Child Psychiatry & Human Development (2018) 49:572–606
https://doi.org/10.1007/s10578-017-0776-2

ARTICLE

A Review of the Clinical Utility of Systematic Behavioral Observations in Attention Deficit Hyperactivity Disorder (ADHD)

Franziska Minder[1] · Agnieszka Zuberer[1] · Daniel Brandeis[1,2,3,4] · Renate Drechsler[1]

Published online: 6 December 2017

Abstract

This review evaluates the clinical utility of tools for systematic behavioral observation in different settings for children and adolescents with ADHD. A comprehensive search yielded 135 relevant results since 1990. Observations from naturalistic settings were grouped into observations of classroom behavior ($n=58$) and of social interactions ($n=25$). Laboratory observations were subdivided into four contexts: independent play ($n=9$), test session ($n=27$), parent interaction ($n=11$), and peer interaction ($n=5$). Clinically relevant aspects of reliability and validity of employed instruments are reviewed. The results confirm the usefulness of systematic observations. However, no procedure can be recommended as a stand-alone diagnostic method. Psychometric properties are often unsatisfactory, which reduces the validity of observational methods, particularly for measuring treatment outcome. Further efforts are needed to improve the specificity of observational methods with regard to the discrimination of comorbidities and other disorders.

Keywords ADHD · Systematic behavioral observation · Psychometric properties · Classroom behavior · Test session behavior

Introduction

Attention deficit hyperactivity disorder (ADHD) is one of the most frequently diagnosed neurobehavioral disorders in childhood, with a prevalence rate of approximately 5% [1]. The chronic nature of the disorder and its long-term impact on the social and academic life of affected individuals substantiate the need for early identification and treatment. Behavioral rating scales and interviews with parents and teachers are the most frequently used diagnostic tools in the assessment of ADHD [2–4], as the diagnostic criteria are of a behavioral nature [2, 5]. Informant ratings offer an efficient summary of natural observations over extended time spans [6]. A symptom of inattention is met if a person "is often easily distracted" or "often does not seem to listen when spoken to" (DSM-5 [7]). However, terms such as "often" and "easily" are not specifically defined and can therefore be subject to rater bias. Misinterpretation of items, inaccurate recall of events [5], halo effects [8], unknown bases of comparison of the informant [9], factors affecting the informant (e.g., maternal depression [6]), and the socioeconomic status of the target subject [10] may influence the validity of rating scales. The issue of subjectivity is particularly detrimental if ratings are used as an indicator of treatment response in unblinded pre-post intervention designs. Alternative methods for ADHD assessment with greater objectivity are therefore highly desirable. While it is also common to conduct laboratory psychological testing as part of the diagnostic process [2], the consensus view on the validity of such instruments suggests that there is no cognitive litmus test for the diagnosis of ADHD [2, 11, 12]. Pelham et al. [2] emphasized the importance of evaluating observable behavior instead of cognitive test performance for ADHD assessment.

A survey of school psychologists in the U.S. revealed that direct observations are among the most commonly

✉ Renate Drechsler
renate.drechsler@kjpd.uzh.ch

[1] Department of Child and Adolescent Psychiatry and Psychotherapy, University Hospital of Psychiatry Zurich, University of Zurich, Zurich, Switzerland

[2] Neuroscience Center Zurich, University of Zurich and ETH Zurich, Zurich, Switzerland

[3] Central Institute of Mental Health, Medical Faculty Mannheim/Heidelberg University, Mannheim, Germany

[4] Center for Integrative Human Physiology, University of Zurich, Zurich, Switzerland

used methodologies in diagnostic processes [13]. Handler and DuPaul [4] consider the use of observations in combination with other assessment methods to be consistent with standards of best practice. For treatment evaluation, blinded observations were claimed to be the gold standard of assessment [14]. The use of qualitative approaches to observation, with anecdotal descriptions of behavior, is widespread among practitioners [15], but these approaches do not allow for psychometric testing [16]. In contrast, systematic direct observation methods are based on standardized scoring procedures that aim to quantify operationally defined, specific behaviors in an objective way, enabling inter-observer agreement to be assessed [17]. These methods employ either a rating scale developed for the purpose of direct observations or a standardized recording strategy. Two typical recording strategies can be distinguished: continuous recording and time sampling of behavior. Continuous recording includes event counting (frequency) and duration recording. In time sampling, a target behavior is coded if it occurs during the whole predefined interval (whole-interval time sampling), at any time within the interval (partial-interval time sampling), or at a fixed moment of time (momentary time sampling) [18]. In ADHD, behavioral categories for observation usually consist of a proxy for attentive and inattentive behavior (i.e., on- and off-task behavior). Additionally, visually detectable aspects of motor activity and indicators of social interactions such as disruptiveness, aggression, and noncompliance constitute common variables in observational approaches.

Systematic observations also represent a common method for diagnostics and treatment evaluation in autism spectrum disorder (ASD) [19]. These observations usually focus on variables describing social play and communication behaviors, challenging behaviors, and stereotypies (e.g., Autism Diagnostic Observation Schedule—ADOS [20]; Early Social Communication Scale—ESCS [21]). Children with conduct disorders (CD) are frequently observed during parent–child interactions (e.g., Parent–Child Interaction Task—PCIT [22]), peer interactions in the laboratory [23] and in the classroom (e.g., Multiple Option Observation System for Experimental Studies—MOOSES [24]). In general, observational methods are more commonly used for externalizing disorders than for internalizing disorders, owing to the overt nature of the behavioral problems. However, behavioral inhibition [25] or avoidance and fear (Anxiety Dimensional Observation Scale [26]) during mother–child interactions can be observational targets for anxiety disorders in children. Some comprehensive observational methods also include scales for internalizing problems (ASEBA-Direct Observation Form (DOF), Test Observation Form (TOF) [27, 28]).

Environments for observation can be roughly divided into *naturalistic* settings, such as the classroom, and *standardized* settings, such as the laboratory or clinic [5]. The high ecological validity of naturalistic settings [6] comes at the expense of uncontrollable contextual factors that might affect behavior. In standardized laboratory situations, by contrast, behavior is limited to a distinct given context, whereby the comparability of observed behavior between individuals and between multiple administrations is increased. Laboratory settings also allow for the application of less obtrusive observational methods through one-way mirrors or video cameras. Behavior in the laboratory may, however, be less generalizable due to the artificial nature of the situation.

To date, the psychometric properties and the diagnostic utility of standardized observations in ADHD have only been selectively delineated [2, 29]. A complete overview of their clinical validity in ADHD is lacking. Therefore, the purpose of this article is to comprehensively review the systematic observational instruments that have been used in studies on ADHD, published between 1990 and 2016. The employed tools were evaluated with respect to four clinically relevant issues:

(1) Basic reliability measures of the methods are reported, namely inter-rater reliability (IRR) and test–retest reliability (TRR) for samples of ADHD subjects.
(2) The predictive validity [2] of observations is discussed, i.e., to what extent such instruments can accurately distinguish between individuals with and without ADHD. The main emphasis is placed on reported classification rates, sensitivity and specificity (sensitivity refers to the ability of a measure to correctly identify cases, whereas specificity refers to the ability to correctly classify individuals without the problem in question).
(3) Findings on convergent validity of observational measures are evaluated, i.e., correlations between observational data and other measures of ADHD (mostly parent and teacher ratings).
(4) The evidence that behavioral observations detect treatment effects is reviewed.

For ease of reference, the numbering of these clinical issues will be retained and indicated accordingly in the corresponding sections of the review.

Method

With the intention to cover all relevant fields in ADHD research in which observational methods were applied, a search strategy ensuring wide coverage was implemented. Search terms included “ADHD or attention deficit hyperactivity disorder or ADD or attention deficit disorder” for subject field and “direct observ*” or “behavioral observ*” in any field (the asterisks served as wild cards). The initial search was extended by manual analysis of the reference lists of articles and by searches based on the names of observational instruments that were detected by the initial search (see Table 1 for overview of the instrument names). Inclusion criteria were publication in English in a peer-reviewed journal from 1990 to 2016 and the administration of a systematic observational instrument in the study of individuals with a diagnosis of ADHD or symptoms of ADHD. Studies with fewer than ten subjects or with adult subjects were excluded. Objective measures of activity by mechanical or infrared devices (for review see [30, 31]), and aspects of language and private speech (for review see [32]) of individuals with ADHD were not reported. Behavioral measures in choice-impulsivity tasks were not considered as a method of direct systematic observation (for review see [33]).

Results

The database search generated 685 peer-reviewed articles using PsycINFO, PsycARTICLES, and Medline, finalized on July 6, 2016. Ninety-seven abstracts from the database search fulfilled the inclusion criteria. The comprehensive search including retrievals from reference lists and additional searches based on names of observational instruments yielded 179 studies for review. Studies applying a standardized observation unattached to a specific instrument were excluded from the review ($n=56$). This resulted in 123 studies for review. Twelve studies comprised results of more than one observational tool (i.e., from different situations, e.g., classroom and playground). These were specified twice in tables with respect to the specific context. Hence, the tables contain 135 individual entries.

Eighty-two studies reported systematic observations of children with ADHD in naturalistic settings, i.e., low-structured situations with few standardization attempts through the study protocol. These were separated into two major sections: classroom observations ($n=58$) and observations of social interactions in natural contexts, e.g., group leisure activities or free play ($n=25$).

Situations that were clearly predefined and specified by the study (e.g., room, group size, materials, instruction) were considered as laboratory (even if the observation occurred at home or in a separate room at school). In 52 studies, behavioral observations of children or adolescents with ADHD were conducted in such standardized, non-naturalistic settings. Tables were generated for different observational contexts (e.g., classroom observation, independent play, test session behavior). Within the tables, studies were sorted by study type, i.e., group discrimination, convergent validity, pharmacological and non-pharmacological interventions, and by year of publication. A separate section at the end of the tables displayed the studies with adolescents ($n=13$).

The results narrative was structured according to the observational tools and the four research questions to be evaluated for each tool. In some cases, studies from before 1990 or non-ADHD studies were cited if no newer reports or no ADHD-specific studies were available to evaluate the respective issue. These studies were not listed in the tables.

Observation Studies of Children and Adolescents with ADHD in Naturalistic Settings

Systematic Classroom Observations

Fifty-four studies with classroom observations of children and four studies with observations of adolescents with ADHD conducted since 1990 were included in this review (Table 2). In 28 studies, the naturalistic concept was only partially applicable because behavior was investigated in a simulated school situation, i.e., a laboratory school or the classroom of a summer treatment program. In total, 11 different specific classroom observational instruments were applied in the study of ADHD.

Classroom Observation Code (COC) The COC is an early, well-established observational system [91], which was applied in seven of the reviewed studies in Table 2. It assesses 3–12 variables in the classroom (interference, off-task behavior, noncompliance, motor activity, aggression, etc.) and applies a partial-interval time-sampling recording method (15-s intervals). (1) High rates of IRR were documented ($\text{phi}=.80-1$, $\text{kappa}=.77–.94$) [36, 48, 67]. In a modified version of the COC, TRR for 32 children with ADHD was highly significant at an interval of 1 day ($r=.37–.72$), but low at an interval of 2 days ($r=.27–.49$) [48]. (2) According to the COC categories of off-task behavior, interference, motor activity, and solicitation, 80% of cases of ADHD were correctly classified (false positive error of 9.8%) (in the original study [91]). All COC categories were exhibited at a significantly higher rate by children with ADHD compared to their typically developing classmates in the large sample of the Multimodal Treatment Study (MTA) [36]. (3) Small to moderate significant correlations were reported between observed negativistic behaviors (interference, noncompliance, aggression) and the Inattention and Overactivity with Aggression (IOWA) Conners teacher ratings of aggression ($r=.37–.60$) [48]. A correlation coefficient of $r=.46$ was reported between classroom off-task behavior and the IOWA inattention scale [48]. The COC categories correlated modestly (all $r<.40$) with performance on neuropsychological tasks [49]. (4) Three studies [52, 56, 67] reported significant improvement on the COC with pharmacological treatment.

