University of Groningen

The Validity of Teacher Rating Scales for the Assessment of ADHD Symptoms in the Classroom

Staff, Anouck I.; Oosterlaan, Jaap; van der Oord, Saskia; Hoekstra, Pieter J.; Vertessen, Karen; de Vries, Ralph; van den Hoofdakker, Barbara J.; Luman, Marjolein

Published in:
Journal of Attention Disorders

DOI:
10.1177/1087054720916839

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2021

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Staff, A. I., Oosterlaan, J., van der Oord, S., Hoekstra, P. J., Vertessen, K., de Vries, R., van den Hoofdakker, B. J., & Luman, M. (2021). The Validity of Teacher Rating Scales for the Assessment of ADHD Symptoms in the Classroom: A Systematic Review and Meta-Analysis. *Journal of Attention Disorders*, *25*(11), 1578-1593. https://doi.org/10.1177/1087054720916839

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Journal of Attention Disorders
2021, Vol. 25(11) 1578–1593
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1087054720916839
journals.sagepub.com/home/jad
SAGE

The Validity of Teacher Rating Scales for the Assessment of ADHD Symptoms in the Classroom: A Systematic Review and Meta-Analysis

Anouck I. Staff[1], Jaap Oosterlaan[1,2], Saskia van der Oord[3,4], Pieter J. Hoekstra[5], Karen Vertessen[1], Ralph de Vries[6], Barbara J. van den Hoofdakker[5], and Marjolein Luman[1]

Abstract

Objective: To assess attention-deficit/hyperactivity disorder (ADHD) symptoms in the classroom, most often teacher rating scales are used. However, clinical interviews and observations are recommended as gold standard assessment. This systematic review and meta-analysis evaluates the validity of teacher rating scales. **Method:** Twenty-two studies (N = 3,947 children) assessing ADHD symptoms using teacher rating scale and either semi-structured clinical interview or structured classroom observation were meta-analyzed. **Results:** Results showed convergent validity for rating scale scores, with the strongest correlations (r = .55–.64) for validation against interviews, and for hyperactive–impulsive behavior. Divergent validity was confirmed for teacher ratings validated against interviews, whereas validated against observations this was confirmed for inattention only. **Conclusion:** Teacher rating scales appear a valid and time-efficient measure to assess classroom ADHD; although validated against semi-structured clinical interviews, there were only a few studies available. Low correlations between ratings and structured observations of inattention suggest that observations could add information above rating scales. *(J. of Att. Dis. 2021; 25(11) 1578-1593)*

Keywords

ADHD, validity, rating scale, clinical interview, structured observation

Introduction

Attention-deficit/hyperactivity disorder (ADHD) is one of the most prevalent childhood psychiatric disorders, characterized by age-inappropriate, pervasive, and persistent levels of inattention, hyperactivity, and/or impulsivity (American Psychiatric Association, 2013). Symptoms can manifest in different ways across settings (McConaughy et al., 2010) and are often first recognized in the classroom (Abikoff et al., 2002; Junod et al., 2006). Besides special demands of the classroom on a child's ability to focus, sustain attention, and/or to control his or her behavior, deviant behaviors are often more rapidly recognized by teachers (Lauth et al., 2006). This emphasizes the need for reliable and valid measures to assess ADHD in the classroom.

In addition to information from the parents, teachers are considered to be important informants for the assessment of ADHD symptoms in children (American Academy of Pediatrics, 2000; American Psychiatric Association, 2013). Different methodologies can be used: rating scales, structured or nonstructured teacher interviews, and/or structured or nonstructured observations. Evidence suggests that up to 85% the clinicians reported using teacher rating scales to assess ADHD symptoms at school (Handler & DuPaul, 2005). Furthermore, 64% of the clinicians reported using teacher interviews (structured and nonstructured), whereas only 38% reported to use classroom observations (structured and nonstructured) (Handler & DuPaul, 2005).

There are several *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*)-based standardized behavior rating scales that can be used in the schools, either as a small-band scale assessing only ADHD or as part of a broadband

[1]Vrije Universiteit Amsterdam, The Netherlands
[2]Emma Children's Hospital, Amsterdam UMC, The Netherlands
[3]University of Amsterdam, The Netherlands
[4]KU Leuven, Belgium
[5]University of Groningen, The Netherlands
[6]Amsterdam UMC, VU Medical Center, Amsterdam, The Netherlands

Corresponding Author:
Anouck I. Staff, Section Clinical Neuropsychology, Vrije Universiteit Amsterdam, Van der Boechorststraat 7-9, 1081 BT Amsterdam, The Netherlands.
Email: a.i.staff@vu.nl

scale assessing a broader range of child psychiatric conditions including ADHD. Small-band scales usually include items displaying *DSM* symptoms of the disorder scored on a Likert-type scale and are highly sensitive and specific in distinguishing between children diagnosed with ADHD and typically developing community controls (sensitivity and specificity $>$ 94%; American Academy of Pediatrics, 2000). However, rating scales are not recommended as a sole diagnostic tool for assessing classroom ADHD (American Academy of Pediatrics, 2000; National Institute for Health and Care Excellence, 2018; Pliszka, 2007), because ratings may be biased by projection bias or halo effects (Burns et al., 2003; DuPaul, 2003; Gomez et al., 2003), and they do not take functional impairment into account (National Institute for Health and Care Excellence, 2018). In addition, teachers are not trained in diagnosing psychopathology in children.

Semi-structured clinical interviews are less sensitive to bias and accepted as the gold standard assessment method for the evaluation of the presence, duration, frequency, severity, and onset of ADHD symptoms in the classroom (Pelham et al., 2005; Pliszka, 2007; Taylor & Sonuga-Barke, 2008; Volpe et al., 2005). Semi-structured clinical interviews require teachers to describe behaviors in several situations, related to ADHD and comorbid conditions. On the basis of the derived descriptions, the clinician rates symptoms as present or absent, taking functional impairment into account. Although published studies into the reliability of semi-structured clinical teacher interviews are lacking, one study showed high test–retest reliability of two interviewers contacting teachers within a 2-week timeframe (r = .79–.94; Valo & Tannock, 2010). Regarding validity, this study showed that children who met criteria for ADHD on the Teacher Telephone Interview (TTI; Tannock et al., 2002) were more likely to score above the clinical range on the corresponding Conners' Teacher Rating Scale—Revised (CTRS-R; Conners, 1997) subscales (χ^2 = 19.39 for inattention and 62.46 for hyperactivity–impulsivity). A fallback of *DSM*-based semi-structured clinical interviews is that they are time-consuming (particularly for teachers) and have to be conducted by a (trained) clinician.

Systematic observations are viewed as one of the most objective and direct measurements of a child's behavior including ADHD (Volpe et al., 2005). They are frequently used to assess ADHD symptoms at school, particularly among school psychologists (Handler & DuPaul, 2005; Shapiro & Heick, 2004; Wilson & Reschly, 1996). The most commonly used observational coding schemes assessing ADHD symptoms focus on inattention by measuring off-task behavior and hyperactivity by measuring motor movement and noisiness. Impulsivity is usually not explicitly assessed in these coding schemes, but is taken into account when scoring disruptive and oppositional behaviors (see, for reviews, Pelham et al., 2005; Volpe et al., 2005). However, also direct observations are time-consuming (e.g., observers need to be trained, observations have to be conducted, and coding behavior is time-consuming). Although studies into the validity of coding schemes are limited, reliability of most coding schemes appears to be acceptable (r = .61–1, ϕ = .60–1, κ = .39–.99; Minder et al., 2017; Pelham et al., 2005).

Little research has been conducted to establish the validity of teacher rating scales in assessing ADHD symptoms at school (Parker & Corkum, 2016). To date, only the validity of combined parent and teacher ratings has been studied in relation to a full diagnostic assessment including classroom observations, review of the child's school records, semi-structured interviews with teacher and parent, and a standardized assessment of the child's cognitive abilities and skills, showing high sensitivity (McGonnell et al., 2009; Parker & Corkum, 2016), but moderate specificity (Parker & Corkum, 2016). This is in line with findings of a review by Snyder et al. (2006) that showed moderate to high overall accuracy of parent and teacher rating scales for classifying children with ADHD. So far, validity studies of teacher rating scales (e.g., not in combination with parent ratings) are lacking, although studies do report positive associations between teacher ratings and structured classroom observations (see, for review, Minder et al., 2017). This review showed small to moderate, occasionally strong, convergent validity (r = .02–.50) for total ADHD symptoms, although they did not look separately at inattention and hyperactivity–impulsivity. Furthermore, other assessment methods (e.g., clinical interviews) were not taken into account.

As there is currently no systematic review of studies into associations between teacher rating scales and either clinical interview or structured classroom observation instruments, this systematic review and meta-analysis was aimed at aggregating available studies on the validity of teacher ratings of inattention and hyperactivity–impulsivity symptoms. We expected correlations between teacher rating scales and semi-structured clinical teacher interviews or structured observations to be stronger for overt behavior (hyperactivity and impulsivity) than for covert behavior (inattention) (Atkins et al., 1989; Lauth et al., 2006; Milich & Landau, 1988; Whalen et al., 1978, 1979). Furthermore, differences in the demands put on boys and girls and differences in the expectations from boys and girls, as well as changing demands and expectations across development, might influence the perception of behavior as reflected in teacher ratings. Teachers are more likely to underestimate ADHD symptoms in girls (Meyer et al., 2020), and identify inattention in girls as attentional or emotional difficulties rather than ADHD symptoms (Groenewald et al., 2009). Similarly, teachers are more likely to interpret behavior of the youngest children in a class as reflecting ADHD symptoms rather than young but age-appropriate behavior (Halldner et al., 2014; Krabbe et al., 2014). To verify if observed relations are stable across sex and age, we studied whether validity differs for boys and girls and across ages.

