustun et al. (2017), and van de Glind et al. (2013) studies. In the Hines et al. (2012) study a randomly selected group of patients presenting at eight different primary care medical practices for a routine appointment (e.g., not attending due to suspected ADHD) were administered the ASRS-v1.1 6-item screening questionnaire (Kessler et al., 2005). Those who scored 4 or more out of 6 on this questionnaire were assumed to have ADHD, and those with lower scores were assigned to the control sample. In the Ward et al. (1993) study, ADHD status was confirmed using the Utah criteria for ADHD, which requires only self-reporting of childhood and current symptoms. Sometimes (e.g., Hines et al., 2012; Pettersson et al., 2018; Solanto et al., 2004; Van Voorhees et al., 2011) the rating scale being evaluated had also been used to inform diagnostic status.

In most other studies, the actual method of ADHD diagnosis for participants was not provided, with most (e.g., Brevik et al., 2020) saying it was a "well validated" group or a group who simply self-identified as ADHD (e.g., J. Suhr et al., 2009). In one study (Kessler et al., 2005) the composition of the groups, final numbers per group, and method of identification were opaque. Nowhere do the authors of the Kessler et al. study actually identify the final number of persons who were or were not considered to have ADHD, and the method by which diagnosis was given is not explained operationally. Notably, in none of the 20 studies reviewed were any performance or symptom validity measures utilized in the assessment or diagnosis phase or when evaluating self-reported symptoms.

Diagnostic Accuracy of Screening Measures

Tables 4 and 5 provide details regarding the classification performance of the various screening measures. While all of the studies (overtly or not) included the specificity and sensitivity of the measure in question, none provided relevant PPV and NPV metrics according to expected base rate of ADHD. Hence, we have provided these in both tables. We chose to provide data for base rates of both 5% (aligning with the higher estimated base rate of adult ADHD in the general population) and 10% (given previous suggestions that the prevalence of adult ADHD in general medical practices is as much as twice the population prevalence (see Kessler et al., 2005)).

Differentiating ADHD From Normal/Non-treatment Seeking Adults

Table 4 presents the results from the 12 evaluations that compared individuals said to have ADHD with non-ADHD individuals. Classification results were, in all but one study, compared with individuals said to be normal, non-ADHD, or adults attending a medical practice for routine complaints other than possible ADHD. Only the Kessler et al. (2005) study was opaque regarding the control group composition (see Table 4 for sample descriptions).

Sensitivity is the true positive value of a test. The higher the score, the fewer false negative results. Table 4 shows that, for about half of the tests reviewed, individuals already known to have ADHD are accurately classified relative to normal individuals. Indeed, nine tests reviewed had a sensitivity of over 90%, whereas 11 screening tests fell below 90% when differentiating non-symptomatic individuals from those said to have ADHD, depending on cut score employed for identification. The lowest sensitivities when differentiating between normal and ADHD individuals were the WURS-25 (J. Suhr et al., 2009) and the ASRS 18 items (Kessler et al., 2005), meaning that a large proportion of those who truly had ADHD were not correctly identified in these studies.

Specificity is a test's ability to correctly identify those without the disease (the true negatives). A highly specific test means that there are few false positive results. Depending on the consequences of incorrect identification, specificity of 90% or higher is often recommended in order to ensure that the false positive rate is low (e.g., Schroeder et al., 2021). When differentiating true ADHD individuals from non-clinical samples, the range was from 99.5%

Table 4.

Estimated rate of ADHD

Table 5. Predictive Values at Two Base Rates of ADHD Comparing Individuals With ADHD and Treatment Seeking/Clinical Samples.

							Estimated Rate of ADHD			
							5%		10%	
Test	Reference	Sample	# Items/ Scale Used	Cut Score Used	Sen (%)	Spe (%)	PPV (%)	NPV (%)	PPV (%)	NPV (%)
ASRS	Dunlop et al. (2018)	40 adult patients with major depressive disorder (42 years), 55 healthy adult control subjects (44								

(Kessler et al., 2005) to a low of 22% (Brevik et al., 2020), with most falling in the mid-range of 40-60% (see Table 4); only six studies found a specificity of 90% or better, meaning that many known normal individuals were falsely identified as having ADHD using these tests.

Of greater interest was the variation in PPV scores when the assumed base rate of ADHD is either 5% or 10%. Here, PPV ranges between a low of 6% (ASRS 18 using a cut-score of ≥16 and a base rate of 5%; Brevik et al., 2020; BAARS-IV when the base rate is 5%; Dvorsky et al., 2016) to a high of 88-94% using ASRS-part A and a 5-10% base rate (Kessler et al., 2005). As may be seen in Table 4, however, a positive score in any of these studies typically had, at best, chance ability to correctly identify those with true ADHD compared with normal adults. By contrast, all screening tests had excellent ability to correctly classify non-ADHD individuals, meaning that there is a very small chance that someone with a score below published cut-offs really has ADHD.

Differentiating ADHD From Other Clinical Samples

Table 5 provides the classification statistics for the 13 studies where individuals said to have ADHD were compared with treatment seeking or clinical samples. The make-up of the clinical samples differed; some were seeking an assessment for ADHD but did not receive a clinical diagnosis, whereas other studies compared individuals with presumed ADHD to those with mental health or other psychiatric conditions (e.g., anxiety disorders, major depressive disorders, substance use disorders). None of the comparator groups were said to be “symptom-free.”

Sensitivity and specificity scores were lower in this sample (see Table 5). Here, sensitivity ranged from 97% (Luty et al., 2009) to 37% (J. Suhr et al., 2009); only six studies found a sensitivity of 90% or greater. Regarding specificity, no test achieved a specificity score above 90%; six were at or above 80% and the lowest two were at 27%.