ADHD School Observation Code (ADHD-SOC) The ADHD-SOC [92] was developed on the basis of the COC [91]. It

Table 1 Summary of the reliability and validity information for 29 observational tools in ADHD

	N	IRR	TRR	Predictive validity	Convergent validity	Sensitivity to treatment
Classroom observations						
ADHD-SOC	1	Adequate	–	Sig. group differences, 91% sensitivity, 80% specificity	–	Sig. with medication
BOSS	9	Adequate	–	Sig. group differences, 71% correct classification	Small–moderate	Sig. with non-pharmacol. intervention
COC	7	Adequate	Low–moderate	Sig. group differences, 80% correct classification	Small–moderate	Sig. with medication
COCADD	9	Low–adequate	–	83% correct classification	Small–moderate	Sig. with medication
DOF	4	Adequate	Low–moderate	61–70% correct classification	Small–moderate	Sig. with non-pharmacol. intervention
GUCCI	2	Adequate	–	Sig. group differences	–	–
HBRS	2	Adequate	–	Sig. group differences	–	–
MAI	1	–	–	Sig. group differences	Moderate–high	–
RIPPS	1	Adequate	–	Sig. group differences	–	–
SOS	1	Adequate	–	Sig. group differences	–	–
SKAMP	21	Low–adequate	Moderate–high	–	Small	Sig. with medication
Naturalistic social interaction observations						
COCA-R	1	Adequate	–	–	Small–moderate	Sig. with non-pharmacol. intervention
COSA/ADHD-SOC	4	Low–adequate	Low	Sig. group differences	Small–moderate	Sig. with medication
ESP	4	Adequate	–	Sig. group differences	Small	–
SRP-Obs.	6	Low–adequate	–	Sig. group differences	Small–moderate	Sig. with medication
ADRCS	10	Low–adequate	–	–	–	Sig. with medication and non-pharmacol. intervention
Independent play observations						
Attention/engagement	3	Adequate	Moderate–high	Sig. group differences	–	Mixed with non-pharmacol. intervention
SOAPS	6	Adequate	–	Sig. group differences, 58–64% correct classification	–	Sig. with medication
*Test session observation*s						
GATSB	1	–	–	81% correct classification	–	–
HBRS	1	Low–adequate	–	Sig. group differences	Small–moderate	–
RAS	23	Low–adequate	Moderate–high	Sig. group differences, 64–86% correct classification	Small–moderate	Sig. with medication and non-pharmacol. intervention
TOF	2	Adequate	Moderate–high	Sig. group differences, 74% correct classification	Small–moderate	–
*Parent–child observation*s						
CRS	2	Adequate	–	Sig. group differences	–	Mixed with non-pharmacol. intervention
DB-DOS	1	Adequate	Moderate–high	87% sensitivity, 79% specificity	Small–moderate	–

Table 1 (continued)

	N	IRR	TRR	Predictive validity	Convergent validity	Sensitivity to treatment
DPICS-R	1	Adequate	–	–	–	Sig. with non-pharmacol. intervention
GIPCI-R	2	Low–adequate	Low–moderate	–	–	Mixed with non-pharmacol. intervention
PAICS	3	Low–adequate	Low	Sig. group differences	–	Mixed with non-pharmacol. intervention
MTA-Obs.	2	Low–adequate	–	–	–	Mixed with medication and non-pharmacol. intervention
Peer–child observations						
ToP	5	–	–	–	–	Sig. with non-pharmacol. intervention

ADHD-SOC ADHD School Observation Code, *ADRCS* All-day response-cost system, *BOSS* Behavioral Observation of Students in Schools, *COC* Classroom Observation Code, *COCA-R* Coder Rating of Child Adaptation-Revised, *COCADD* Classroom Observations for Conduct and Attention Deficit Disorder, *COSA* Code for Observing Social Activity, *CRS* Conflict Rating Scale, *DB-DOS* Disruptive Behavior Diagnostic Observation Schedule, *DOF* Direct Observation Form, *DPICS-R* Dyadic Parent–Child Interaction Coding System-Revised, *ESP* Early Screening Project, *GATSB* Guide to Assessment of Test Session Behavior, *GIPCI-R* Global Impressions of Parent–Child Interaction-revised, *GUCCI* Ghent University Classrooom Coding Inventory, *HBRS* Hillside Behavior Rating Scale, *MAI* Munich Observation of Attention Inventory, *MTA-Obs*. Multimodal Treatment of ADHD study observational code, *PAICS* Parent and Adolescent Interaction Coding System, *RAS* Restricted Academic Situation, *RIPPS* Responses to Interpersonal and Physically Provoking Situations, *SKAMP* Swanson, Kotkin, Agler, M-Flynn, and Pelham impairment rating scale, *SOAPS* Structured Observations of Academic and Play Setting, *SOS* Student Observation System, *SRP-Obs*. Summer research program observational code, *TOF* Test Observation Form, *ToP* Test of Playfulness

was used in one study of Table 2 [57]. The ADHD-SOC assesses interference, motor movement, noncompliance, aggression, and off-task behavior in a 15-s partial-interval time-sampling procedure. (1) IRR was acceptable (kappa = .57–.84) [57]. TRR was not specifically evaluated for the ADHD-SOC. (2) All classroom observational categories of the ADHD-SOC were shown to discriminate children with ADHD and comorbid tic disorder from controls on the group level. A combination of off-task behavior, interference, and noncompliance yielded correct identification of 91% of the subjects, but also misclassification of 20% of peers [57]. (3) There are no reports on the convergent validity of the ADHD-SOC. (4) The ADHD-SOC was sensitive to stimulant drug effects, with observed normalized classroom behavior in approximately 75% of children with ADHD and tic disorder [57].

Behavioral Observation of Students in Schools (BOSS) The BOSS [93] separates on-task behavior into active (AET) and passive engagement (PET), while off-task behavior is divided into three subcategories: motor, verbal, and passive. Engagement is coded with a momentary time-sampling method (every 15 s), while off-task behavior is coded using a 15-s partial-interval time-sampling procedure. The BOSS was applied with different modifications in nine of the reviewed studies. (1) Adequate IRR was reached, with kappas ranging from .77 to .98 [37, 82]. The TRR was not investigated. Steiner et al. [85] noted a significant improvement in classroom off-task behavior over time in an untreated ADHD control group, which indicates some instability. A dependability study revealed two 30-min observations on two separate days, providing acceptable levels of dependability for progress monitoring purposes [94]. A third of the variance in BOSS on-task behavior within 30 min on 2 days was attributable to individual differences [94]. Single observations, even for a duration of 60 min, could not reach the same dependability as two 30-min observations on 2 days in the same academic subject [94]. (2) The rates of PET and off-task behavior significantly differentiated ADHD children from controls [37, 41, 47]. Based on a regression model, 71% of subjects were correctly classified into the groups of ADHD and peers by the BOSS categories of off-task behavior [41]. (3) BOSS off-task behavior was also a significant predictor of reading achievement in students with ADHD [37]. Inter-correlations between BOSS categories and the teacher AD/HD rating scale-IV were not significant (r = .02–.20) [37]. Hosterman et al. [10] reported moderate significant correlations between some BOSS categories and the teacher AD/HD rating scale-IV (r = .27–.40) and between some BOSS categories and the Conners Teacher Rating Scale (r = .25–.47). (4) Non-pharmacological intervention studies have shown significant behavioral improvement using the BOSS [82, 84, 85].

Table 2 Classroom observation studies of children and adolescents with ADHD ($n = 58$)

Author(s)	Age	*N*	Diagnosis	Control group	Intervention		Duration (min)	Observational instrument	Results
Group discrimination studies									
Lett and Kamphaus [34]	mean 7–8	55	ADHD ± comorbidity	TD	–	DO TS	15 RC	SOS (13 variables): four adaptive behaviors, nine problem behaviors (IRR: *r* = .69–1)	Inappropriate movement, problem behavior composite: ADHD > TD*, (SOS variables did not improve discriminant analyses)
Skansgaard and Burns [35]	5–6	24	ADHD subtypes	Classmates	–	DO TS RS	40 RC	DOF: narrative description, on-/off-task, rating on 96 problem items (IRR: *r* = .69 − 1)	Inattentive behavior: ADHD > classmates* Hyperactive/impulsive and ODD/CD behavior: ADHD-C > ADHD-IN/classmates*
Abikoff et al. [36]	7–10	502	ADHD ± DBD, ± ANX	Classmates	–	DO TS	16 RC	COC (12 variables): interference, off-task, motor standing, motor vigorous, physical aggression (IRR: phi = .80–1)	All observed measures: ADHD > classmates* Interference, aggression, motor movement: ADHD boys > ADHD girls*
DuPaul et al. [37]	6–10	189	ADHD	TD	–	DO TS	30 RC	Adapted BOSS (six variables): active, passive engagement, off-task motor, verbal, passive, noncompliance (IRR: κ = .93–.98)	Passive engagement: ADHD < TD* Off-task: ADHD > TD*
Antrop et al. [38]	6–11	28	ADHD	TD	–	VT RS	75 RC	Adapted HBRS (six variables): restlessness, noisiness, interaction, disturbing, frustration tolerance, search for extra stimulation (IRR: *r* = .70–.98)	Restlessness, noisiness, frustration, stimulation seeking, disturbance: ADHD > TD* (levels of activity increased during waiting)
Antrop et al. [39]	6–11	28	ADHD	Classmates	–	VT RS	60 RC	Adapted HBRS (five variables): out-of-seat, repetitive movement, off-task, noisiness, disturbing behavior (IRR: *r* = .92–.97)	Off-task, fidgeting, out-of-seat: ADHD > classmates* Noisiness increase in the afternoon: ADHD > classmates*
DuPaul et al. [40]	6–10	175	ADHD	None	–	DO TS	30 RC	Adapted BOSS (six variables): active, passive engagement, off-task motor, verbal, and passive, noncompliance (IRR: κ = .88–.98)	No sig. gender differences in observed behavior
Vile Junod et al. [41]	6–10	155	ADHD	TD, classmates	–	DO TS	30 RC	BOSS (five variables): active, passive engagement, off-task motor, verbal, and passive (IRR: κ = .93–.98)	Engagement: ADHD < TD, classmates* Off-task: ADHD > TD, classmates* Correct classification of 71% of the sample into ADHD and classmates, no significant correlations between BOSS and teacher ADHD rating (*r* = .02–.20)
Lauth et al. [42]	7–11	106	ADHD	Classmates	–	DO TS	12 lessons RC	MAI (five variables): inattentive off-task, disruptive off-task, inconspicuous on-task, self-initiated on-task, other-initiated on-task (IRR: not reported)	Off-task, self- and other-initiated on-task: ADHD > classmates* Inconspicuous on-task: ADHD < classmates* Moderate to strong correlations between MAI and teacher ratings (*r* = .31–.71)
McConaughy et al. [43]	6–11	163	ADHD subtypes	TD, clin. ref.	–	DO TS RS	40 RC	DOF: narrative description, on-/off-task, rating on 89 problem items (IRR: *r* = .70–.97)	Intrusive and oppositional scale, ADHD scale, hyperactivity/impulsivity scale, total problems: ADHD-C > TD/clin.ref.* On-task: ADHD-C/IN < TD* Sluggish cognitive tempo, attention problem scale, inattention scale, total problems: ADHD-IN > TD* Correct classification: 61–67% into ADHD-C versus TD/clin ref Correct classification: 70% into ADHD-IN versus TD
Hart et al. [44]	7–12	33	ADHD	None	–	DO TS	daily STP	Adapted COCADD (one variable): on-/off-task (IRR: κ = .42–.78)	On-task: small-group instruction > whole-group instruction and independent seatwork*
Imeraj et al. [45]	6–12	62	ADHD	Classmates	–	VT CS	1 day RC	GUCCI (four variables): activity, nosiness, disruptive behavior (IRR: κ = .74–.99)	Activity, disruptive behavior, noisiness: ADHD > classmates*, Noisiness/hyperactivity during idle time: ADHD > classmates*