Method

Study Selection and Description

This systematic review and meta-analysis was carried out in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement (www.prisma-statement.org). Studies had to meet the following criteria to be included: (a) The study was published in the English language in an academic peer-reviewed journal between 1980 (introduction of ADHD in the *DSM* [3rd ed.; *DSM-III*; American Psychiatric Association, 1980]) and January 2020 (final literature search), (b) participants in the study attended elementary school (average age between 6 and 12 years), and (c) ADHD-related behavior was evaluated using a teacher rating scale aimed to assess ADHD symptoms based on the description according to *DSM-III* or more recent editions, and either (1) a semi-structured clinical teacher interview assessing ADHD symptoms, or evaluated using (2) a structured classroom observation of ADHD behavior by an independent observer, according to Hintze's (2005) definition of systematic direct observation (see below). Both clinical and community samples were included. With regard to medication use, studies were included only if the different methods of assessment were administered under the same medication condition. Studies were excluded if (a) the study sample mainly consisted of children diagnosed with psychiatric disorders other than ADHD, such as oppositional defiant disorder, conduct disorder, autism spectrum disorder, anxiety disorder, and tic disorder, or of children with neurological dysfunctions, such as epilepsy; (b) the study sample consisted of only children with intellectual disability (with or without comorbid ADHD symptoms); (c) the clinical interview or classroom observation did not yield a quantitative outcome measure (e.g., only descriptions of behavior, without frequency or duration); or (d) the reporting pertained to (single) case studies. In the case multiple articles were published using the same sample, and if dependent variables did not differ between studies, we included the study with the largest sample size, the most comprehensive description of the assessment of ADHD symptoms, or the first published paper on the sample, respectively. In case insufficient data were reported on the association between teacher ratings and interviews and/or observation scores, the authors of the relevant studies were contacted.

A comprehensive search was performed in the bibliographic databases PubMed, EMBASE, Ebsco/ERIC, and Ebsco/PsycINFO, from inception up to January 17, 2020. The search was conducted in collaboration with a medical librarian. The following terms were used (including synonyms and closely related words) as index terms or free-text words: "Attention Deficit Disorder with Hyperactivity," "School/Classroom Observation," and "Teacher Interview." Duplicate articles were excluded. The full search strategies for all databases can be found in Supplementary File S1. The first author (A.I.S.) and a second independent assessor (K.V.) screened all articles for eligibility on title and abstract. Thereafter, full-text articles were screened for eligibility. Conflicts were resolved by consensus. Reference lists of included articles were searched for additional articles satisfying the inclusion criteria.

Definitions and Outcome Measures

Characteristics of the sample and information on the rating scale, interview, and/or classroom observation method were extracted from the included studies. Ideally, for each instrument, scores on the inattention and hyperactivity–impulsivity subscales as well as total ADHD symptom score were extracted.

Teacher ratings of ADHD behavior. Teacher ratings of ADHD behavior included rating scales purported to assess the symptom domains of ADHD (e.g., symptoms of inattention, hyperactivity–impulsivity, and/or total ADHD symptoms). If studies reported on two or more rating scales, then we included the rating scale that had the highest validity for measuring *DSM*-related symptoms and/or that included data for all subscales (i.e., inattention, hyperactivity–impulsivity, and total ADHD scale), or assessed the largest sample, respectively. When higher scores on the rating scale indicated fewer problems, the scores on the rating scale were reversed by multiplying the score by −1.

Clinical teacher interviews. Teacher interviews included semi-structured clinical teacher interviews, assessing ADHD symptom domains according to *DSM-III* or more recent releases. All interviews were with the primary teacher of the child (e.g., the teacher spending most time with the child). ADHD symptoms were rated by a clinician or trained interviewer according to a protocol or manual of the clinical interview.

Structured classroom observations. We have used Hintze's (2005) definition of systematic direct observation to include studies using observational measures, meaning that studies were included if (a) ADHD symptom domains (e.g., on-task and/or off-task behavior as a measure of inattention, hyperactivity, and impulsivity) were assessed; (b) to be observed (e.g., coded or scored) symptoms were fully operationalized; (c) observations were conducted using standardized procedures; (d) observations were conducted in the classroom (not during an individual test session); and (e) the observations were rated by independent observers (e.g., observers were unaware of a child's diagnostic and treatment status) using standardized instructions. Furthermore, (f) either ADHD symptoms were rated on a quantitative scale in terms of frequency, duration, or percentage of total time, or descriptions of behavior were rated on a scale (e.g., Likert-type scale). For observation methods that report on different behaviors measured on the same scale (e.g., both interval coding or continuous coding within the same time period) that pertain to a similar symptom domain (e.g., repetitive movements, noisiness, and interrupting behavior as a measure of hyperactivity–impulsivity), raw scores were aggregated to obtain a single score for the corresponding ADHD symptom domain, by calculating the sum of frequency or duration of these behaviors (within the same time period), following previous studies (Epstein et al., 2005; Junod et al., 2006). When inattention was operationalized as being on-task rather than off-task behavior, the inattention subscale was reversed by multiplying the score by −1. Scales measuring only disruptive behavior and scales assessing both hyperactive and rule-breaking behavior (e.g., oppositional or aggressive behavior) were excluded.

Background variables. The background variables sex, age, medication use, and comorbid psychiatric diagnosis were extracted from the articles or requested from the authors. Sex was defined as the percentage of male participants in the study sample. Age was the mean age in years. When the study sample included children on medication, the percentage of children on medication at baseline was extracted. Comorbidity was defined as all diagnoses other than ADHD, according to the *DSM* guidelines.

Statistical analyses. Statistical analyses were performed using SPSS version 24.0 (IBM Corp., 2016) and Comprehensive Meta-Analysis (Borenstein et al., 2005). For all included studies, we extracted correlations between rating scale scores and interviews or observational measures calculated across the full study sample, number of participants (N), the reported measure of association (*Pearson's r* or *Spearman's rho*), and the accompanying significance level (p). To maximize homogeneity between study samples, we used raw correlations without any covariates (e.g., partial correlations). If raw correlations were not available, authors were contacted. For studies reporting correlations only for subgroups (e.g., ADHD and controls), authors were contacted to provide measures of association pertaining to the full study sample to maximize the distribution of scores within samples to allow a dimensional rather than a categorical approach.

First, meta-analytic effect sizes were calculated for the association between rating scale scores and interview or observational measure. A minimum of three studies were used to calculate meta-analytic effect sizes (Borenstein et al., 2011). Correlations of .10, .30, and .50 were interpreted as small, medium, and large effects, respectively (Cohen, 1988). Differences between meta-analytic effects for symptom domains and assessment methods were tested for significance using Fisher's r-to-z transformations (Borenstein et al., 2011). Exploratory meta-regression analyses were used to test whether background variables confounded the results. A minimum of 10 studies were used to calculate meta-regression effects (Borenstein et al., 2011). Sensitivity analyses were used to check whether specific instrument characteristics (e.g., type of rating scale, interview or observation method, number of observations) or sample (e.g., clinical sample or community sample) affected the meta-analytic effects. For example, it was examined whether correlations between teacher ratings and observations differed for observations conducted on single or multiple schooldays, given that observations conducted on multiple days are more representative for the child's behavior as rated by teachers using a rating scale.

All meta-analytic effect sizes were computed using the random-effects model (method of moments estimation), because heterogeneity may have been introduced using data from different instruments to assess ADHD symptoms according to different editions of the *DSM*. Q and I^2 tests were used to calculate heterogeneity (Borenstein et al., 2011). A significant p value of the Q test indicates heterogeneity. I^2 test values of 25%, 50%, and 75% represent low, moderate, and high heterogeneity, respectively (Higgins et al., 2003).

Publication bias. The possibility of publication bias was assessed for all meta-analytic outcomes based on a minimum of 10 studies (Cochrane Collaboration, 2011). Two methods were used: (a) The degree of funnel plot asymmetry was determined with the method as proposed by Egger et al. (1997) (two sided, $\alpha = .05$), and (b) Rosenthal's fail-safe N was calculated to determine the number of studies needed to nullify the meta-analytic effect (Rosenthal, 1995).

Study quality. To assess study quality, four items of the Critical Appraisal Checklist for Diagnostic Test Accuracy Studies (Joanna Briggs Institute, 2017) were used regarding patient selection, parallel assessment of both measures, and bias due to missing data. The selected items are presented in Supplementary File S4, and each item was independently scored "yes," "no," or "unclear" by two of the authors (A.I.S. and K.V.). Conflicts were resolved by consensus. The sum of items scored "yes" was taken as the measure of study quality (range 0–4). Correlational analyses were conducted to assess whether study quality was related to effect size findings for convergent validity of inattention, hyperactivity–impulsivity, and total ADHD.