In almost all cases the self-report screening tests had extremely good NPV when differentiating between ADHD individuals and a clinical sample. At either estimated base rate, a negative score on these measures very rarely misses true cases of ADHD, even in those with comorbid conditions. Exceptions were the ability of the CAARS and the WURS-25 to differentiate substance abuse treatment participants diagnosed retrospectively with ADHD from those who did not screen positive for ADHD (Luty et al., 2009).

The positive predictive value of a screening test score in these clinical samples, by contrast, had only weak ability to correctly classify true cases of ADHD. When tasked with differentiating true ADHD from psychiatric or assessment-seeking populations, the tests with the highest correct classification accuracy at 10% or 5% base rates were: the CAARS (60% and 57% chance that a substance abuse client also had ADHD given a high score; Luty et al., 2009); and the WURS (61% and 59%; Luty et al., 2009). No other studies found that an ADHD screening test/interview had a better than chance ability to correctly identify true ADHD when compared with clinical samples. Indeed, the second-best positive prediction scores were found for the CAARS at a 10% base rate (a high score has a 34% chance of accurate classification; Harrison et al., 2019) and the WURS-25 at the same base rate (33%; Ward et al., 1993). Most had less than a 10% chance of accurate diagnosis given a positive test score (see Table 5).

scale both when differentiating between normal adults and those with ADHD and, of more clinical relevance, when attempting to differentiate individuals with ADHD from those with other clinical conditions or concerns.

It was noteworthy that only about half (nine) of the studies/manuals reviewed actually provided PPV and NPV data for the screening measure being evaluated. For those that did, they almost always reported only PPV and NPV based

conditions lead to false-positive diagnoses in young adults. The discrepancy in positive predictive value between initial development and practical application of a screening test demonstrates why it is vital for such screening measures to be independently validated against clinical samples.

Clinically, differentiating between ADHD and other,

administer semi-structured interviews need to be aware that a positive screening outcome, especially in a clinical setting, has an extremely high false positive rate and a low positive predictive value. This means that clinicians must undertake a rigorous evaluation of clients with positive screening scores, including objective reviews of past history, obtaining opinions from knowledgeable collateral sources, evaluating whether symptoms have caused substantial impairment both historically and currently, and most importantly, ruling out the causal influence of many other, higher base rate disorders such as anxiety, depression, addictions, or symptom overreporting. Furthermore, those who develop ADHD screening measures have a responsibility to evaluate how well these measures predict actual ADHD when compared with a sample of assessment-seeking clients and provide data regarding the positive and negative predictive values of their tests at expected population base rates. Without this validation, clinicians run the risk of inappropriately diagnosing and treating clients for ADHD.

Appendix

Chamberlain, S. R., Cortese, S., & Grant, J. E. (2021). Screening for adult ADHD using brief rating tools: What can we conclude from a positive screen? Some caveats. Comprehensive Psychiatry, 106, 152224. https://doi.org/10.1016/j.comppsych.2021.152224

Chung, W., Jiang, S. F., Paksarian, D., Nikolaidis, A., Castellanos, F. X., Merikangas, K. R., & Milham, M. P. (2019). Trends in the prevalence and incidence of attention-deficit/hyperactivity disorder among adults and children of different racial and ethnic groups. JAMA network open, 2(11), e1914344. https://doi.org/10.1001/jamanetworkopen.2019.14344

Cook, C., Kurtz-Costes, B., & Burnett, M. (2021). Nonprescription stimulant use at a public university: Students' motives, experiences, and guilt. Journal of Drug Issues, 51(2), 376–390.

Copeland, E. (1989). Copeland symptom checklist for adult attention deficit disorders. Southeastern Psychological Institute.

Dandachi-FitzGa

opportunitiesCampus Mental Health: Community of Practice (CoP). Canadian Association of Colleges and University Student Services. file:///C:/Users/RARCUS~1/AppData/

A., Bu, E. T., Franck, J., Konstenius, M., Moggi, F., Dom, G., Verspreet, S., Demetrovics, Z., Kapitány-Fövény, M., Fatséas, M., Auriacombe, M., Schillinger, A., Seitz, A., . Levin, F. R. (2013). Validity of the adult ADHD self-report scale (ASRS) as a screener for adult ADHD in treatment seeking substance use disorder patients. Drug and Alcohol Dependence, 132(3), 587–596. https://doi.org/10.1016/j.drugalcdep.2013.04.010

Van Voorhees, E. E., Hardy, K. K., & Kollins, S. H. (2011). Reliability and validity of self- and other-ratings of symptoms of ADHD in adults. Journal of Attention Disorders, 15(3), 224–234. https://doi.org/10.1177/1087054709356163

Ward, M. F., Wender, P. H., & Reimherr, F. W. (1993). The Wender Utah Rating Scale: An aid in the retrospective diagnosis of childhood attention deficit hyperactivity disorder. American Journal of Psychiatry, 150, 885–890.

Weis, R., Till, C. H., & Erickson, C. P. (2019). ADHD assessment in college students: Psychologists' adherence to DSM-5 criteria and multi-method/multi-informant assessment. Journal of Psychoeducational Assessment, 37(2), 209–225. https://doi.org/10.1177/0734282917735152

Weyandt, L. L., Oster, D. R., Marraccini, M. E., Gudmundsdottir, B. G., Munro, B. A., Rathkey, E. S., & McCallum, A. (2016). Prescription stimulant medication misuse: Where are we and where do we go from here? Experimental and Clinical Psychopharmacology, 24(5), 400–414erime7L4142TTA3W. T2.9400 2T5aiJ/ 1 Tf 0.0ealth Organiz: Whe