Table 2 (continued)

Author(s)	Age	N	Diagnosis	Control group	Intervention		Duration (min)	Observational instrument	Results
Imeraj et al. [46]	6–12	62	ADHD	Classmates	–	VT CS	1 day RC	GUCCI (one variable): on-task (IRR: κ = .77–.84)	On-task during whole-group, individual work, academic task, transition time: ADHD < classmates*
Steiner et al. [47]	2nd/4th gr.	104	ADHD	Classmates	–	DO TS	45 RC	Adapted BOSS (three variables): engagement, off-task motor and verbal, inattentive behavior (IRR: mean κ = .86)	Engagement: ADHD < classmates* Off-task: ADHD > classmates*
Convergent validity									
Nolan and Gadow [48]	5–13	34	ADHD	None	–	DO TS	3–4 × 50 RC	Modified COC (five variables): interference, motor movement, noncompliance, aggression, off-task (IRR: κ = .77–.86;TRR: ICC = .33–.57, r = .27–.72)	Sig. correlations (r = .37–.60) between COC interference, aggression, and noncompliance and teacher aggression ratings, r = .46 between observed off-task and teacher-rated inattention-overactivity
Solanto et al. [49]	7–10	106	ADHD	TD	–	DO TS	16 RC	COC (five variables): interference, off-task, motor standing, motor vigorous, physical aggression (IRR: phi = .80–1)	Modest (all r < .4) correlations between COC and performance in stop-signal and delay aversion tasks
Hosterman et al. [10]	mean 8–9	172	ADHD	TD	–	DO TS	30 RC	BOSS (five variables): active, passive engagement, off-task motor, verbal, and passive (IRR:κ = .89–.96)	Moderate, sig. (r = .21–.56) correlations between BOSS and teacher ADHD ratings
Sonuga-Barke et al. [50]	6–12	184	ADHD	None	–	DO TS	7× daily LS	SKAMP (13 items) rating of attention and deportment (IRR: not reported)	Small correlations (all r < .3) between SKAMP and parent ratings
McConaughy et al. [51]	6–12	310	ADHD subtypes	TD	–	DO TS RS	40 RC	DOF (89 items) (IRR: r = .71–.80)	Small to moderate correlations between DOF ADHD problems and parent ADHD ratings (r = .09–.20) and teacher ADHD ratings (r = .22–.36)
Pharmacological intervention studies									
Gadow et al. [52]	5–11	11	Aggressive hyperactive	Classmates	MPH	DO TS	6 × 2 h RC	Modified COC (five variables): interference, noncompliance, motor movement, aggression, off-task (IRR: κ = .76–1)	MPH improved* aggression, off-task, noncompliance, disruptiveness
Carlson et al. [53]	6–12	24	ADHD	None	MPH, BM	DO TS CS	10 STP	Adapted COCADD (eight variables): aggression, verbal abuse/teasing, inappropriate use of property, cheating, intrusion, talking to self, out-of-seat, off-task (IRR: r = .92–.96)	MPH improved* off-task, disruptive behavior
Pelham et al. [54]	5–9	31	ADHD	None	MPH, BM	DO TS CS	10 STP	Adapted COCADD (eight variables): aggression, verbal abuse/teasing, inappropriate use of property, cheating, intrusion, talking to self, out-of-seat, off-task (IRR: κ = .69–.75)	MPH improved* off-task, disruptive behavior BM improved* disruptive behavior
Gadow et al. [55]	6–11	34	ADHD + Tic	None	MPH	DO TS	3–4 × 30 RC	Modified COC (five variables): interference, noncompliance, motor movement, aggression, off-task (IRR: κ = .57–.84)	MPH improved* interference, motor movement, off-task, noncompliance
Klein and Abikoff [56]	6–12	89	ADHD	None	MPH, BT, comb	DO TS RS	13 × 16 RC	COC (14 variables): interference, out-of-chair, noncompliance, off-task, minor motor, gross motor etc. (IRR: not reported)	MPH and BT combined improved* 7 out of 14 variables over BT alone, MPH and BT normalized COC behaviors
Nolan and Gadow [57]	6–11	34	ADHD + Tic	Classmates	MPH	DO TS	3–4 × 60 RC	ADHD-SOC (five variables): interference, motor movement, noncompliance, aggression, off-task (IRR: κ = .57–.84)	Off-task, interference, motor movement, noncompliance: ADHD > classmates*, with MPH* normalization of approx. 75% of children Sensitivity = 91%, specificity = 80%

Table 2 (continued)

Author(s)	Age	*N*	Diagnosis	Control group	Intervention		Duration (min)	Observational instrument	Results
Wigal et al. [58]	7–12	34	ADHD	None	MPH	DO RS	4× daily LS	10-item SKAMP rating of attention and deportment (IRR: not reported; TRR: $r = .63–.78$)	MPH improved* SKAMP ratings Moderate to strong correlations between SKAMP ratings and IOWA inattention-overactivity ratings ($r = .50–.84$), for same time periods rated by same observer
Swanson et al. [59]	7–14	33	ADHD	None	Adderall	DO RS	6× daily LS	10-item SKAMP rating of attention and deportment (IRR: not reported)	Adderall improved* SKAMP ratings
Pelham et al. [60]	5–12	25	ADHD	None	MPH, Adderall	DO TS CS	All-day STP	Adapted COCADD (two variables): on-task, disruptive behavior (IRR: κ = .80–.82)	MPH and Adderall improved* on-task, disruptive behavior
Pelham et al. [61]	6–12	21	ADHD	None	MPH, Adderall	DO TS CS	All-day STP	Adapted COCADD (two variables): on-task, disruptive behavior (IRR: not reported)	MPH and Adderall improved* on-task, disruptive behavior, and all point system measures (except positive peer behaviors and attention)
Pelham et al. [62]	6–12	68	ADHD	None	MPH	DO TS CS	9× daily LS	Adapted COCADD (two variables): on-task, disruptive behavior (IRR: phi = .60–.74)	MPH improved* on-task, disruptive behavior
Swanson et al. [63]	7–12	32	ADHD	None	MPH	DO RS	10× daily LS	10-item SKAMP rating of attention and deportment (IRR: not reported)	MPH improved* SKAMP ratings
Greenhill et al. [64]	7–12	12	ADHD	None	Adderall	DO RS	9× daily LS	10-item SKAMP rating of attention and deportment (IRR: not reported)	Adderall improved* SKAMP ratings
Lopez et al. [65]	6–12	36	ADHD	None	MPH	DO RS	8× daily LS	10-item SKAMP rating of attention and deportment (IRR: not reported)	MPH improved*SKAMP ratings
McCracken et al. [66]	6–12	51	ADHD	None	Adderall	DO RS	8× daily LS	10-item SKAMP rating of attention and deportment (IRR: not reported)	Adderall improved* SKAMP ratings
Abikoff et al. [67]	7–9	103	ADHD	None	MPH	DO TS	6× over 2 years, RC	COC (three variables): interference, off-task, gross-motor (IRR: phi = .83–.92)	MPH improved* interference, off-task, gross-motor (no sig. long-term changes)
Döpfner et al. [68]	8–15	82	ADHD	None	MPH	DO RS	5× daily LS	10-item SKAMP rating of attention and deportment (IRR: $r = .61–.74$)	MPH improved* SKAMP ratings
Swanson et al. [69]	6–12	184	ADHD	None	MPH	DO RS	7× daily LS	13-item SKAMP rating of attention and deportment (IRR: not reported)	MPH improved* SKAMP ratings
Biederman et al. [70]	6–12	57	ADHD	None	Amph., atomox	DO RS	6× daily LS	13-item SKAMP rating of attention and deportment (IRR: not reported)	Mixed amphetamine improved* SKAMP ratings more than Atomoxetine
McGough et al. [71]	6–12	79	ADHD	None	MPH	DO RS	9× daily LS	13-item SKAMP rating of attention and deportment (IRR: not reported)	MPH* improved SKAMP ratings
Silva et al. [72]	6–12	53	ADHD	None	MPH	DO RS	10× daily LS	13-item SKAMP rating of attention and deportment (IRR: not reported)	MPH improved* SKAMP ratings
Silva et al. [73]	6–12	54	ADHD	None	d-MPH	DO RS	10× daily LS	13-item SKAMP rating of attention and deportment (IRR: not reported)	d-MPH improved* SKAMP ratings
Brams et al. [74]	6–12	86	ADHD	None	d-MPH	DO RS	7× daily LS	13-item SKAMP rating of attention and deportment (IRR: not reported)	d-MPH improved* SKAMP ratings
Muniz et al. [75]	6–12	84	ADHD	None	d-MPH	DO RS	10× daily LS	13-item SKAMP rating of attention and deportment (IRR: not reported)	d-MPH improved* SKAMP ratings

Table 2 (continued)

Author(s)	Age	*N*	Diagnosis	Control group	Intervention		Duration (min)	Observational instrument	Results
Schulz et al. [76]	6–14	147	ADHD	None	MPH	DO RS	5× daily LS	13-item SKAMP rating of attention and deportment (IRR: not reported)	MPH improved* SKAMP ratings
Brams et al. [77]	6–12	158	ADHD	None	d-MPH	DO RS	8× daily LS	13-item SKAMP rating of attention and deportment (IRR: not reported)	d-MPH improved* SKAMP ratings
Wigal et al. [78]	6–12	20	ADHD	None	MPH	DO RS	9× daily LS	13-item SKAMP rating of attention and deportment (IRR: not reported)	MPH improved* SKAMP ratings
Williamson et al. [79]	9–12	135	ADHD ± LD	None	MPH	DO RS	daily LS	13-item SKAMP rating of attention and deportment (IRR: not reported)	MPH improved* SKAMP ratings
Childress et al. [80]	6–12	107	ADHD	None	Amph. sulfate	DO RS	7× daily LS	13-item SKAMP rating of attention and deportment (IRR: not reported)	Amphetamine sulfate improved* SKAMP ratings
Manos et al. [81]	6–12	25	ADHD	None	Lisdex-amph., BM	DO RS	6× daily STP	13-item SKAMP rating of attention and deportment (IRR: not reported)	Lisdexamph. and comb. with BM improved* SKAMP ratings
Non-pharmacological intervention studies									
DuPaul et al. [82]	6–11	28	ADHD	Classmates	CT	DO TS	45 weekly RC	BOSS (four variables): active and passive engagement, off-task, fidgeting (IRR: κ = .77–.92)	Peer tutoring improved* engagement
Barkley et al. [83]	3–5	158	Disruptive behavior	WL	PT, CT, comb.	DO RS	1 × 60 RC	89 items DOF rating scale of classroom behavior (IRR: *r* = .69–.80)	CT improved* DOF externalizing problem scale
Pfiffner et al. [84]	Mean 8	57	ADHD	None	CT	DO TS	2 × 45 RC	BOSS (one variable): on-task (IRR: κ = .86)	School-home intervention improved* on-task
Steiner et al. [85]; Steiner et al. [86]	2nd/4th gr.	104	ADHD	BAU	NF, AT	DO TS	3 × 45 RC	BOSS (two variables): on-task, off-task motor/verbal (IRR: mean κ = .89)	NF improved* off-task more than AT and control condition; no further improvement at follow-up
Studies with adolescents									
Carroll et al. [87]	13–17	58	ADHD	TD	-	DO RS	20 RC	RIPPS: coding of responses and triggers, rating of severity (IRR: mean 89%)	Classroom off-task, challenging behaviors: ADHD > TD*
Evans et al. [88]	11–14	25	ADHD	Classmates	-	DO TS	2–3× monthly RC	COCADD (two variables): off-task, disruptive behavior (IRR: mean phi = .84)	Small to moderate correlations between teacher ratings and on-task (*r* = .19–.34), smaller correlations with disruptive behavior (*r* = .08–.24)
Evans and Pelham [89]	11–15	9	ADHD	None	MPH	DO TS	daily STP	COCADD (two variables): off-task, disruptive behavior (IRR: κ = .39–.83)	MPH improved* classroom disruption and inattention
Evans et al. [90]	12+	45	ADHD	None	MPH	DO TS	daily STP	COCADD (two variables): off-task, disruptive behavior (IRR: not reported)	MPH improved* classroom disruption and inattention