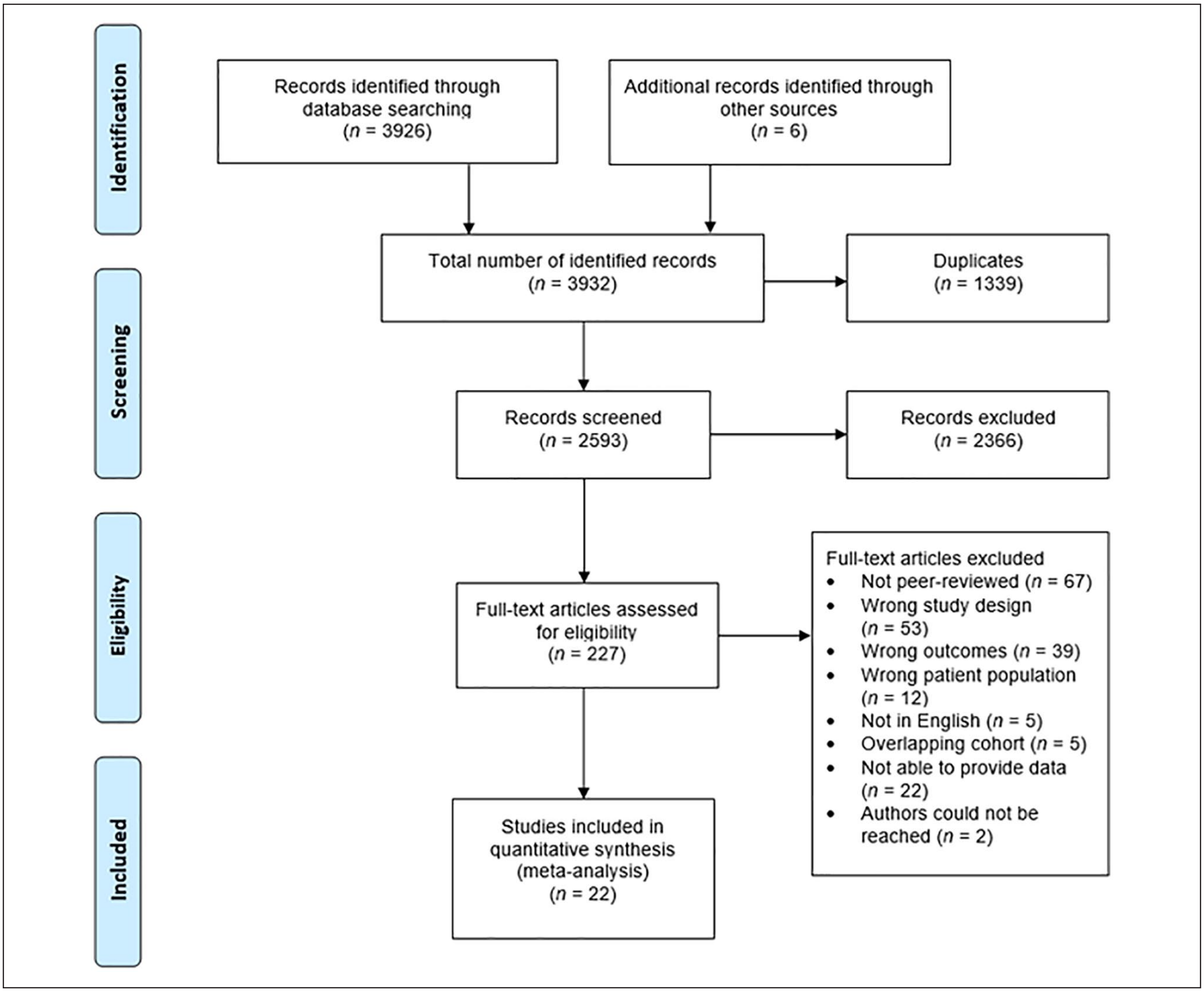
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 3926)
Additional records identified through other sources (n = 6)
Total number of identified records (n = 3932)
Duplicates (n = 1339)
Records screened (n = 2593)
Records excluded (n = 2366)
Full-text articles assessed for eligibility (n = 227)
Full-text articles excluded
• Not peer-reviewed (n = 67)
• Wrong study design (n = 53)
• Wrong outcomes (n = 39)
• Wrong patient population (n = 12)
• Not in English (n = 5)
• Overlapping cohort (n = 5)
• Not able to provide data (n = 22)
• Authors could not be reached (n = 2)
Studies included in quantitative synthesis (meta-analysis) (n = 22)

Figure 1. PRISMA flowchart of the study selection procedure.
Note. PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

Results

The initial search yielded 51 studies eligible for inclusion in our meta-analysis. Five papers reported data on the same sample, and from the remaining 46 articles, 24 did not report on all necessary outcome measures of interest for which data were either not available from the authors upon request ($n = 22$), or authors could not be reached ($n = 2$). A total of 3,947 children from the 22 remaining studies were included in this meta-analysis. Figure 1 presents an overview of the inclusion of studies and reasons for exclusion.

Meta-Analytic Associations Between Rating Scale and Clinical Interview Scores

A total of 1,744 children from only four different studies contributed data to the meta-analysis of the associations between rating scale and clinical interview scores. Table 1 provides an overview of study characteristics. All studies included primary school children (5–13 years) at risk for ADHD, including both boys and girls (the percentage of boys ranged from 69% to 76%). Teacher ratings were collected using two different rating scales (CTRS-R, Conners, 1997; Strengths and Weaknesses of ADHD-Symptoms and Normal-Behaviors [SWAN], Swanson et al., 2012). The semi-structured interview measure used in all studies was the TTI (Tannock et al., 2002). A summary of the characteristics of the included instruments is provided in Supplementary File S3.

Convergent validity. Meta-analytic results and heterogeneity statistics are described in Table 2 and Supplementary File S2 (Supplementary Figure S2a). Meta-analytic correlations between rating scales and interview measures assessing the

Table 1. Studies Assessing ADHD Symptoms Using Teacher Rating Scales and Clinical Interview.

Study	Sample size	Age range in years (*M*)	%Boys	%On ADHD medication	Comorbid psychiatric diagnosis	Sample description	Exclusion criteria	Rating scale	Interview measure	Remarks on methodology
Charach et al. (2009)	1,038	6–12 (8.8)	76	19	14% ODD 4% CD 27% RD 14% LI 15% IQ < 85	Referred for attention and behavioral problems, 87% met *DSM-IV-TR* criteria for ADHD based on TTI	Psychotropic medications other than stimulants, attending full-time residential or day treatment program	CTRS-R:L (59 items): *DSM-IV* Inattentive, *DSM-IV* Hyperactive–impulsive, *DSM-IV* Total symptoms	TTI-IV: Inattention, Hyperactivity–impulsivity, Total	Spearman
Valo & Tannock (2010)[a]	321	6–13 (8.4)	74	10	34% ODD 7% CD	Referred for inattention and disruptive behavior, 66% were given a *DSM-IV* ADHD diagnosis	PDD, schizoid disorders, developmental or intellectual delays	CTRS-R:S (28 items): *DSM-IV* Inattentive, *DSM-IV* Hyperactive–impulsive, *DSM-IV* Total symptoms	TTI-IV: Inattention, Hyperactivity–impulsivity, Total scale	Partial correlation[b]
Parker & Corkum (2016)	279	5–12 (8.49)	69	0	48% LD	Referred to an ADHD clinic, 52% were given a *DSM-IV* ADHD diagnosis	Psychotropic medication, psychoeducational assessment <2 years	CTRS-R:L (59 items): *DSM-IV* Inattentive, *DSM-IV* Hyperactive–impulsive, *DSM-IV* Total symptoms	TTI-IV: Inattention, Hyperactivity–impulsivity, Total	Pearson
Veenman et al. (2019)	114	6–13 (8.94)	76	0	1% PDD 1% CD	Population sample screened for ADHD, 10% had *DSM-IV-TR* diagnosis and 53% met criteria based on TTI	ADHD treatment (<6 months), neurological or severe physical condition, IQ < 80	SWAN (18 items): Inattention, Hyperactivity–impulsivity, Total $\alpha = .91$	TTI-IV: Inattention, Hyperactivity–impulsivity, Total	Spearman

Note. ODD = oppositional deviant disorder; CD = conduct disorder; RD = reading disability; LI = language impairment; IQ = Intelligence Quotient; *DSM-IV-TR* = *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text rev.); TTI = Teacher Telephone Interview; CTRS-R:L = Conners' Teacher Rating Scale—Revised: Long Form; *DSM-IV* = *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.); PDD = pervasive developmental disorder; CTRS-R:S = Conners' Teacher Rating Scale—Revised: Short Form; LD = learning disorder; SWAN = Strengths and Weaknesses of ADHD-Symptoms and Normal-Behaviors.

[a] For this study, data were provided by the author, which contained a larger sample than the sample described in this article. [b] For this study, only correlations controlling for sex, age, and total scores of the remaining dimensions of the CTRS-R:S were available.

Table 2. Overview of Meta-Analytic Results of Correlations Between ADHD Reports on Teacher Rating Scales and Clinical Interview.

		Meta-analytic effect size					Heterogeneity		
Rating scale	Clinical interview	*N*	#Studies	*r*	95% CI	*p*	*Q*	I^2	*p*
Inattention	Inattention	1,740	4	.548	[.506, .588]	<.001	3.90	23.12	.272
	Hyperactivity–impulsivity	1,740	4	.175	[.038, .305]	.013	19.16	84.34	<.001
	Total	1,742	4	.426	[.344, .502]	<.001	9.21	67.44	.027
Hyperactivity–impulsivity	Inattention	1,741	4	.200	[−.050, .426]	.116	65.37	95.41	<.001
	Hyperactivity–impulsivity	1,741	4	.641	[.540, .723]	<.001	24.23	87.62	<.001
	Total	1,743	4	.539	[.331, .697]	<.001	68.71	95.63	<.001
Total	Inattention	1,744	4	.450	[.339, .549]	<.001	17.37	82.73	.001
	Hyperactivity–impulsivity	1,744	4	.505	[.378, .613]	<.001	25.15	88.07	<.001
	Total	1,747	4	.580	[.446, .688]	<.001	33.87	91.14	<.001

Note. CI = confidence interval.

same symptom domain of ADHD ($n = 4$) were all positive and strong ($r = .548$, $p < .001$ for inattention; $r = .641$, $p < .001$ for hyperactivity–impulsivity; $r = .580$, $p < .001$ for total ADHD symptoms). The aggregated correlation for rating scale and interview scores of hyperactivity–impulsivity was significantly stronger than that for inattention ($z = 4.25$, $p < .001$). Effect sizes for hyperactivity–impulsivity and total ADHD symptoms showed significant heterogeneity ($I^2 = 87.62$ and $I^2 = 91.14$, respectively), indicating that there is variability in the magnitude of observed correlations. Heterogeneity for inattention was low ($I^2 = 3.90$).

Divergent validity. The meta-analytic correlation between teacher ratings on the inattention subscales and the hyperactivity–impulsivity of the interview measure was significant ($r = .175$, $p = .013$), but significantly weaker than the aggregated correlation between the inattention scales of the two assessment measures ($r = .548$; $z = 12.93$, $p < .001$), supporting divergent validity. Heterogeneity was high for the meta-analytic correlation between ratings of inattention and hyperactivity–impulsivity as measured by interview ($I^2 = 84.34$).

The meta-analytic correlation between teacher ratings on the hyperactivity–impulsivity subscale and the inattention scale of the interview was not significant ($r = .200$, $p = .116$) and significantly weaker than the aggregated association between the hyperactivity–impulsivity subscales of the two assessment measures ($r = .641$; $z = 16.42$, $p < .001$), supporting divergent validity. Heterogeneity was high for the meta-analytic correlation between hyperactivity–impulsivity ratings and by interview-assessed inattention ($I^2 = 95.41$).

Meta-Analytic Associations Between Rating Scale and Structured Observation Scores

A total of 2,203 children from 18 different studies were included in this meta-analysis. Study characteristics are provided in Table 3. All studies included primary school children (5–14 years) from either community samples, samples at risk for ADHD, or samples of children with a clinical ADHD diagnosis. Samples included both boys and girls (percentage boys ranged from 47% to 100%). Seven different ADHD teacher rating scales were used to measure ADHD symptoms of children in the classroom, and studies used nine different structured classroom observational instruments to assess ADHD symptoms in the classroom (see Table 3). A summary of the characteristics of the included instruments is provided in Supplementary File S3.

Convergent validity. Meta-analytic results and heterogeneity statistics are described in Table 4 and Supplementary File S2 (Supplementary Figure S2b) and showed that meta-analytic correlations between subscales of the rating scale and observational measures measuring inattention ($n = 12$), hyperactivity–impulsivity ($n = 10$), or total ADHD symptoms ($n = 6$) were all significant and small to moderate ($r = .211$, $p < .001$ for inattention; $r = .294$, $p < .001$ for hyperactivity–impulsivity; $r = .261$, $p < .001$ for total ADHD symptoms). The aggregated correlation between hyperactivity–impulsivity subscales of the two assessment measures was significantly stronger than the correlation between the inattention subscales of the two measures ($z = 2.59$, $p = .010$). Effect sizes for all subscales showed high heterogeneity: $I^2 = 77.94$ for inattention, $I^2 = 71.54$ for hyperactivity–impulsivity, and $I^2 = 78.68$ for total ADHD symptoms.

Divergent validity. The meta-analytic correlation between teacher ratings of inattention and observed hyperactivity–impulsivity ($n = 12$) was not significant (trend effect) ($r = .120$, $p = .075$), and this significantly differed from the convergent validity measure for inattention ($r = .211$; $z = 2.74$, $p = .006$), indicating divergent validity for the inattention scales of teacher rating scales. For ratings of inattention compared with hyperactivity–impulsivity as measured by structured observation, heterogeneity was high ($I^2 = 83.67$).

Table 3. Studies Assessing ADHD Symptoms Using Teacher Rating Scales and Structured Classroom Observations.

Study	Sample size	Age range in years (*M*)	%Boys	%On ADHD medication	Comorbid psychiatric diagnosis	Sample description	Exclusion criteria	Rating scale	Observational measure	Remarks on methodology
Brewis (2002)	219	6–12 (no information)	50	0	No information	Population sample, no information on %ADHD	Psychotropic medication use	BASC-TRS (19 items): Attention Problems, Hyperactivity	Total duration of Attention states, Hyperactivity–impulsivity states, ADHD total score. $\kappa =$ not reported	Spearman. OBS: 10 × 2 min on 2 days, CS
DuPaul (1991)	55	6–12 (8.9)	47	No information	No information	Population sample, no information on %ADHD	No information	ACTRS (10 items): Total score	On-task percentage. $\kappa = .74$	Pearson. OBS: 3 days × 20 min, CA, TS
DuPaul et al. (1998b)	53	5–14 (9.8)	47	No information	No information	Population sample, no information on %ADHD	Psychotropic medication use <6 months	ADHD-RS-IV (18 items): Inattention, Hyperactivity–impulsivity	ADHD behavior code: Off-task, Fidgets. κ (mean) $= .56$	Pearson. OBS: 1 × 15 min, CA, TS
DuPaul et al. (2006)	241 (175 ADHD, 66 MC)	6–12 (8.58)	76	22	38% ODD 15% CD	Population sample who met criteria for ADHD and MC	ASD, developmental disabilities, brain damage, visual or hearing impairments, MR	CTRS-R:L (59 items): *DSM-IV* Inattentive, *DSM-IV* Hyperactive–impulsive, *DSM-IV* Total symptoms	BOSS: Engagement (Active engaged time + Passive engaged time), Off-task (Off-task motor + Off-task verbal), ADHD total score (Engagement + Off-task). $\kappa = .88$–$.98$	Spearman. OBS: 2 × 15 min (M, R) on 1 day, TS
Epstein et al. (2005)	528	7–9 (8.4)	81	0	56% DBD	Participants of the MTA study, all met *DSM-IV* ADHD-C criteria	No information	CTRS-R (39 items): *DSM-IV* Inattentive, *DSM-IV* Hyperactive–impulsive, *DSM-IV* Total symptoms	COC: Off-task, Hyperactivity–impulsivity (Interference + Interference to teacher + Gross motor standing + Gross motor vigorous), ADHD total score (Off-task + Hyperactivity–impulsivity). $\phi = .80$–1	Spearman. OBS: 1–2 × 16 min (baseline, on 1 day), CA, TS
Imeraj et al. (2016)	62	6–12 (8.94)	81	84 children with ADHD	8% ODD 3% CD	Children with formal ADHD-C diagnosis and MC	IQ < 80, PDD, neurological disorder, psychotropic medication use except MPH	DBDRS (18 items): Inattention, Hyperactivity–impulsivity	GUCCI: Off-task, Hyperactivity (Hyperactivity + Noisiness), ADHD total score (Off-task + Hyperactivity). $\kappa = .77$–$.99$	Spearman. OBS: 4 × 60 min, on 2 days, CA, CS
Jiang et al. (2019)	135	No information (8.39)	71	0	No comorbid diagnosis	Children referred for ADHD, who met study criteria for ADHD symptoms and impairment	IQ < 80	CSI (18 items): Inattention, Hyperactivity–impulsivity	BOSS: Engagement (Active engaged time + Passive engaged time). $\kappa = .83$	Pearson. OBS: 3 days × 48 15-s intervals, TS
Johnson et al. (2020)	42 (21 ADHD 21 MC)	8–11 (9.7)	76	5	No comorbid diagnosis	Population sample who met study criteria for ADHD and impairment, and MC	ASD	SWAN (18 items): Inattention, Hyperactivity–impulsivity	DOF: On-task, norm-referenced *T* score. $\kappa =$ not reported	Spearman. 4 × 10 min (academic subject) on 3 days, TS
Kennerley et al. (2018)	43	6–12 (8.69)	80	27	47% ODD	Children with or referred for ADHD, all met formal criteria for ADHD	No information	ADHD-RS-IV (18 items): Inattention, Hyperactivity–Impulsivity, Total	BOSS: Engagement (Active engaged time + Passive engaged time), Off-task (Off-task motor + Off-task verbal), ADHD total score (Engagement + Off-task).$\kappa = .84$–$.90$	Pearson. OBS: 1 × 30 min (academic subject), TS
Lauth et al. (2006)	106 (53 ADHD, 53 MC)	7–11 (8.48)	72	0	No information	Population sample who met criteria for subclinical ADHD and MC	No information	CTRS:S (12 items): Total score	MAI: Expected behavior, Actively disruptive. $\kappa =$ not reported	Pearson. OBS: 4 days × 3–9 5-s intervals, TS

(continued)

Table 3. (continued)