Studies are sorted by year of publication within each section

ADHD-C ADHD combined type, *ADHD-IN* ADHD inattentive type, *ADHD-SOC* ADHD School Observation Code, *ANX* anxiety disorder, *AT* attention training, *BAU* business as usual, *BM* behavior modification, *BOSS* Behavioral Observation of Students in Schools, *BT* behavioral treatment, *COC* Classroom Observation Code, *COCADD* Classroom Observation of Conduct and Attention Deficit Disorder, *CS* continuous sampling, *CT* classroom treatment, *d-MPH* dexmethylphenidate, *DBD* disruptive behavior disorder, *DO* direct observation, *DOF* Direct Observation Form, *gr.* grade, *GUCCI* Ghent University Classroom Coding Inventory, *HBRS* Hillside Behavior Rating Scale, *improved** improved significantly ($p < .05$), *LS* laboratory school, *LD* learning disability, *MAI* Munich Observation of Attention Inventory, *MPH* methylphenidate, *MR* mental retardation, *NF* neurofeedback, *PT* parent training, *RC* regular classroom, *RIPPS* Response to Interpersonal and Physically Provoking Situations, *RS* rating scale, *SKAMP* Swanson, Kotkin, Agler, M-Flynn, and Pelham, *SOS* Student Observation System, *STP* summer treatment program, *TD* typically developing, *TS* time sampling, *VT* videotape, *WL* waitlist

Direct Observation Form (DOF) The DOF [27] is composed of a narrative part, a whole-interval sampling recording of on- and off-task behavior (5-s intervals), and an 89-item rating scale of problem behaviors to be completed for observations of 10 min. It assesses five syndrome scales and a DSM-oriented ADHD problem scale with subscales of inattention and hyperactivity-impulsivity. Table 2 includes four studies using the DOF. (1) Correlations between observers ranged from $r = .69$ to .57 [83] and from $r = .97$ to 1 [35]. The test–retest coefficients for the DOF scales and its on-task measure ranged between $r = .25$ and $r = .77$ (mean for problem scales $r = .56$) in a sample of 27 clinically referred children, as indicated in the instrument's manual [27]. According to Volpe et al. [95], five 10-min DOF observations are required to reach acceptable generalizability and dependability for the DOF scales of ADHD problems and hyperactivity-impulsivity, whereas 11–14 observations are necessary for the sluggish cognitive tempo syndrome scale and the attention problem subscale. (2) ADHD subtype differences were demonstrated by the DOF [35]. Discriminant analyses based on DOF classroom observations revealed correct classification rates ranging from 61 to 67% for ADHD combined type versus clinically referred children without ADHD and normal controls, as well as 70% correct classification of ADHD inattentive type versus controls (no significant difference versus the non-ADHD referred clinical sample) [43]. (3) Regarding convergent validity, low to moderate correlations were found between the DOF ADHD scale and the parent AD/HD rating scale-IV ($r = .09$–.33) and between the DOF ADHD scale and the teacher AD/HD rating scale-IV ($r = .21$–.36). For two subscales of the DOF (oppositional and intrusive), some incremental validity was demonstrated, as indicated by 2–6% of additional variance accounted for in parent- and teacher-rated ADHD symptoms [51]. (4) As treatment outcome variable, the DOF showed significantly lower levels of externalizing behavior in a treated group of ADHD preschoolers compared to untreated controls [83].

Classroom Observations for Conduct and Attention Deficit Disorder (COCADD)—Children The COCADD [96] consists of 32 measures in five domains of classroom behavior (position, physical-social orientation, vocal activities, non-vocal activities, play), which are coded using a 2-s whole-interval sampling procedure. Since 1990, modified versions of the COCADD have been applied in six summer treatment program studies with ADHD children. (1) Kappa indices of IRR ranged between .42 and .78 [44] and .69 to .75 [54]. The TRR was not reported. (2) Teacher-identified students with ADHD were predicted in 83% of the cases (with 9% false positives and 24% false negatives) by using three variables of the COCADD (sitting, verbal intrusion, and talking to self) and three measures of desk checks and academic performance in the original study [96]. (3) COCADD overactive behavior correlated significantly with the IOWA Conners teacher rating of inattention-overactivity ($r = .23$) and COCADD verbal disruptive behavior with teacher-rated inattention-overactivity ($r = .21$) and aggression ($r = .41$) in the classroom in a sample of mixed ADHD/disruptive and unselected boys. Otherwise, no significant correlations emerged (e.g., the correlation between COCADD attending and inattention-overactivity was $r = .02$) [97]. (4) Sensitivity to pharmacological interventions was shown in the analogue classroom of several summer treatment program studies [53, 54, 60, 61] and a laboratory school study [62] for the modified version of the COCADD.

COCADD—Adolescents Three studies employed the measures of off-task behavior and disruptive behavior of the COCADD in adolescents with ADHD. (1) IRR was low to adequate (phi = .84 [88]; kappa = .39–.83 [89]). No adolescent-specific TRR was reported, but off-task behavior seemed to vary considerably between different school subjects with different teachers (science versus math class; $r = .25$) [88]. (2) No group comparisons between adolescents with and without ADHD were conducted using the COCADD (no sensitivity/specificity analyses). (3) Student off-task behavior correlated moderately with the teacher AD/HD rating scale-IV (strongest correlation between total ADHD symptoms and on-task behavior $r = -.27$) [88]. (4) The COCADD variables of off-task and disruptive behavior proved to be sensitive to pharmacological interventions in the analogue junior high school lecture classroom in two summer treatment program studies [89, 90].

Swanson, Kotkin, Agler, M-Flynn, and Pelham Scale (SKAMP) The 13 items of the SKAMP [98] are highly time- and situation-specific. The SKAMP is used to assess classroom-specific observable symptoms of inattention (e.g., staying seated) and deportment (e.g., getting started) over short time spans of 30–45 min [50]. It has been employed in the laboratory analogue classroom in 20 medication trial studies since 1990. (1) Döpfner et al. [68] found an IRR of $r = .61$ for the deportment scale and $r = .74$ for the inattention scale; otherwise, IRR was not reported for the SKAMP. TRR coefficients of the SKAMP were moderate to high ($r = .63$–.78) [58]. (2) In a large US sample of elementary school students, SKAMP teacher ratings did not predict later diagnosis of ADHD [99]. These SKAMP ratings were, however, based on the teachers' observations over the previous 4 weeks and thus differ conceptually from direct observational ratings as administered in the laboratory classroom. (3) The SKAMP scales correlated moderately to strongly with the IOWA inattention-overactivity ratings ($r = .50$–.84), which were rated for the same observation periods by the same observer

[58] (i.e., concurrent rather than convergent validity). Swanson et al. [50] reported small to moderate agreements ($r = .21–.25$) with parent SNAP-IV ratings. (4) Sensitivity to various pharmacological interventions has been repeatedly shown for the SKAMP [58, 59, 63, 66, 68–73].

Student Observation System (SOS) The SOS uses a 30-s momentary time-sampling method to assess adaptive behavior categories (e.g., responding to teacher) and maladaptive behavior categories (e.g., inattention, movement) [34]. It was applied in one study of Table 2 [34]. (1) IRR was acceptable ($r = .69–1$); TRR was not examined. (2) The category of inappropriate movement and the maladaptive behavior composite differed significantly between ADHD and controls, but discriminant analyses showed that the SOS failed to add information above and beyond that obtained by teacher ratings alone [34]. Convergent validity (3) and treatment sensitivity (4) were not examined for the SOS.

Hillside Behavior Rating Scale (HBRS) An adapted version of the HBRS was applied in two studies [38, 39] of Table 2. It collected ratings of restlessness, noisiness, interactions, disturbance, frustration, and stimulation search (1) with adequate IRR ($r = .70–.98$) from videotaped classrooms [38]. TRR was not reported. (2) ADHD children displayed higher rates of behavior on all scales (except interactions) than typically developing peers; no ADHD prediction was calculated [38]. Convergent validity (3) and treatment sensitivity (4) were not examined for the HBRS in the classroom.

Ghent University Classroom Coding Inventory (GUCCI) The GUCCI is a continuous sampling coding scheme for behaviors of activity, nonsocial vocalization, and social behavior [45] or time on-task [46] (applied in two studies of Table 2). (1) IRR was high (kappa = .74–.99) [45]. TRR was not reported. (2) Significant group differences were found, but no predictive analysis of ADHD was conducted. Convergent validity (3) and sensitivity to change (4) were not evaluated.

Munich Observation of Attention Inventory (MAI) The MAI measures off- and on-task behavior with the use of a 5-s time-sampling procedure. It was applied in one study [42] (Table 2). (1) IRR and TRR were not assessed. (2) Children with ADHD differed significantly from controls by displaying more off-task behavior, but also initiating more on-task behavior. Passive inattention explained most variance in teacher ratings. Predictive validity was not assessed [42]. (3) Observed off-task behavior was moderately related to teacher DSM-III-R ADHD ratings ($r = .41–.50$) and inconspicuous on-task behavior (e.g., reading, writing) reached a correlation coefficient of $r = -.71$ with teacher ADHD ratings [42]. (4) No treatment evaluation study has applied the MAI.

Responses to Interpersonal and Physically Provoking Situations (RIPPS) The RIPPS is a classroom observation schedule that was applied with ADHD adolescents in one study [87]. It records the student's emotional responses and triggers. (1) The IRR was high (80%). TRR was not reported. (2) The RIPPS revealed higher rates of off-task and disruptive behavior in adolescents with ADHD than in controls [87]. Predictive validity, convergent validity (3) and treatment response (4) were not evaluated using the RIPPS.

Short Summary Eleven different systematic tools for classroom observation were used in a total of 58 studies (Table 2: ADHD-SOC [$n = 1$], BOSS [$n = 9$], COC [$n = 7$], COCADD [$n = 9$], DOF [$n = 4$], GUCCI [$n = 2$], HBRS [$n = 2$], MAI [$n = 1$], RIPPS [$n = 1$], SKAMP [$n = 21$], SOS [$n = 1$]).

- IRR: mostly acceptable ($r = .61–1$, phi = .60–1, kappa = .39–.99). The lowest Pearson r was reported for the SKAMP [68], the lowest phi and kappa coefficients were reported for the COCADD [44, 62, 89]. Not reported in 24 studies.
- TRR: reported for two instruments (COC, SKAMP), ranging between $r = .27$ and .78 [48, 58]; between $r = .25$ and .77 on the DOF for clinically referred children [27].
- Correct classification: ranged between 61% (DOF [43]) and 86% (ADHD-SOC [57]); analyzed for seven instruments (COC [91], ADHD-SOC [57], BOSS [41], DOF [43], COCADD [96], SKAMP [99], SOS [34]).
- Convergent validity: reported in nine studies for six different tools (COC [48], BOSS [10, 37], DOF [51], COCADD [88, 97], SKAMP [50, 58], MAI [42]); poor agreements with parent ratings ($r = .09–25$), moderate to occasionally strong agreements with teacher ratings ($r = .21–.93$), moderate agreements with neuropsychological tests ($r = .26–.40$).
- Treatment outcome: significant pharmacological intervention effects were found using the ADHD-SOC ($n = 1$), COC ($n = 4$), COCADD ($n = 7$), and the SKAMP ($n = 20$); significant effects of non-pharmacological interventions were found using the BOSS ($n = 4$) and DOF ($n = 1$).