Study	Sample size	Age range in years (*M*)	%Boys	%On ADHD medication	Comorbid psychiatric diagnosis	Sample description	Exclusion criteria	Rating scale	Observational measure	Remarks on methodology
McConaughy et al. (2010)	310	6–12 (8.2)	69	0	No information	Population sample who met criteria for subclinical ADHD and MC	Parent-reported physical or medical problems, MR, ASD, PDD	ADHD-RS-IV (18 items): Inattention, Hyperactivity–impulsivity	DOF: Rating on problem items of Attention problems, Intrusive, Attention deficit/hyperactivity. IRR = .72–.80	Pearson. OBS: 3–4 × 10 min on 2 days, TS
Minder et al. (2018)	77	8–15 (10.76)	65	31	26% ODD 4% AD	Population sample who met study criteria for clinically relevant symptoms of ADHD	ASD, TD, other psychiatric disorders, neurological diseases, psychotropic medication except MPH, IQ $<$ 80	Conners-3 (111 items): Inattention, Hyperactivity–impulsivity, Total score	[a]BOSS: Engagement (Active engaged time + Passive engaged time), Off-task (Off-task motor + Off-task verbal), ADHD total score (Engagement + Off-task). κ = .64–.73	Spearman. OBS: No information on length (baseline), TS
Nolan & Gadow (1994)	34	5–13 (9.1)	91	0	53% tics 8% MR	Participants diagnosed with ADHD using criteria of *DSM-III* or *DSM-III-R*	Psychosis, seizure disorder, major organic brain dysfunction, major medical illness, PDD	ATRS (10 items): Total score	COC: Off-task, Hyperactivity–impulsivity (Mean motor movement, Interference). κ = .77–.94	Pearson. OBS: 3–4 days × 20 min (placebo), CA, TS
Pelham et al. (1993)	31	5–9 (8.23)	100	0	32% ODD 48% CD	Participants diagnosed with ADHD using *DSM-III-R*	No information	IOWA (five items): Inattention–overactivity	COCADD: On-task. κ = .75	Pearson. OBS: 38 × 15-s intervals (adaptation period) on 10 days, CA, TS
Pelham et al. (1999)	25	5–12 (9.6)	84	0	52% ODD 32% CD	Participants diagnosed with ADHD using *DSM-IV*	No information	IOWA (five items): Inattention–overactivity	COCADD: On-task. κ = .80	Pearson. OBS: 38 × 15-s intervals (placebo) on 5 days, TS
Pfiffner et al. (2013)	57	5–12 (8.1)	70	7	No comorbid diagnosis	Population sample who met study criteria for ADHD	PDD, full-day special education, visual or hearing impairments, LD, psychosis	CSI (18 items): Inattention, Hyperactivity–impulsivity, Total	BOSS: Engagement (Active engaged time + Passive engaged time), Off-task (mean Off-task motor, Off-task verbal), ADHD total score (Engagement + Off-task). κ ADHD Composite score = .86	Pearson. OBS: 3 days × 15 min (baseline), TS
Stevenson et al. (2010)	144	8–9 (8.85)	52	0	No information	Population sample, none had ADHD diagnosis	No information	ADHD-RS (10 items): Total score	COC: ADHD total score. IRR $>$ .87	Pearson. OBS: 3 days × 8 min (baseline), CA, TS
Veenman et al. (2017)	114	6–13 (8.94)	76	0	1% PDD 1% CD	Population sample screened for ADHD, 10% had *DSM-IV-TR* diagnosis and 53% met criteria based on TTI	ADHD treatment ($<$6 months), neurological or severe physical condition, IQ $<$ 80	SWAN (18 items): Inattention, Hyperactivity–impulsivity, Total score	COC: Off-task, Hyperactivity–impulsivity (Interference + Interference to teacher + Gross motor standing + Gross motor vigorous), ADHD total score (Off-task + Hyperactivity–impulsivity). κ = not reported	Spearman. OBS: 2 × 8 min (baseline) on 1 day, CA, TS

Note. BASC-TRS = Behavior Assessment for Children—Teacher Rating Scale; OBS = observation; CS = continuous sampling; ACTRS = Abbreviated Conners' Teacher Rating Scale; CA = concurrent assessment of the rating scale and observational measure; TS = time sampling; ADHD-RS-IV = ADHD Rating Scale-IV: School Version; MC = matched controls; ODD = oppositional deviant disorder; CD = conduct disorder; ASD = autism spectrum disorder; MR = mental retardation; CTRS-R:L = Conners' Teacher Rating Scale—Revised: Long Form; *DSM-IV* = *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.); BOSS = Behavioral Observation of Students in Schools; DBD = disruptive behavior disorder; MTA = Multimodal Treatment of Attention Deficit Hyperactivity Disorder; ADHD-C = attention-deficit/hyperactivity disorder, combined; CTRS-R = Conners' Teacher Rating Scale—Revised; COC = Classroom Observation Code; IQ = Intelligence Quotient; PDD = pervasive developmental disorder; MPH = methylphenidate; DBDRS = Disruptive Behavior Disorders Rating Scale; GUCCI = Ghent University Classroom Coding Inventory; CSI = Child Symptom Inventory; SWAN = Strengths and Weaknesses of ADHD-Symptoms and Normal-Behaviors; DOF = Direct Observation Form; T = teacher; CTRS:S = Conners' Teacher Rating Scale—Short Form; Conners-3 = Conners 3rd Edition; MAI = Munich Observation of Attention Inventory; IRR = incidence rate ratio; AD = anxiety disorder; TD = tic disorder; *DSM-III* = *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed.); *DSM-III-R* = *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed., rev.); ATRS = Abbreviated Teacher Rating Scale; COCADD = Classroom Observations for Conduct and Attention Deficit Disorder; LD = language delay; ADHD-RS = Abbreviated ADHD Rating Scale; *DSM-IV-TR* = *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text rev.); TTI = Teacher Telephone Interview.

[a]Correlations reported in this article were standardized associations of the observations as compared with peer measures. For our meta-analysis, we used raw scores of target children.

Table 4. Overview of Meta-Analytic Results of Correlations Between ADHD Reports on Teacher Rating Scales and Structured Classroom Observations.

		Meta-analytic effect size					Heterogeneity		
Rating scale	Structured observation	N	#Studies	r	95% CI	p	Q	I^2	p
Inattention	Inattention	1,807	12	.211	[.105, .312]	<.001	49.87	77.94	<.001
	Hyperactivity–impulsivity	1,629	10	.120	[−.012, .248]	.075	55.13	83.67	<.001
	Total	1,520	8	.164	[.008, .312]	.039	57.33	87.79	<.001
Hyperactivity–impulsivity	Inattention	1,807	12	.252	[.160, .340]	<.001	38.31	71.29	<.001
	Hyperactivity–impulsivity	1,629	10	.294	[.198, .384]	<.001	31.62	71.54	<.001
	Total	1,520	8	.304	[.219, .384]	<.001	18.33	61.81	.011
Total	Inattention	1,446	14	.305	[.184, .416]	<.001	57.47	79.12	<.001
	Hyperactivity–impulsivity	1,201	9	.247	[.133, .354]	<.001	25.80	69.00	.001
	Total	1,095	6	.261	[.121, .391]	<.001	23.46	78.68	<.001

Note. CI = confidence interval.

Meta-analytic results of the correlation between teacher ratings of hyperactivity–impulsivity and observed inattention (n = 12) showed a significant, small to moderate correlation (r = .252, p < .001). Against expectations, this correlation was not significantly weaker than the aggregated association between hyperactivity–impulsivity assessed by both measures (convergent validity) (r = .294; z = 1.33, p = .184). Heterogeneity was moderate to high (I^2 = 71.29).

Meta-regression. Testing mean age of the sample at baseline and percentage of boys on their effect on our meta-analytic results using meta-regression ($n \geq 10$ for these factors) did not significantly influence our results. Sample sizes were too small for meta-regression analyses on medication use and comorbidity.

Sensitivity analyses. Sensitivity analyses were conducted to investigate whether associations between rating scale scores and observational measures were dependent on the type of instrument used and the type of sample. The available data allowed these sensitivity analyses to be carried out on ($n \geq 3$) (a) studies using the Behavioral Observation of Students in Schools (BOSS; Shapiro, 2004) as compared with the full sample of studies, (b) studies conducting observations on a single school day (n = 5) versus multiple school-days (n = 6), as well as (c) studies using samples including only children with ADHD symptoms as compared with the full sample of studies.

For teacher ratings of inattention and inattention assessed with the BOSS, the meta-analytic correlation between subscales of the two instruments was significant (r = .207, p < .001). Meta-analytic correlation for the BOSS did not differ significantly from the meta-analytic correlation including the full sample of studies (z = −0.09, p = .928). Heterogeneity was low (I^2 = 30.66). The meta-analytic correlation between the hyperactivity–impulsivity subscale of rating scales and (aggregated) subscale of the BOSS assessing hyperactivity and impulsivity (r = .330, p = .001) was significant and did not significantly differ from the full sample of studies (z = 0.71, p = .478). Heterogeneity was moderate to high (I^2 = 69.42).

For all outcomes, correlations between rating scale scores and observations conducted on a single day were lower (r = .100, p = .184 for inattention; r = .296, p < .001 for hyperactivity–impulsivity; r = .232, p = .006 for total ADHD) compared to correlations between rating scale scores and observations conducted on multiple days (r = .334, p < .001 for inattention; r = .367, p < .001 for hyperactivity–impulsivity; r = .385, p < .001 for total ADHD), with the difference in correlation being significant for inattention (z = −4.92, p < .001).

Sensitivity analyses in ADHD-only samples revealed similar results compared with the full sample of studies for convergent validity outcomes: r = .165, p = .004 for inattention; r = .320, p < .001 for hyperactivity–impulsivity; and r = .232, p = .006 for total ADHD. Heterogeneity remained high: I^2 = 75.16, 75.57, and 80.04, respectively.

Publication bias. Due to the small number of studies available, it was possible to perform publication bias analyses only for the associations with structured observations. There was no evidence for publication bias neither for teacher ratings of inattention, nor for the ratings of hyperactivity–impulsivity. More specifically, for inattention, the funnel plot was symmetric (Egger's Ps = .12) and the fail-safe N value was 166. Regarding hyperactivity–impulsivity, the Egger funnel plot asymmetry was not significant (Ps = .99) and the fail-safe N value was 307. Funnel plots are presented in Supplementary File S5.

Study quality. Results showed that study quality was maximal for all interview studies (M = 4.00, SD = 0.00) and ranged between 1 and 4 for observational studies (M = 2.72, *SD* = 1.07). Study quality was not related to effect sizes for observational studies (range $r = -.051$ to .572). Results are presented in Supplementary File S4.