Observations in Naturalistic Social Interaction Settings

An overview of social interaction observation studies ($n = 25$) is given in Table 3. Six different specific observational instruments were employed:

Summer Research Program Observations Six studies of Table 3 applied all-day observational schedules during summer research programs involving two (i.e., noncompliance,

Table 3 Social interaction observation studies of children with ADHD ($n = 25$)

Author(s)	Age	*N*	Diagnosis	Control group	Intervention		Duration (min)	Observational instrument	Results
Group discrimination studies									
Erhardt and Hinshaw [100]	6–12	49	ADHD	TD	–	DO TS	All-day	SRP-Obs. (five variables) on-task, prosocial, noncompliant-disruptive, aggression, social isolation (IRR: κ = .30–.73)	Noncompliant-disruptive, aggression: ADHD > TD*
Anderson et al. [101]	6–12	86	ADHD	TD	–	DO TS	All-day	SRP-Obs. (two variables) noncompliance, aggression (IRR: κ = .65–.73)	Noncompliant-disruptive, aggression: ADHD > TD*
Hinshaw et al. [102]	6–12	176	ADHD	TD	–	DO TS	All-day	SRP-Obs. (two variables) noncompliance, aggression (IRR: κ = .65–.73)	Noncompliant-disruptive, aggression: ADHD > TD* (stealing, property destruction: ADHD > TD*)
Hinshaw et al. [103]	6–12	133	ADHD	TD	–	DO TS	All-day	SRP-Obs. (five variables) aggression, noncompliance, social isolation, prosocial, compliance (IRR: κ = .65–.72)	Noncompliant-disruptive, aggression: ADHD > TD* (stealing, property destruction: ADHD > TD*)
DuPaul et al. [104]	3–5	94	ADHD	TD	–	DO TS	60 PS	Adapted ESP (three variables): negative social behaviors, positive social behaviors, activity change (IRR: mean κ = .81)	Negative social behavior: ADHD > TD*
Mikami and Hinshaw [105]	6–12	149	ADHD	TD	–	DO TS	All-day	SRP-Obs. (five variables) aggression, noncompliance, socializing, solitary play, alone (IRR: > .70)	Aggression: ADHD > TD* Solitary play: ADHD < TD*
Riley et al. [106]	3–5	102	ADHD subtypes	None	–	DO TS	15 PS	ESP (two variables): off-task, disruptive behavior (IRR: κ = .74–.83)	No sig. subtype differences in observed behavior
Pollack et al. [107]	3–5	107	ADHD ± ODD, ANX	None	–	DO TS	20 PS	ESP (five variables): solitary play, parallel play, positive social engagement, negative verbal behavior, negative physical behavior (IRR: κ = .82–.88)	No sig. subgroup differences in play behavior
Convergent validity									
Nolan and Gadow [48]	5–13	23	ADHD	None	–	DO TS	2 × 20 PG/LR	COSA (five variables): appropriate social interaction, physical aggression, nonphysical aggression, verbal aggression, noncompliance (IRR: κ = .79–.92; TRR: ICC = −.20–.24, *r* = −.43–.68)	Few sig. moderate correlations (*r* = .41–.66) between observed aggression and noncompliance and teacher aggression and inattention-overactivity rating
Nigg et al. [108]	6–12	38	ADHD	TD	–	DO TS	All-day	SRP-Obs. (three variables) aggression, noncompliance, nonsocial (IRR: κ = .40–.73)	Modest correlations between observed aggression and CPT performance (*r* = .38) and parent ratings (*r* = .40)

Table 3 (continued)

Author(s)	Age	N	Diagnosis	Control group	Intervention		Duration (min)	Observational instrument	Results
Thomas et al. [109]	3–5	137	At risk for ADHD	None	–	DO TS	20 min PS	ESP (six variables): parallel play, positive social engagement, solitary play, negative verbal, negative physical, disruptive (IRR: κ = .81–.91)	Small correlations (all $r < .30$) between ESP and teacher ratings
Pharmacological intervention studies									
Gadow et al. [52]	5–11	11	Aggressive hyperactive	Classmates	MPH	DO TS	6 × 2 h PG/LR	COSA (five variables): appropriate social interaction, physical aggression, nonphysical aggression, verbal aggression, noncompliance (IRR: κ = .76–1)	MPH improved* verbal and physical aggression on the playground
Pelham et al. [110]	8–13	22	ADHD	None	MPH, amph., pemol	DO CS	All-day STP	ADRCS (five variables): following rules, positive peer behavior, noncompliance, conduct problems, negative verbalizations (IRR: not reported)	All medication conditions improved* all observed variables
Murphy et al. [111]	6–11	26	ADHD ± aggression	None	MPH	DO CS	All-day STP	ADRCS (three variables): negative verbalizations, conduct problems, negative interactions (IRR: r = .69–.99)	Negative verbalizations, conduct problems: ADHD high aggression > ADHD low aggression* MPH improved* all observed variables
Gadow et al. [55]	6–11	25	ADHD + Tic	None	MPH	DO TS	3–4 × 60 PG/LR	COSA (five variables): appropriate social interaction, physical aggression, nonphysical aggression, verbal aggression, noncompliance (IRR: κ = .57–.84)	MPH improved* aggression in the lunchroom and on the playground
Nolan and Gadow [57]	6–11	34	ADHD + Tic	Classmates	MPH	DO TS	3–4 × 60 PG/LR	ADHD-SOC (five variables): appropriate social behavior, nonphysical aggression, noncompliance, physical aggression, verbal aggression (IRR: κ = .57–.84)	Noncompliance, aggression on playground and in lunchroom: ADHD > classmates* MPH* improved deviant behaviors
Pelham et al. [112]	7–9	117	ADHD	None	BT, MPH + BT	DO CS	All-day STP	ADRCS (eight variables): following rules, rule violations, sportsmanship, noncompliance, interruption, complaining, positive peer behavior, conduct problems; classroom (one variable): rule following (IRR: r = .66–.98, κ = .65–.89)	Rule-following behavior, good sportsmanship: combined treatment group > behavioral treatment*

Table 3 (continued)

Author(s)	Age	N	Diagnosis	Control group	Intervention		Duration (min)	Observational instrument	Results
Pelham et al. [113]	7–12	136	ADHD	None	MPH	DO CS	All-day STP	ADRCS (seven variables): following rules, noncompliance, interruption, complaining, positive peer behavior, conduct problems, negative verbalizations; classroom (three variables): rule following, on-task, disruptive behavior (IRR: mean $r = .82$, mean $\kappa = .75$)	MPH improved* all observed variables except positive peer behavior
Chacko et al. [114]	5–6	36	ADHD	None	MPH	DO CS	All-day STP	ADRCS (four variables): rule following, compliance, conduct problems, negative verbalizations; classroom (one variable): rule following (IRR: not reported)	MPH improved* all observed variables
Pelham et al. [115]	6–12	27	ADHD	None	MPH, BM	DO CS	All-day STP	ADRCS (seven variables): following rules, noncompliance, interruption, complaining, positive peer behavior, conduct problems, negative verbalizations; classroom (one variable): rule following (IRR: mean $r = .79$, mean $\kappa = .48$)	MPH and BM (alone) improved* all observed variables
Pelham et al. [116]	7–12	36	ADHD	None	MPH	DO CS	All-day STP	ADRCS (six variables): rule following, compliance, interruption, complaining, conduct problems, negative verbalizations; classroom (one variable): rule following (IRR: $r = .82–.88$, mean $\kappa = .59$)	MPH improved* all observed variables
Non-pharmacological intervention studies									
Chronis et al. [117]	6–13	44	ADHD + DBD	None	BM	DO CS	All-day STP	ADRCS (seven variables): following rules, noncompliance, interruption, complaining, conduct problems, negative verbalizations, rule violations; classroom (one variable): rule following (IRR: mean $r = .88$)	Sig. deterioration on all observed measures in the withdrawal phase without behavior modification
Fabiano et al. [118]	6–12	71	ADHD	None	TO	DO CS	All-day STP	ADRCS (three variables): noncompliance, aggression, property destruction (IRR: not reported)	Time-out procedures improved* all observed variables, normalization of noncompliant behavior in 75%
Mrug et al. [119]	5–13	268	ADHD	None	BM	DO CS	All-day STP	ADRCS (12 variables): rule following, compliance, sharing, aggression, teasing, interrupting, etc. (IRR: $r = .61–.93$)	BM improved* 6 out of 12 observed variables

Table 3 (continued)

Author(s)	Age	*N*	Diagnosis	Control group	Intervention		Duration (min)	Observational instrument	Results
Webster-Stratton et al. [120]	4–6	99	ADHD	WL	PT	DO RS	2×30 PS	26 items of COCA-R (dimensions: cognitive concentration, authority acceptance, social contact) (IRR: ICC=.87–.93)	Parent training improved* observed social contact Small to moderate correlations between COCA-R and teacher ratings (*r*=.26–.39)

Studies are sorted by year of publication within each section

ADRCS all-day response-cost system, *ANX* anxiety, *BM* behavior modification, *BT* behavioral treatment, *COCA-R* Coder Observation of Child Adaptation-Revised, *COSA* Code for Observing Social Activity, *CPT* continuous performance task, *CS* continuous sampling, *CT* combined treatment, *DBD* disruptive behavior disorder, *DO* direct observation, *ESP* Early Screening Project, *improved** improved significantly (*p*<.05), *LR* lunchroom, *MPH* methylphenidate, *ODD* oppositional defiant disorder, *PG* playground, *PS* preschool, *PT* parent training, *RS* rating scale, *SRP-Obs*. Summer research program observation, *STP* summer treatment program, *TD* typically developing, *TO* time-out, *TS* time sampling, *VT* videotape, *WL* waitlist

aggression) to five (i.e., prosocial behavior, social isolation, nonsocial behavior) variables to measure social interactions during different activities by a 5-s whole-interval sampling procedure. (1) IRR did not reach adequate levels in some of the variables (e.g., for prosocial behavior kappa = .31 [100], nonsocial behavior kappa = .40 [108]). Better agreement was reported for noncompliance (kappa = .65) [103] and aggression (kappa = .73) [108]. TRR was not reported. (2) Four summer research program studies reported higher rates of noncompliant and aggressive behavior in ADHD children than in comparison children [100–103]. Sensitivity and specificity were not evaluated. (3) Correlations between observed aggression and continuous performance task (CPT) scores were moderate but significant (*r* = .38) [108]. Observed aggression was also significantly correlated with mother-rated externalizing problems on the Child Behavior Checklist (CBCL). Noncompliance and nonsocial behavior revealed no significant associations with parent ratings [108]. (4) Medication had a significant attenuating effect on observed noncompliance and aggression [108].

Early Screening Project (ESP) The play behavior of young children with ADHD was observed in preschools with the observation component of the ESP [121] in four studies of Table 3. The code uses a partial-interval, a whole-interval, and a momentary-interval time-sampling system with 15-s intervals. (1) The ESP allows different aspects of positive (e.g., positive social engagement) and negative social interactions (e.g., disruptive behavior) to be recorded and has shown adequate IRR (kappas = .81–.93) [106, 107, 109]. (2) During unstructured free play, preschoolers with ADHD displayed significantly more negative social behavior than typically developing children [104]. ADHD subtype and comorbidity did not lead to significant differences on the ESP [106, 107]. Predictive validity was not examined. (3) Teacher ratings on the Social Skills Rating System in children at risk of ADHD correlated weakly (*r* < .30) with observed ESP solitary play and aggression [109]. (4) The ESP was not used for treatment evaluation.

Coder Observation of Child Adaptation-Revised (COCA-R) The COCA-R is a preschool observational instrument in a rating scale format. It was applied in one study of Table 3 [120]. (1) Observers achieved adequate IRR (*r* = .87–.93) on the COCA-R. TRR was not reported. (2) COCA-R scores of an ADHD sample were not compared to healthy controls; no discriminant analyses were conducted. (3) Correlations with the Conners Teacher Rating Scale were moderate and significant (*r* = .26–.39). (4) Combined parent and child training induced a significant improvement on the COCA-R social contact scale compared to an ADHD waitlist group [120].

Code for Observing Social Activity (COSA) and ADHD-SOC In the playground and in the lunchroom, observations of children with ADHD were conducted with the COSA in three studies and with its precursor—the ADHD-SOC—in one study of Table 3. Both codes use a 15-s partial-interval time-sampling procedure to record aggression, noncompliance, and appropriate social interactions (30-s intervals in [52]). (1) These observations yielded low to adequate IRR (kappa = .57–.94) [48, 57]. For lunchroom measures, the TRR was almost entirely non-significant at an interval of 1 day but stronger at an interval of 2 days (r = .35–.68). Playground behavior of children with ADHD was highly unstable [48]. (2) Children with ADHD and comorbid tic disorder exhibited higher levels of observed aggressive and noncompliant behavior in the lunchroom and higher levels of physical aggression in the playground than classmates. Specificity and sensitivity analyses of the ADHD-SOC lunchroom and playground behavior were not conducted [57]. (3) Observed aggressive behavior in the lunchroom was significantly correlated with aggression in the IOWA Conners Teacher Rating Scale (r = .41–.66). The IOWA rating of inattention-overactivity was negatively correlated with playground physical aggression (r = − .52) [48]. (4) Significant reductions in aggression in the playground and in the lunchroom with stimulant medication were repeatedly reported [52, 55, 57].

Response-Cost Systems All-day response-cost systems target directly observable behaviors. These systems produce a frequency count of undesirable behaviors across daily classroom and recreational periods. Ten studies of Table 3 included such point systems, which assessed the frequency of 3–12 variables (e.g., noncompliance, rule following, negative verbalizations). (1) Agreement between raters seemed rather variable (e.g., r = .44–.96 [115]) and TRR is unknown. (2) The predictive validity and (3) the convergent validity were not examined. (4) This instrument was found to be sensitive to various pharmacological [110, 111, 114] and behavioral treatments [118, 119].

Short Summary Six different systematic tools for naturalistic social interaction observations were used in a total of 25 studies (Table 3: ADHD-SOC [n = 1], response-cost systems [n = 10], COCA-R [n = 1], COSA [n = 3], ESP [n = 4], Summer Research Program Observations [n = 6]).

- IRR: mostly acceptable (r = .61–.99, ICC = .87–.95, kappa = .30 – 1). The lowest Pearson r was reported for response-cost systems [119], the lowest kappa was reported for Summer Research Program Observations [100]. Not reported in three studies [110, 114, 118].
- TRR: reported for one instrument (COSA playground and lunchroom observations) [48], ranging between ICC = − .20 and .24.
- Correct classification: not examined beyond significant differences on the group level (Summer Research Program Observations, ESP, COSA).
- Convergent validity: reported in four studies for four instruments (COSA [48], COCA-R [120], ESP [109], Summer Research Program Observations [108]); small to moderate agreements with parent ratings (r = .07–.40), teacher ratings (r = .20–.66), and CPT scores (r = .13–.38).
- Treatment outcome: significant pharmacological intervention effects were found using the ADHD-SOC (n = 1), the COSA (n = 1), and response-cost systems (n = 7); significant effects of non-pharmacological interventions were found using response-cost systems (n = 3) and the COCA-R (n = 1).

Observation Studies of Children and Adolescents with ADHD in Laboratory Settings

Independent Play Observations

Since 1990, nine studies with ADHD children employing a specific observational tool during independent play have been published (Table 4).

Structured Observation of Academic and Play Settings (SOAPS) Different versions of the instrument SOAPS [131] were applied in six of the reported studies. Its original version consists of a free and a restricted 15-min play session, in which the duration of on-task behavior, fidgeting, out-of-seat, vocalizing, and the number of task shifts and position changes (i.e., floor grid crossings) is recorded [122]. SOAPS behaviors were time-sampled by the use of a 10-s partial-interval method [122]. Continuous sampling of the duration of behavior (i.e., off-task) and the frequency of behaviors (i.e., grid crossings) was conducted from videotapes [123, 124]. (1) Acceptable IRR was reached (kappas = .73–.99 [124]; 85–99% agreement [127]). Significant long-term stability was reported among a sample of clinic-referred boys for the playroom measures of position changes, on-task behavior, out-of-seat, and vocalizations over 2 years (r = .40–.52 [132]). TRR was otherwise not assessed. (2) Roberts [122] reported 64 and 58% correct classifications of hyperactive, aggressive, and hyperactive and aggressive boys in free play and restricted play, respectively. The original SOAPS was later modified for use with preschoolers. The addition of "forbidden" toys to the playroom differentiated preschoolers with and without ADHD quite strongly [124], but this effect could not be replicated [126]. Seventy percent of mentally retarded children with ADHD were

Table 4 Independent play observation studies of children with ADHD ($n=9$)

Author(s)	Age	N	Diagnosis	Control group	Intervention		Duration (min)	Observational instrument	Results
Group discrimination studies									
Roberts [122]	5–12	36	Hyperactive ± aggression	Aggressive	–	DO TS	30	SOAPS (six variables): on-task, fidgeting, out-of-seat, vocalizations, task shifts, grid crossings (IRR: ICC = .85–.95)	Free play: grid crossings, out-of-seat: hyperactive > aggressive*, correct prediction: 64% Restricted play: task shifts: Hyperactive > aggressive*, correct prediction: 58%
Paternite et al. [123]	6–12	132	ADHD subtypes	TD	–	VT CS	15	SOAPS (four variables): grid crossings, out-of-seat, vocalizations, on-task (IRR: not reported)	Grid crossings, off-task, vocalizations: ADHD-HI > ADHD-IN* Grid crossings, off-task: ADHD-C > ADHD-IN*
Byrne et al. [124]	3–5	26	ADHD	TD	–	VT CS	60	Modified SOAPS (five variables): out-of-seat, grid crossings, task shifts, vocalizations, parent attention seeking (additionally: language test behavioral observation) (IRR: κ = .73–.99)	Play with off-limit toys: ADHD > TD* Language test grabbing, examiner commands: ADHD > TD*
Handen et al. [125]	6–12	42	ADHD + MR ± CD	MR	–	VT TS	10	Modified SOAPS (eight variables): intense activity, vocalizations, movement, play with nontoys, noninvolvement with toys, picking up toys, leaving toys, duration of play (IRR: 85–99%)	Vocalizations, toy changes: ADHD + MR > MR* Sensitivity = 57%, specificity = 88%
DeWolfe et al. [126]	3–6	50	ADHD	TD	–	VT CS	60	Modified SOAPS (six variables): off-task, duration of play episode, play shift, grid changes, verbalization, off limits toy (IRR: κ = .74–.88)	Vocalizations, grid crossings, off-task: ADHD > TD*, duration of play episodes: ADHD < TD*
Pharmacological intervention studies									
Handen et al. [127]	6–12	22	ADHD + MR	None	MPH	VT TS	3 × 10	Modified SOAPS (eight variables): intense activity, vocalizations, movement, play with nontoys, noninvolvement with toys, picking up toys, leaving toys, duration of play (IRR: 85–99%)	MPH improved* intense activity, vocalization, and movement
Non-pharmacological intervention studies									
Sonuga-Barke et al. [128]	3	78	ADHD	WL, TD	PT	VT CS	3 × 10	Index of attention/engagement: time on task/number of switches between zones (two variables) (IRR: $r=.76$; TRR: $r=.81$)	Index of attention/engagement: ADHD < TD* sig. less pronounced decrease in observed attention after parent training
Daley and O'Brien [129]	4–11	43	ADHD	WL	PT	VT CS	2 × 25	Index of attention/engagement: time on task/number of switches between toys (two variables) (IRR: $r=.76–.92$; TRR: $r=.49–.68$)	No sig. treatment effect on play behavior

Table 4 (continued)

Author(s)	Age	*N*	Diagnosis	Control group	Intervention		Duration (min)	Observational instrument	Results
Abikoff et al. [130]	3–4	164	ADHD	WL	PT	VT CS	10	Index of attention/engagement: time on task/number of switches between zones (two variables) (IRR: not reported)	No sig. treatment effect on play behavior

Studies are sorted by year of publication within each section

ADHD-C ADHD combined type, *ADHD-HI* ADHD hyperactive-impulsive type, *ADHD-IN* ADHD inattentive type, *CD* conduct disorder, *CS* continuous sampling, *DO* direct observation, *improved** improved significantly ($p<.05$), *MPH* methylphenidate, *MR* mental retardation, *PT* parent training, *SOAPS* Structured Observations of Academic and Play Setting, *TAU* treatment as usual, *TD* typically developing, *TS* time sampling, *VT* videotape, *WL* waitlist

classified correctly by the SOAPS as cases [125]. Children of the ADHD inattentive type could not be discriminated from controls based on their playroom behavior [123]. (3) Convergent validity was not assessed. (4) Significant effects of MPH were reported for the SOAPS free play behavior in ADHD children with mental retardation (e.g., fewer vocalizations, less movement) [127].

Index of Attention/Engagement Three intervention studies [128–130] applied the observational measure of the index of attention/engagement while children played with a standardized toy. To calculate the index, the observed time on-task was divided by the number of attention switches. The higher the index, the more attention and the less switching were displayed. (1) Acceptable IRR ($r=.76$–91) [128, 129] and a high TRR coefficient ($r=.81$) were reported [128]. Another—much lower—TRR score of .54 was reported in a waitlist ADHD group of 19 subjects [129]. (2) Preschoolers with ADHD had a significantly lower index of attention/engagement than preschoolers without ADHD [128]. Specificity and sensitivity of the index were not evaluated. (3) Convergent validity was not assessed. (4) One study [128] revealed a significantly less pronounced decrease on the attention/engagement index in the treatment group than in the control group. Otherwise, no treatment-related changes were reported [129, 130].

Short Summary Two different systematic tools for independent play observations were used in a total of nine studies (Table 4: Index of attention/engagement [$n=3$], SOAPS [$n=6$]).

- IRR: good (all coefficients > .70). Not reported in two studies [123, 130].
- TRR: reported for one instrument (index of attention/engagement) [128, 129], ranging between $r=.49$ and .81.
- Correct classification: ranged between 58 and 70% on the SOAPS [122, 125].
- Convergent validity: not examined.
- Treatment outcome: significant pharmacological intervention effects were found using the SOAPS [127]; no significant effects of non-pharmacological interventions were detected using the index of attention/engagement [128–130].

Test Session Behavioral Observations

Table 5 displays 27 studies in which children's or adolescents' test or task behavior was assessed using an observational instrument.