Discussion

This systematic review and meta-analysis examined convergent and divergent validity of teacher rating scales to assess ADHD symptoms, against two gold standard methods: semi-structured clinical teacher interviews and structured classroom observations. Data of 3,947 participants derived from 22 peer-reviewed articles were aggregated. Studies regarding clinical interviews were limited, but results support convergent validity of teacher rating scales when validated against semi-structured clinical interview, with strong correlations for all (sub)scales: inattention, hyperactivity–impulsivity, and total ADHD. Also divergent validity was confirmed for rating scale measures validated against semi-structured clinical interview: Meta-analytic convergent correlations were significantly larger than the divergent correlations. Validated against structured observations, convergent validity of rating scales was further confirmed, although correlations with teacher rating scales were only small to moderate. Divergent validity was supported only for the inattention symptom domain. Finally, as expected, overall, independent of the type of instrument, convergent validity was larger for ratings of hyperactivity–impulsivity than for ratings of inattention.

Whereas studies discouraged the sole use of rating scales because of biases in ratings of teachers (Burns et al., 2003; DuPaul, 2003; Gomez et al., 2003), our results indicate that teacher ratings of a child's ADHD symptoms may seem largely in line with the clinician's ratings on a semi-structured clinical interview, with a large percentage of shared variance (41% for hyperactivity–impulsivity and 30% for inattention). Although teacher rating scales, unlike semi-structured clinical interviews, do not take functional impairment and behavior in relation to the context into account, they appear valid, and useful, to assess symptoms of ADHD in the classroom. Given the high sensitivity and specificity of small-band rating scales, such as the CTRS-R (Conners, 1997) as compared with more broadband screeners such as the Teacher's Report Form (TRF; American Academy of Pediatrics, 2000), especially the use of small-band rating scales is recommended here. Our findings provide useful clinical information, as teacher ratings are time efficient and involve relatively low teacher burden, and are used frequently in clinical practice. However, for the diagnostic assessment of ADHD, multiple informants, such as parents, and assessment by a clinician are necessary (American Academy of Pediatrics, 2000; American Psychiatric Association, 2013; National Institute for Health and Care Excellence, 2018; Pliszka, 2007).

Results regarding the validity of rating scales against structured observations show that convergent validity is supported for ratings of hyperactivity–impulsivity (shared variance 9%) and to a lesser extent for inattention (weak convergent validity, shared variance is 4%). Our results correspond to the findings reported by Minder et al. (2017) that showed that the convergent validity of teacher ratings compared with structured classroom observations is limited, although in that study only total ADHD symptom scales were taken into account. This study extends these findings by showing lower validity for ratings of inattention compared with ratings of hyperactivity–impulsivity, although divergent validity is confirmed for inattention. Low convergent validity for inattention, however, suggests that teachers report on different behaviors using rating scales compared with a more objective measure of this behavior (i.e., on-task behavior) using structured observational measures. This is in line with the finding that teachers experience difficulties in reporting and observing symptoms of inattention (Poznanski, Hart, & Cramer, 2018). Evidence suggests that systematic behavioral observations are better in determining low frequent or covert behavior (e.g., attention shifts) than rating scales (Fabiano et al., 2004), and therefore classroom observations may provide unique information over and above rating scales, specifically regarding moments of inattention. However, structured observations usually describe on-task behavior as paying visual attention to a task, whereas rating scale items also contain *DSM*-based items such as paying close attention to details or making careless mistakes, the latter being difficult to code for an independent observer. However, divergent validity of rating scales compared with classroom observations was only confirmed for inattention scales. Inattentive and hyperactive–impulsive behaviors are highly related to each other in children with ADHD (for teacher ratings: $r = .67$; Wilcutt et al., 2012). For example, the hyperactivity *DSM* symptom "often unable to play or engage in leisure activities quietly" also requires the ability to stay focused to this activity. As a result, hyperactive behavior as measured by classroom observations is often seen as off-task (although there are some observational systems that take these differences into account by differentiating between passive and active on-task, Shapiro, 2004), which may account for correlations between inattentive and hyperactive–impulsive behaviors as assessed with rating scales and observations, and difficulties in distinguishing the two symptom dimensions. This raises fundamental questions about whether the symptom dimensions of ADHD can be judged fully in isolation. Future studies should examine what unique information may be provided by rating scales as well as clinical interviews and structured observations. All results for validity of teacher rating scales compared with structured observations showed significant heterogeneity. Sensitivity analysis for validity scores with only BOSS observational scores showed similar levels of associations compared with analyses including all studies, with lower levels of heterogeneity, suggesting that part of the full-sample heterogeneity can be explained by variability in the type of measures to assess ADHD. Furthermore, sensitivity analyses revealed that when the observational measure consisted of multiple days of assessments, scores corresponded more closely with teacher ratings. This finding may thus explain heterogeneity in our results and confirms the validity of teacher ratings, given that these ratings cover multiple days. In addition, study quality was not related to the magnitude of observed effect sizes. Moreover, the type of sample did not affect correlations or heterogeneity in the observed meta-analytic findings, indicating that observed correlations are stable among different samples. Future studies may look at other factors such as the type of rating scale and comorbidity to address heterogeneity, because our sample size did not allow for such further analysis.

Clearly, our finding that associations between rating scales and semi-structured clinical interview were stronger than those between rating scales and structured observational measures is evident, given that the teacher is involved as informant for both rating scales and clinical interviews. Furthermore, the coder of behavior usually involves a trained observant that has no relationship with the child and is potentially less biased regarding diagnostic status or personal relations (Burns et al., 2003; DuPaul, 2003; Gomez et al., 2003). Finally, the timeframe in which behavior is assessed is comparable for rating scales and semi-structured clinical interviews, usually being about the last week or last month(s) (i.e., ADHD Rating Scale-IV: School Version [ADHD-RS-IV], DuPaul, Power, Anastopoulos, & Reid, 1998a; Disruptive Behavior Disorders Rating Scale [DBDRS], Pelham et al., 1992; SWAN, Swanson et al., 2012; TTI, Tannock et al., 2002), whereas observations assess behavior over a shorter period of time (15–60 min) (i.e., Classroom Observation Code [COC], Abikoff & Gittelman, 1985; Ghent University Classroom Coding Inventory [GUCCI], Imeraj et al., 2013). Moreover, according to the meta-analysis by Achenbach et al. (1987) into cross-informant correlations of measures of behavioral and emotional problems in children, correlations between different informants do not exceed a moderate to strong correlation (r = .42 for teacher ratings and structured observations of behavioral and emotional problems). Clearly, factors such as situational specificity of behavior and diversity in informants (Burns et al., 2003; De Los Reyes & Kazdin, 2005; Gomez et al., 2003) are associated with lower correlations, limiting the maximum expected correlation between multiple informants. Our results are remarkably strong in this light, especially for the association between teacher ratings and clinical interview although the clinician's rating is based on information of the teacher, taking context, peers, intellectual abilities of the child, and impairment into account.

Despite clear strengths, our study has limitations. First, there are (unfortunately) few studies available to address the important issue of validity of teacher ratings as a source of information on ADHD behavior as compared with semi-structured clinical interviews (i.e., four studies, all using the same semi-structured clinical interview). Despite this small number of studies, convergent and divergent validity were confirmed, with only high-quality studies, highlighting the strength of this result. However, this also indicates a current lack of studies addressing such an important topic. Although we believe that this meta-analysis provides important information regarding the validity of teacher rating scales, clearly more studies are needed to demonstrate whether teacher rating scales can be seen as "another gold standard" assessment method, like teacher interviews (Pelham et al., 2005; Pliszka, 2007; Taylor & Sonuga-Barke, 2008; Volpe et al., 2005). Furthermore, the available studies assessing ADHD symptoms by semi-structured clinical interviews all included the same interview (TTI; Tannock et al., 2002), in referred samples. Results could therefore not be generalized to other samples, and the development of another teacher interview may be worth considering. Second, a larger number of included studies would have allowed us to perform meta-regression analyses to study the impact of the type of sample and of specific instrument characteristics. Although meta-regression analyses performed so far showed that neither age nor sex moderated the effects, these studies could not control for the likelihood of aggregation bias for percentage of males as a moderator (Higgins & Thompson, 2002). Individual participant data meta-analysis (IPDMA) would be a powerful method to examine the effect of sex on the relations between measures more accurately.

To summarize, this systematic review and meta-analysis shows the validity and thus utility of teacher rating scales as validated against clinical interviews, being an easy-to-administer (and relatively cheap) method for the assessment of ADHD symptoms in the classroom, with stronger associations for hyperactivity–impulsivity than for inattention. However, the number of studies investigating the validity of teacher rating scales against clinical interviews was small and more studies into its psychometric properties are clearly needed to further confirm validity. Moreover, our findings suggest that symptom domains of ADHD may not be judged in full isolation and that rating scales (particularly for inattention) measure different aspects of behavior than structured observations. This suggests that observations could add information over and above rating scales, by assessing specific or detailed (on-task) behavior of an individual child. The stronger correlations for teacher rating scales and observations conducted on multiple occasions compared with observations on one day only provides evidence for the validity of teacher rating scales, given that these ratings cover multiple days. Furthermore, future research is needed to gain more insight into the predictive validity of rating scales with the aim (a) to identify children at risk for ADHD diagnosis or (b) to predict potential escalation of (problem) behavior (i.e., by looking at later outcomes such as school dropout or referral to mental health services, Wentzel, 1993). In addition, it is of importance to investigate whether (subscales of) rating scales could predict the response to different treatment options and whether impairment rating scales filled out by teachers are related to actual functional impairments (e.g., academic achievement and performance as well as social impairment). The finding that rating scales can be considered valid as the assessment of ADHD problem behavior at school is an important clinical message, as teacher ratings are frequently used in clinical practice and studies into its validity were scarce so far.

Acknowledgments

The authors thank Alexandra Brewis, Alice Charach, Penny Corkum, Renate Drechsler, George DuPaul, Jeffery Epstein, Kenneth Gadow, Dione Healey, Bernd Heubeck, Lindita Imeraj, Stephanie McConaughy, William Pelham Jr., Linda Pfiffner, Jim Stevenson, Rosemary Tannock, and Betty Veenman for providing them with their data and their valuable contribution to the meta-analysis.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by The Netherlands Organization for Health Research and Development (ZonMw; Grant No. 729300013).