Table 5 Test session observation studies of children and adolescents with ADHD ($n=27$)

Author(s)	Age	N	Diagnosis	Control group	Intervention		Duration (min)	Observational instrument	Results
Group discrimination studies									
Barkley et al. [133]	6–11	140	ADD ± H	LD, TD	–	VT TS	24	RAS measures (five variables) (IRR: 74–87%)	CPT off-task: ADD + H > LD/TD/ ADD* CPT vocalizations: ADD > LD/ TD* RAS off-task: ADD ± H > LD*
Roberts [122]	5–12	36	Hyperactive ± aggression	Aggressive	–	DO TS	15	RAS measures (six variables): on-task, fidgeting, out-of-seat, vocalizations, task shifts, grid crossings (IRR: ICC = .85–.95)	On-task: hyperactive ± aggression < aggressive* Grid crossings, out-of seat: hyperactive + aggression > aggressive* Correct prediction of 86% of cases
Pliszka [134]	6–12	107	ADHD ± ANX	TD	–	DO TS	15	RAS measures (five variables) (IRR: not reported)	RAS total score, off-task: ADHD > ADHD/ANX* > TD* Sig. correlation between RAS total score and CPT commission errors ($r = .39$)
Paternite et al. [123]	6–12	132	ADHD subtypes	TD	–	VT CS	15	RAS measures (five variables) (IRR: not reported)	Object play: ADHD-C > ADHD-HI* Out-of-seat, object play: ADHD-C > ADHD-IN*
Glutting et al. [135]	6–12	98	ADHD	Normative sample	–	DO RS	IQ test duration	GATSB 29 items (IRR: not reported)	IQ test inattentiveness, avoidance, uncooperative mood: ADHD > normative sample* Sensitivity = 88%, specificity = 76%
Mariani and Barkley [136]	4–5	64	ADHD	TD	–	DO TS	20	RAS measures (total score of five variables) (IRR: not reported)	RAS total score: ADHD > TD*
Handen et al. [125]	6–12	42	ADHD + MR ± CD	MR	–	VT TS	10	RAS measures (five variables) (IRR: 85–99%)	Forbidden toy touches, off-task: ADHD + MR > MR* Sensitivity = 61%, specificity = 68%
Bauermeister et al. [137]	6–11	98	ADHD subtypes	TD	–	VT TS	30	RAS measures (five variables) (IRR: mean 91–92%)	RAS object play, out-of-seat: ADHD > TD* CPT RAS total score: ADHD > TD*
McConaughy et al. [138]	6–11	177	ADHD subtypes	Clin. ref., TD	–	DO RS	180	TOF 125 items (IRR: not reported)	Oppositional scale, attention scale, ADHD scale, inattention scale, hyperactivity-impulsivity scale, externalizing scale: ADHD-C > ADHD-IN/TD/ clin.ref.* Correct classification: 74% into ADHD-C versus clin.ref
Convergent validity									
Barkley [6]	6–11	140	ADHD	TD	–	VT TS	15	RAS measures (five variables) (IRR: not reported)	Small to moderate correlations between CPT scores and RAS ($r = .12–.34$) Small correlations between parent ratings and RAS ($r = .04–.28$) Small correlations between teacher ratings and RAS ($r = –.01–.28$)

Table 5 (continued)

Author(s)	Age	*N*	Diagnosis	Control group	Intervention		Duration (min)	Observational instrument	Results
Willcutt et al. [139]	3–7	252	ADHD	TD	–	DO RS		HBRS 7 items (IRR: $r = .58–.68$)	Sig. associations between direct behavior ratings and parent and teacher ratings ($r = .32$-.50)
McConaughy et al. [51]	6–12	310	ADHD subtypes	TD	–	DO RS	180	TOF 125 items (IRR: $r = .60–.77$)	Small to moderate correlations between TOF ADHD problems and parent ADHD ratings ($r = .19–.33$) and teacher ADHD ratings ($r = .21–.30$)
Pharmacological intervention studies									
Barkley et al. [140]	6–11	40	ADD ± H	None	MPH	VT TS	15	RAS measures (five variables) (IRR: 67–87%)	RAS total score: ADD + H > ADD-H* MPH improved* off-task, fidgeting, object play, out-of-seat, total
Fischer and Newby [141]	2–17	161	ADHD	None	MPH	DO TS	10	RAS measures (five variables) (IRR: not reported)	MPH improved* all variables
DuPaul et al. [142]	6–12	40	ADHD + internalizing	None	MPH	VT TS	15	RAS measures (five variables) (IRR: 67–87%)	ADHD + high internalizing symptoms improved less with MPH than low internalizing groups
Ialongo et al. [143]	7–11	69	ADHD	TD	MPH	VT TS	30	RAS (two variables) off-task, out-of-seat (IRR: κ = .48–1)	MPH improved* off-task behavior
Handen et al. [127]	6–12	22	ADHD + MR	None	MPH	VT TS	3 × 10	RAS measures (five variables) (IRR: 85–99%)	MPH improved* off-task, object play, out-of-seat
Fischer and Newby [144]	5–17	149	ADHD ± LD	None	MPH	DO TS	3 × 10	RAS measures (five variables) (IRR: not reported)	MPH improved* all variables
Grizenko et al. [145]	6–12	147	ADHD	None	MPH	DO TS	2 × 15	RAS measures (five variables total score) (IRR: not reported)	MPH improved* total score
Gorman et al. [146]	6–12	75	ADHD	TD	MPH	DO TS	3 × 15	RAS measures (five variables total score) (IRR: not reported)	RAS total score: ADHD > TD*, no sig. group differences with MPH
Karama et al. [147]	6–12	138	ADHD	None	MPH	DO TS	2 × 30	RAS measures (five variables) (IRR: ICC = .97–.99; TRR: $r = .61–.67$)	MPH improved* all variables
Grizenko et al. [148]	6–12	493	ADHD subtypes	None	MPH	DO TS	2 × 30	RAS measures (five variables) (IRR: not reported)	Larger effect size of MPH on RAS than on parent and teacher ratings
Non-pharmacological intervention studies									
Green et al. [149]	7–14	26	ADHD	ADHD + placebo training	WMT	VT CS	2 × 15	RAS measures (five variables) (IRR: κ = .95–1)	WM training improved* off-task and object play

Table 5 (continued)

Author(s)	Age	*N*	Diagnosis	Control group	Intervention		Duration (min)	Observational instrument	Results
Studies with adolescents									
Fischer et al. [150]	12–20	160	Hyperactive	TD	–	VT TS	15	RAS measures (five variables) (IRR: 83–99%)	All RAS measures: hyperactive group > TD* age-related decline in off-task, object play, out-of-seat
Barkley [6]	12–20	77, 159	ADHD	TD	–	VT TS	15	RAS measures (five variables) (IRR: not reported)	Small to moderate correlations between CPT scores and RAS (r = .17–.41) Small to moderate correlations between parent ratings and RAS (r = − .01–.36) Small correlations between teacher ratings and RAS (r = − .05–.15)
Barkley et al. [151]	12–17	161	ADHD	TD	–	VT TS	15	RAS measures (five variables) (IRR: not reported)	Off-task, fidgeting, out-of-seat, RAS total score: ADHD > TD*
McGrath et al. [152]	12–17	57	ADHD	Clin. ref.	–	VT TS	15	RAS measures (five variables) (IRR: 74–88%)	No sig. group differences in observed behavior, no sig. correlations with other diagnostic measures

Studies are sorted by year of publication within each section

ADHD-C ADHD combined type, *ADHD-IN* ADHD inattentive type, *ANX* anxiety, *CPT* continuous performance task, *clin. ref* clinically referred, *CD* conduct disorder, *CS* continuous sampling, *DO* direct observation, *GATSB* Guide to Assessment of Test Session Behavior, *HBRS* Hillside Behavior Rating Scale, *improved** improved significantly ($p < .05$), *LD* learning disabled, *MPH* methylphenidate, *MR* mentally retareded, *RAS* Restricted Academic Situation, *RS* rating scale, *TD* typically developing, *TOF* Test Observation Form, *TS* time sampling, *VT* videotape, *WMT* working memory training

Restricted Academic Situation (RAS)—Children The RAS was implemented in 23 of the reviewed studies from Table 5. Originally, the RAS was an extension of the free and restricted play observations SOAPS [131, 153]. Individuals perform written academic math problems in playroom surroundings for 15 min as a laboratory analogue to classroom seatwork. A time-sampling strategy is applied to record the occurrence of usually five behavioral categories within 30-s intervals: off-task behavior, out-of-seat, fidgeting, vocalizing behavior, and object play (hereafter referred to as RAS measures). These same variables and methodology have also been applied to observe behavior during CPTs [133, 137]. (1) Acceptable IRR was reached for the RAS (e.g., ICC = .97–.99 [147], kappa = .86–1 [57]). Significant TRR in school-aged children with ADHD was reported by Karama et al. [147] (factor task disengagement *r* = .67; factor motor activity *r* = .61). An earlier study reported a TRR coefficient of *r* = .86 for RAS total ADHD behavior [153]. (2) The proportion of time on-task of the RAS most effectively separated hyperactive from aggressive children (86%) [122]. A correct classification rate of 64% was reported for children with mental retardation and ADHD by the RAS [125]. However, consistent evidence of discriminatory power for this paradigm is missing, as it was not possible to significantly distinguish between girls with and without ADHD [154], and another study failed to find significant between-group differences in off-task behavior between ADHD children and healthy controls during academic seatwork [133]. Findings are inconsistent regarding subtype differences [123, 137]. (3) Correlations between the RAS behavioral codes and CPT omission and commission errors were low to moderate (*r* = .26–.34) [6]. Pliszka [134] reported a significant correlation of *r* = .39 between CPT commission errors and RAS total score. Total ADHD behavior correlated significantly with parent hyperactivity ratings on the CBCL (*r* = .28), while observed off-task behavior correlated significantly with inattention on the Child Attention Problems Inattention rating scale for teachers (*r* = .28) and on the Conners Teacher Rating Scale (*r* = .26) in the same sample. (4) Significant positive treatment outcome was repeatedly shown in the RAS measures [127, 140–149], also when the same observational categories were applied in the regular classroom [155, 156].

RAS—adolescents The RAS was adapted for adolescents by adding distracting music to the playroom. Four studies with adolescent participants are shown in Table 5. (1) Adequate IRR was reached [150, 152]. Adolescent-specific TRR was not evaluated. (2) Adolescents with ADHD were successfully discriminated from healthy controls by all RAS measures [150], although not consistently [151]. An age-related decline was found in most observational variables [150]. Compared to a clinical control group without ADHD, adolescents with ADHD were not found to display higher scores on the RAS [152]. (3) In the same study, no significant correlations between the RAS measures and other diagnostic instruments were found [152]. However, Barkley [6] reported low to moderate correlations (*r* = .26–.36) between the impulsive-hyperactive factor of the Conners parent rating scale and RAS measures in a mixed sample of adolescents with and without ADHD. (4) Medication functioned as a significant covariate in between-group comparisons, which suggests some sensitivity to pharmacological treatment for the adolescent RAS [152].

Guide to Assessment of Test Session Behavior (GATSB) The GATSB [157] is a normed 29-item rating scale that is completed by examiners after the administration of intelligence tests. It yields scores on the subjects' avoidance, inattentiveness, and uncooperative mood during testing and was applied in one study of Table 5 [157]. (1) Reliability was not evaluated. (2) Classification analysis based on the GATSB revealed a hit rate of 81%, sensitivity of 88%, and specificity of 76% for differentiating children with ADHD hyperactive-impulsive type from non-ADHD controls [135]. (3) Convergent validity and (4) treatment sensitivity were not assessed.

Test Observation Form (TOF) The TOF is a comprehensive direct behavioral rating scale [28], which consists of 125 items describing the child's behavior, affect, and test-taking style. It was employed in two studies of Table 5. (1) IRR ranged between *r* = .60 (for oppositional problems) and *r* = .77 (for ADHD problems) [51]. TRR in a sample of 130 typically developing children was acceptable (*r* = .53–.87, mean *r* = .80 [28]). (2) Children with ADHD combined type differed significantly on six TOF scales from a clinically referred group and a typically developing group. An overall correct classification rate of 74% was reached for the combined type versus a clinically referred sample without ADHD. The predominantly inattentive subtype could not be validly discriminated from the non-ADHD referred sample and healthy controls [138]. (3) The TOF DSM-oriented scale of ADHD problems was significantly correlated with parent ratings of inattention (*r* = .19) and hyperactivity-impulsivity (*r* = .33) on the AD/HD rating scale-IV. Correlations with teacher-rated inattention (*r* = .21) and hyperactivity-impulsivity (*r* = .31) on the AD/HD rating scale-IV were also significant [51]. (4) The TOF has not been employed for treatment evaluation.

Hillside Behavior Rating Scale (HBRS) The seven items of the HBRS were assessed during test sessions in one study. The items were rated after the completion of tests of intelligence and academic achievement in ADHD preschoolers [139]. (1) Significant IRR coefficients (*r* = .58–.68) were reached.