ORCID iDs

Anouck I. Staff https://orcid.org/0000-0002-2333-7189
Ralph de Vries https://orcid.org/0000-0002-2075-7495

Supplemental Material

Supplemental material for this article is available online.

References

Studies included in this meta-analysis are marked with an *.

Abikoff, H., & Gittelman, R. (1985). Classroom observation code: A modification of the Stony Brook Code. *Psychopharmacological Bulletin*, *21*, 901–909.

Abikoff, H., Jensen, P. S., Arnold, L. L. E., Hoza, B., Hechtman, L., Pollack, S., Martin, D., Alvir, J., March, J. S., Hinshaw, S., Vitiello, B., Newcorn, J., Greiner, A., Cantwell, D. P., Conners, C. K., Elliott, G., Greenhill, L. L., Kraemer, H., Pelham, W. E. Jr., Severe, J. B., Swanson, J. M., Wells, K., & Wigal, T. (2002). Observed classroom behavior of children with ADHD: Relationship to gender and comorbidity. *Journal of Abnormal Child Psychology*, *30*, 349–359. https://doi.org/10.1023/a:1015713807297

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

American Academy of Pediatrics. (2000). Clinical practice guideline: Diagnosis and evaluation of the child with attention-deficit/hyperactivity disorder. *Pediatrics*, *105*, 1158–1170. https://doi.org/10.1542/peds.105.5.1158

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.).

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

Atkins, M. S., Pelham, W. E., & Licht, M. H. (1989). The differential validity of teacher ratings of inattention/overactivity and aggression. *Journal of Abnormal Child Psychology*, *17*, 423–435. https://doi.org/10.1007/BF00915036

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2005). *Comprehensive meta-analysis* (Version 2). Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley.

*Brewis, A. (2002). Social and biological measures of hyperactivity and inattention: Are they describing similar underlying constructs of child behavior? *Social Biology*, *49*, 99–115. https://doi.org/10.1080/19485565.2002.9989052

Burns, G. L., Walsh, J. A., Gomez, R., & de Moura, M. A. (2003). Understanding source effects in ADHD rating scales: Reply to DuPaul (2003). *Psychological Assessment*, *15*, 118–119. https://doi.org/10.1037/1040-3590.15.1.118

*Charach, A., Chen, S., Hogg-Johnson, S., & Schachar, R. J. (2009). Using the Conners' Teacher Rating Scale-Revised in school children referred for assessment. *The Canadian Journal of Psychiatry*, *54*(4), 232–241. https://doi.org/10.1177/070674370905400404

Cochrane Collaboration. (2011). *Cochrane Handbook for Systematic Reviews of Interventions* (Version 5.1.0). Cochrane Collaboration.

Cohen, J. (1988). *Statistical power analysis for the behavioral science* (2nd ed.). Lawrence Erlbaum.

Conners, C. K. (1997). *Conners' Rating Scales-Revised: Technical manual*. Multi-Health Systems.

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

*DuPaul, G. J. (1991). Parent and teacher ratings of ADHD symptoms: Psychometric properties in a community-based sample. *Journal of Clinical Child Psychology*, *20*, 245–253. https://doi.org/10.1207/s15374424jccp2003_3

DuPaul, G. J. (2003). Assessment of ADHD symptoms: Comment on Gomez et al. (2003). *Psychological Assessment*, *15*, 115–117. https://doi.org/10.1037/1040-3590.15.1.115

*DuPaul, G. J., Jitendra, A. K., Tresco, K. E., Junod, R. E. V., Volpe, R. J., & Lutz, J. G. (2006). Children with attention deficit hyperactivity disorder: Are there gender differences in school functioning? *School Psychology Review*, *35*(2), 292–308.

DuPaul, G. J., Power, T., Anastopoulos, A. D., & Reid, R. (1998a). *Manual for the ADHD Rating Scale-IV*. Guilford Press.

*DuPaul, G. J., Power, T. J., McGoey, K. E., Ikeda, M. J., & Anastopoulos, A. D. (1998b). Reliability and validity of the parent and teacher ratings of attention-deficit/hyperactivity disorder symptoms. *Journal of Psychoeducational Assessment*, *16*, 55–68. https://doi.org/10.1177/073428299801600104

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple graphical test. *British Medical Journal*, *315*, 629–634. https://doi.org/10.1136/bmj.315.7109.629

*Epstein, J. N., Willoughby, M., Valencia, E. Y., Toney, S. T., Arnold, L. E., & Hinshaw, S. P. (2005). The role of children's ethnicity in the relationship between teacher ratings of attention-deficit/hyperactivity disorder and observed classroom behavior. *Journal of Consulting and Clinical Psychology*, *73*(3), 424–434. https://doi.org/10.1037/0022-006X.73.3.424

Fabiano, G. A., Pelham, W. E., Manos, M. J., Gnagy, E. M., Chronis, A. M., Onyango, A. N., Lopez-Williams, A., Burrows-MacLean, L., Coles, E. K., Meichenbaum, D. L., Caserta, D. A., & Swain, S. (2004). An evaluation of three time-out procedures for children with attention-deficit/hyperactivity disorder. *Behavior Therapy*, *35*, 449–469. https://doi.org/10.1016/S0005-7894(04)80027-3

Gomez, R., Burns, G. L., Walsh, J. A., & de Moura, M. A. (2003). A multitrait-multisource confirmatory factor analytic approach to the construct validity of ADHD rating scales. *Psychological Assessment*, *15*, 3–16. https://doi.org/10.1037/1040-3590.15.1.3

Groenewald, C., Emond, A., & Sayal, K. (2009). Recognition and referral of girls with attention deficit hyperactivity disorder: Case vignette study. *Child: Care, Health and Development*, *35*(6), 767–772. https://doi.org/10.1111/j.1365-2214.2009.00984.x

Halldner, L., Tillander, A., Lundholm, C., Boman, M., Langström, N., Larsson, H., & Lichtenstein, P. (2014). Relative immaturity and ADHD: Findings from nationwide registers, parent- and self-reports. *Journal of Clinical Psychology and Psychiatry*, *55*(8), 987–904. https://doi.org/10.1111/jcpp.12229

Handler, M. W., & DuPaul, G. J. (2005). Assessment of ADHD: Differences across psychology specialty areas. *Journal of Attention Disorders*, *9*, 402–412. https://doi.org/10.1177/1087054705278762

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539-1558. https://doi.org/10.1002/sim.1186

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*, 557–560. https://doi.org/10.1136/bmj.327.7414.557

Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review*, *34*, 507–519.

IBM Corp. (2016). *IBM SPSS Statistics for Windows*.

*Imeraj, L., Antrop, I., Roeyers, H., Deboutte, D., Deschepper, E., Bal, S., & Sonuga-Barke, E. J. S. (2016). The impact of idle time in the classroom: Differential effects on children with ADHD. *Journal of Attention Disorders*, *20*(1), 71–81. https://doi.org/10.1016/j.jsp.2013.05.004

Imeraj, L., Antrop, I., Sonuga-Barke, E. J. S., Deboutte, D., Deschepper, E., Bal, S., & Roeyers, H. (2013). The impact of instructional context on classroom on-task behavior: A matched comparison of children with ADHD and non-ADHD classmates. *Journal of School Psychology*, *51*, 587–498.

*Jiang, Y., Capriotti, M., Beaulieu, A., Rooney, M., McBurnett, K., & Pfiffner, L. J. (2019). Contribution of the Behavioral Observation of Students in Schools to ADHD assessment. *School Mental Health*, *11*, 464–475. https://doi.org/10.1007%2Fs12310-019-09313-5

Joanna Briggs Institute. (2017). *Critical appraisal tools for use in JBI systematic reviews: Checklist for Diagnostic Test Accuracy Studies*. The University of Adelaide.

*Johnson, K. A., White, M., Sum Wong, P., & Murrihy, C. (2020). Aspects of attention and inhibitory control are associated with on-task classroom behavior and behavioral assessments, by both teachers and parents, in children with high and low symptoms of ADHD. *Child Neuropsychology*, *26*, 219–241. https://doi.org/10.1080/09297049.2019.1639654

Junod, R. E. V., DuPaul, G. J., Jitendra, A. K., Volpe, R. J., & Cleary, K. S. (2006). Classroom observations of students with and without ADHD: Differences across types of engagement. *Journal of School Psychology*, *44*, 88–104. https://doi.org/10.1016/j.jsp.2005.12.004

*Kennerley, S., Jaquiery, B., Hatch, B., Healey, M., Wheeler, B. J., & Healey, D. (2018). Informant discrepancies in the assessment of attention-deficit/hyperactivity disorder. *Journal of Psychoeducational Assessment*, *36*(2), 136–147. https://doi.org/10.1177/0734282916670797

Krabbe, E. E., Thoutenhoofd, E. D., Conradi, M., Pijl, S. J., & Batstra, L. (2014). Birth month as predictor of ADHD medication use in Dutch school classes. *European Journal of Special Needs Education*, *29*(4), 571–578. https://doi.org/10.1080/08856257.2014.943564

*Lauth, G. W., Heubeck, B. G., & Mackowiak, K. (2006). Observation of children with attention-deficit hyperactivity (ADHD) problems in three natural classroom contexts. *British Journal of Educational Psychology*, *76*, 385–404. https://doi.org/10.1348/00709905X43797

*McConaughy, S. H., Harder, V. S., Antshel, K. M., Gordon, M., Eiraldi, R., & Dumenci, L. (2010). Incremental validity of test session and classroom observations in a multimethod assessment of attention hyperactivity disorder. *Journal of Clinical Child and Adolescent Psychology*, *39*(5), 650–666. https://doi.org/10.1080/15374416.2010.501287

McGonnell, M., Corkum, P., McKinnon, M., MacPherson, M., Williams, T., Davidson, C., & Stephenson, D. (2009). Doing it right: An interdisciplinary model for the diagnosis of ADHD. *Journal of the Canadian Academy of Child Adolescent Psychiatry*, *18*, 283–286.

Meyer, B. J., Stevenson, J., & Sonuga-Barke, E. J. S. (2020). Sex differences in the meaning of parent and teacher ratings of ADHD behaviors: An observational study. *Journal of Attention Disorders*, *24*, 1847–1856. https://doi.org/10.1177/1087054717723988

Milich, R., & Landau, S. (1988). Teacher ratings of inattention/overactivity and aggression: Cross-validation with classroom observations. *Journal of Clinical Child Psychology*, *17*, 92–97. https://doi.org/10.1207/s15374424jccp1701_12

Minder, F., Zuberer, A., Brandeis, D., & Drechsler, R. (2017). A review of the clinical utility of systematic behavioral observations in attention deficit hyperactivity disorder (ADHD). *Child Psychiatry & Human Development*, *49*, 572–606. https://doi.org/10.1007/s10578-017-0776-2

*Minder, F., Zuberer, A., Brandeis, D., & Drechsler, R. (2018). Informant-related effects of neurofeedback and cognitive training in children with ADHD including a waiting control phase: A randomized-controlled trial. *European Child & Adolescent Psychiatry*, *27*, 1055–1066. https://doi.org/10.1007/s00787-018-1116-1

National Institute for Health and Care Excellence. (2018). *Attention deficit hyperactivity disorder: Diagnosis and management* [NICE Guideline NG87].

*Nolan, E. E., & Gadow, K. D. (1994). Relation between ratings and observations of stimulant drug response in hyperactive children. *Journal of Clinical Child Psychology*, *23*(1), 78–90. https://doi.org/10.1207/s15374424jccp2301_10

*Parker, A., & Corkum, P. (2016). ADHD diagnosis: As simple as administering a questionnaire or a complex diagnostic process? *Journal of Attention Disorders*, *20*(6), 478–486. https://doi.org/10.1177/1087054713495736

*Pelham, W. E., Aronoff, H. R., Midlam, J. K., Shapiro, C. J., Gnagy, E. M., Chronis, A. M., Onyango, A. N., Forehand, G., Nguyen, A., & Waxmonsky, J. (1999). A comparison of ritalin and adderall: Efficacy and time-course in children with attention-deficit/hyperactivity disorder. *Pediatrics*, *103*, 1–14. https://doi.org/10.1542/peds.103.4e43

*Pelham, W. E., Carlson, C., Sams, S. E., Vallano, G., Dixon, J., & Hoza, B. (1993). Separate and combined effects of methylphenidate and behavior modification on boys with attention deficit-hyperactivity disorder in the classroom. *Journal of Consulting and Clinical Psychology*, *61*, 506–515. https://doi.org/10.1037/0022-006X.61.3.506

Pelham, W. E., Fabiano, G. A., & Massetti, G. M. (2005). Evidence-based assessment of attention deficit hyperactivity disorder in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *34*, 449–476. https://doi.org/10.1207/s15374424jccp3403_5

Pelham, W. E., Gnagy, E. M., Greenslade, K. E., & Milich, R. (1992). Teacher ratings of DSM-III-R symptoms for the disruptive behavior disorders. *Journal of the American Academy of Child & Adolescent Psychiatry*, *31*, 210–218. https://doi.org/10.1097/00004583-199203000-00006

*Pfiffner, L. J., Villodas, M., Kaiser, N., Rooney, M., & McBurnett, K. (2013). Educational outcomes of a collaborative school-home behavioral intervention for ADHD. *School Psychology Quarterly*, *28*(1), 25–36. https://doi.org/10.1037/spq0000016

Pliszka, S. (2007). Practice parameter for the assessment and treatment of children and adolescents with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *46*, 894–921. https://doi.org/10.1097/chi.0b13e318054e724

Poznanski, B., Hart, K. C., & Cramer, E. (2018). Are teachers ready? Preservice knowledge of classroom management and ADHD. *School Menthal Health*, *10*(3), 301-313. https://doi.org./10.1007/s12310-018-9259-2

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, *118*, 183–192. https://doi.org/10.1037/0033-2909.118.2.183

Shapiro, E. S. (2004). *Academic skills problems workbook* (Rev. ed.). Guilford Press.

Shapiro, E. S., & Heick, P. (2004). School psychologist assessment practices in the evaluation of students referred for social/behavioral/emotional problems. *Psychology in the Schools*, *41*, 551–561. https://doi.org/10.1002/pits.10176

Snyder, S. M., Hall, J. R., Cornwell, S. L., & Quintana, H. (2006). Review of clinical validation of ADHD Behavior Rating Scales. *Psychological Reports*, *99*, 363–378. https://doi.org/10.2466/pr0.99.2.363-378

*Stevenson, J., Sonuga-Barke, E., McCann, D., Grimshaw, K., Parker, K. M., Rose-Zerilli, M. J., Holloway, J. W., & Warner, J. O. (2010). The role of histamine degradation gene polymorphisms in moderating the effects of food additives on children's ADHD symptoms. *American Journal of Psychiatry*, *167*(9), 1108–1115. https://doi.org/10.1176/appi.ajp.2010.09101529

Swanson, J. M., Schuck, S., Mann Porter, 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 SWAN rating scales. *The International Journal of Educational and Psychological Assessment*, *10*(1), 51–70.

Tannock, R., Hum, M., Masellis, M., Humphries, T., & Schachar, R. (2002). *Teacher Telephone Interview for children's academic performance, attention, behavior, and learning*. Toronto, Canada: The Hospital for Sick Children.

Taylor, R., & Sonuga-Barke, E. (2008). Disorders of attention and activity. In M. Rutter, D. V. M. Bishop, D. S. Pine, S. Scortt, J. Stevenson, E. Taylor, & A. Thapar (Eds.), *Rutter's child and adolescent psychiatry* (5th ed., p. 521-542). Blackwell.

Valo, S., & Tannock, R. (2010). Diagnostic instability of DSM-IV ADHD subtypes: Effects of informant source, instrumentation, and methods for combining symptom reports. *Journal of Clinical Child and Adolescent Psychology*, *39*(6), 746–760. https://doi.org/10.1080/15374416.2010.517172

*Veenman, B. Y., Luman, M., Hoeksma, J., Pieterse, K., & Oosterlaan, J. (2019). A randomized effectiveness trial of a behavioral teacher program targeting ADHD symptoms. *Journal of Attention Disorders*, *23*, 293–304. https://doi.org/10.1177/1087054716658124

*Veenman, B. Y., Luman, M., & Oosterlaan, J. (2017). Further insight into the effectiveness of a behavioral teacher program targeting ADHD symptoms using actigraphy, classroom observations and peer ratings. *Frontiers in Psychology*, *8*, 1–10. https://doi.org/10.3389/fpsyg.2017.01157

Volpe, R. J., DiPerna, J. C., Hintze, J. M., & Shapiro, E. S. (2005). Observing students in classroom settings: A review of seven coding schemes. *School Psychology Review*, *34*(4), 454–474.

Wentzel, K. R. (1993). Does being good make the grade? Social behavior and academic competence in middle school. *Journal of Educational Psychology*, *85*(2), 357.

Whalen, C. K., Collins, B. E., Henker, B., Alkus, R., Adams, D., & Stapp, J. (1978). Behavior observations of hyperactive children and methylphenidate (Ritalin) effects in systematically structured classroom environments: Now you see them, now you don't. *Journal of Pediatric Psychology*, *3*, 177–187. https://doi.org/10.1093/jpepsy/3.4.117

Whalen, C. K., Henker, B., Collins, B. E., Finck, D., & Dotemoto, S. (1979). A social ecology of hyperactive boys: Medication effects in structured classroom environments. *Journal of Applied Behavior Analysis*, *12*, 65–81. https://doi.org/10.1901/jaba.1979.12-65

Wilcutt, 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

Wilson, M. S., & Reschly, D. J. (1996). Assessment in school psychology training and practice. *School Psychology Review*, *25*, 9–23.

Author Biographies

Anouck I. Staff, MSc, is a PhD candidate in clinical neuropsychology at Vrije Universiteit Amsterdam.

Jaap Oosterlaan, PhD, is a full professor in clinical neuropsychology at Vrije Universiteit Amsterdam, and is affiliated with Emma Children's Hospital, Amsterdam UMC, University of Amsterdam, Emma Neuroscience Group, Department of Pediatrics, Amsterdam Reproduction & Development.

Saskia van der Oord, PhD, is an associate professor in Clinical Psychology at KU Leuven and in Developmental Psychology, University of Amsterdam. She is also a behavior therapist.

Pieter J. Hoekstra, MD, PhD, is a full professor in the Department of Child and Adolescent Psychiatry, University of Groningen, University Medical Center Groningen, and is also a psychiatrist.

Karen Vertessen, MD, MSc, is a PhD candidate in clinical neuropsychology at Vrije Universiteit Amsterdam.

Ralph de Vries, MSc, is a medical librarian in the Medical Library at Amsterdam UMC, VU Medical Center.

Barbara J. van den Hoofdakker, PhD, is a full professor in the Department of Clinical Psychology and Experimental Psychopathology, University of Groningen, and in the Department of Child and Adolescent Psychiatry, University of Groningen, University Medical Center Groningen. She is also a clinical psychologist and a behavior therapist.

Marjolein Luman, PhD, is an associate professor in clinical neuropsychology at Vrije Universiteit Amsterdam.