TRR was not reported. (2) The composite ADHD score of the HBRS (with items directly corresponding to DSM-IV) was significantly higher in preschoolers with ADHD than in comparison children. HBRS ratings provided small but significant incremental validity in the prediction of functional impairment over parent and teacher reports [139]. Sensitivity and specificity were not evaluated. (3) Findings for convergent validity between the HBRS DSM-oriented ADHD scale and the number of ADHD symptoms reported by parents on the Diagnostic Interview Schedule for Children (DISC) and teachers on the DSM-IV version of the Disruptive Behavior Disorder (DBD) checklist ranged from $r = .32$ to .50. Correlation coefficients were higher for parent ratings [139]. (4) HBRS test session observations were not used for treatment evaluation.

Short Summary Four different observational tools for observations of test behavior were used in a total of 27 studies (Table 5: GATSB [$n = 1$], HBRS [$n = 1$], RAS [$n = 23$], TOF [$n = 2$]).

- IRR: mostly acceptable (agreement 67–92%, $r = .58$–.68, kappa = .48–1, ICC = .97–.99). The lowest percentage agreement was reported for the RAS [140, 142], the lowest Pearson r was reported for the HBRS [139], the lowest kappa was reported for the RAS [143]. Not reported in 13 studies.
- TRR: reported for one instrument (RAS), ranging between $r = .61$ and .86 [147, 153]; ranging between $r = .53$ and .87 on the TOF for typically developing children [28].
- Correct classification: ranged between 64% (RAS [125]) and 88% (GATSB [135]); analyzed for three instruments (GATSB, RAS, TOF) [122, 125, 135, 138].
- Convergent validity: reported in four studies for three instruments (RAS, HBRS, TOF); small to moderate agreement with parent ratings ($r = .19$–.50), teacher ratings ($r = .21$–.38), and CPT scores ($r = .26$–.39) [6, 51, 134, 139].
- Treatment outcome: significant pharmacological intervention effects were found using the RAS ($n = 10$); significant effects of a non-pharmacological intervention were found using the RAS [149].

Parent–Child Interaction Observations

Eleven studies conducted since 1990 have included behavioral observations of children or adolescents with ADHD while interacting with their parents in the laboratory (Table 6). Only observed child behavior (not parenting) is focused on here.

Disruptive Behavior Diagnostic Observation Schedule (DB-DOS) The DB-DOS was applied in one study of Table 6. Extending the DB-DOS, ten items on ADHD symptoms were added, and a total of 31 items were then rated from 5-min taped interactions between the target child and the parent (parent context) or an examiner (examiner context) [158]. (1) IRR was good (ICC = .88–.95). TRR of the DB-DOS scales was moderate (ICC = .52–.80) in a group of mixed referred and typically developing children. (2) The ADHD scale of the DB-DOS reached sensitivity and specificity of 87 and 79%, respectively, as well as a 75% agreement between DB-DOS and best-estimate ADHD diagnosis. (3) Correlation coefficients between different parent and teacher ratings (Kiddie Disruptive Behavior Disorder Scale, Clinical Global Assessment Scale, CBCL, Teacher Report Form) and the ADHD scale of the DB-DOS were significant ($r = .28$–.42) and slightly more pronounced for parent ratings. (4) No reports on the sensitivity to change of the DB-DOS in ADHD are available.

Global Impressions of Parent–Child Interaction-Revised (GIPCI-R) The GIPCI-R rating scale was applied in one study in preschoolers [160] and one in school-aged children with ADHD [129]. (1) The ratings of child behavior showed adequate IRR ($r = .71$–.84) [129], but lower IRR was achieved in the study in preschoolers (ICC = .48–.77) [160]. TRR was rather low ($r = .41$–.50) [129] ($r = .20$) [160]. (2) Predictive validity and (3) convergent validity were not evaluated for the GIPCI-R. (4) Parent training did not significantly improve GIPCI-R observed child behavior during parent–child interactions [129, 160].

Dyadic Parent–Child Interaction Coding System-Revised (DPICS-R) The DPICS-R was applied in one study of Table 6 [120]. (1) It reached adequate IRR for child deviance and child positive behavior (ICC = .70 and .96, respectively) [120]. TRR, (2) predictive validity and (3) convergent validity were not reported. (4) A significant decrease in child deviance after combined parent and child training for preschoolers with ADHD was reported [120].

MTA Parent–Child Interaction Wells et al. [159] investigated the effects of the multimodal treatment on four rated child behaviors during parent–child interactions (complaining, verbal abuse, compliance, likable). The same observational tool was applied in another non-pharmacological treatment study [161]. (1) The IRR for these direct ratings were reasonable ($r = .62$–.85) [159, 161]. TRR, (2) predictive validity, and (3) convergent validity were not reported. (4) Significant treatment-related changes in observed child behavior were reported by Babinski et al. [161], but not by Wells et al. [159].

Table 6 Parent–child interaction observation studies of children and adolescents with ADHD ($n = 11$)

Author(s)	Age	*N*	Diagnosis	Control group	Intervention		Duration (min)	Observational instrument	Results
Group discrimination studies									
Bunte et al. [158]	3–5	251	ADHD ± DBD	DBD, TD	–	VT RS	15	DB-DOS (extended) 31 items on behavior regulation, anger regulation, and ADHD symptoms (IRR: ICC = .86–.96; TRR: ICC = .52–.80)	All scores: ADHD ± DBD/DBD > TD* Sig. correlations between ADHD items of the DB-DOS and parent and teacher ratings ($r = .28–.42$) 75% agreement between best-estimate ADHD diagnosis and DB-DOS Sensitivity = 87%, specificity = 79%
Pharmacological intervention studies									
Wells et al. [159]	7–9	553	ADHD	TAU	MPH, BT, Comb	VT RS	2 × 30	MTA observation: ratings of 4 child behaviors (complaining, verbal abuse, compliance, likable) (IRR: ICC = .62–.85)	No sig. treatment effect on child behavior in parent–child interaction
Non-pharmacological intervention studies									
Thompson et al. [160]	2–6	41	ADHD	TAU	PT	VT RS	2 × 15	GIPCI-R 7 items (child respect, destruction, disruptive, noncompliance, social skills, valance, disconnection) (IRR: ICC = .48–.77 ; TRR: $r = .20$)	No sig. treatment effect on child behavior in parent–child interaction
Webster-Stratton et al. [120]	4–6	99	ADHD	WL	PT	DO RS	2 × 20	DPICS-R ratings (two variables): deviance, positives (IRR: ICC = .70–.97)	PT improved* child deviance (but no sig. group difference at post-test)
Daley and O'Brien [129]	4–11	43	ADHD	WL	PT	VT RS	2 × 15	GIPCI-R 7 items (child respect, destruction, disruptive, noncompliance, social skills, valance, disconnection) (IRR: $r = .71–.84$, TRR: $r = .41–.50$)	No sig. treatment effect on child behavior in parent–child interaction
Babinski et al. [161]	6–12	12	ADHD	None	PT	VT RS	4 × 13	MTA observation: Ratings of four child behaviors (complaining, verbal abuse, compliance, likeable) (IRR: ICC = .75–.83)	PT improved* compliance, complaining
Studies with adolescents									
Barkley et al. [162]	12–20	160	ADHD ± ODD	TD, no ADHD/ODD	-	VT TR	30	PAICS (six variables): commands/put-downs, defends/complains, problem solution, facilitates, defines/evaluates, talks (IRR: 53–81%)	Neutral discussion: commands/put-downs: ADHD + ODD > ADHD, TD* Conflict discussion: general talking: ADHD < TD*

Table 6 (continued)

Author(s)	Age	*N*	Diagnosis	Control group	Intervention		Duration (min)	Observational instrument	Results
Barkley et al. [163]	12–17	160	ADHD ± ODD	TD	-	VT TR	20	PAICS (six variables): commands/put-downs, defends/complains, problem solution, facilitates, defines/evaluates, talks (IRR: mean κ = .68)	Neutral discussion commands/put downs, defends/complains: ADHD + ODD > TD* Facilitation: ADHD + ODD < TD*
Edwards et al. [164]	12–18	119	ADHD + ODD	TD	-	VT RS	30	CRS with 15 dimensions (10 negative/5 positive communication) (IRR: ICC = .64–.82)	Discussion: negative behavior: ADHD + ODD > TD*
Barkley et al. [165]	12–17	61	ADHD	None	BM, CT, FT	VT TR	3 × 20	PAICS (six variables): commands/put-downs, defends/complains, problem solution, facilitates, defines/evaluates, talks (IRR: 67–85%)	Not uniformly positive effects of treatment on behavior in parent–child interaction
Barkley et al. [166]	12–18	97	ADHD/ODD	None	CT, comb	VT RS	4 × 15	CRS with 15 dimensions (10 negative/5 positive communication) (IRR: ICC = .69–.77)	No sig. effect of treatment on behavior in parent–child interaction

Studies are sorted by year of publication within each section

BM behavioral management, *BT* behavioral treatment, *comb* combination treatment, *CT* child training, *CRS* conflict rating system, *DB-DOS* Disruptive Behavior Diagnostic Observation Schedule, *DBD* disruptive behavior disorders, *DO* direct observation, *DPICS-R* Dyadic Parent–Child Interactive Coding System-Revised, *FT* family training, *GIPCI-R* Global Impressions of Parent–Child Interaction-Revised, *improved** improved significantly ($p < .05$), *MPH* methylphenidate, *MTA* Multimodal Treatment Study, *PAICS* Parent-Adolescent Interaction Coding System, *PT* parent training, *RS* rating scale, *TAU* treatment as usual, *TD* typically developing, *TR* transcript, *VT* videotape, *WL* waitlist

Parent and Adolescent Interaction Coding System (PAICS) The PAICS was applied in three studies with adolescents with ADHD [162, 163, 165]. The PAICS codes six behavior categories from transcribed discussions between adolescents and their parents. Typically, a 10-min neutral discussion about a vacation was followed by a 10-min discussion about conflicts. (1) Agreements between coders ranged between 53 and 85% [162, 163]. Two-week TRR was reported to be low [165] (although not specified numerically). (2) Between-group differences in negative communicative behavior between adolescents with and without ADHD were to a great extent accounted for by comorbid oppositional defiant behavior [162, 163]. The predictive validity and (3) convergent validity were not examined. (4) Changes in observed adolescent communicative behavior after different non-pharmacological interventions were not uniformly positive [165, 166].

Conflict Rating Scale (CRS) The CRS [167] was originally used to rate marital conflict interactions. (1) It was applied with adequate IRR (ICC = .64–.82) in two studies of Table 6 to rate 15 dimensions of positive and negative communication during parent–teen conflict and neutral discussions in samples of adolescents with ADHD and ODD [164, 166]. TRR was not examined for the CRS. (2) The ADHD/ODD group showed significantly more negative behavior and less positive behavior than comparison teens during the conflict discussion [164] (sensitivity/specificity were not assessed). (3) Convergent validity was not assessed. (4) No uniformly positive treatment effects were found on the CRS-rated teen behavior after completion of communication training [166].

Short Summary Six different observational tools for observations of parent–child interactions were used in a total of 11 studies (Table 6: CRS [n = 2], DB-DOS [n = 1], DPICS-R [n = 1], GIPCI [n = 2], MTA observation [n = 2], PAICS [n = 3]).

- IRR: mostly acceptable (ICC = .48–.97, r = .71–.84, kappa = .68, agreement = 53–81%). The lowest percentage agreement and kappa were reported for the PAICS [162, 163], the lowest ICC was reported for the GIPCI-R [160]. All studies reported IRR.
- TRR: reported for two instruments (DB-DOS [158], GIPCI-R [129, 160]), ranging between r = .20 and .50 for the GIPCI-R and between ICC = .52 and .80 for the DB-DOS.
- Correct classification: DB-DOS had 75% agreement with ADHD diagnosis [158].
- Convergent validity: reported for one instrument (DB-DOS); small to moderate agreements with parent ratings (r = .30–.42) and teacher ratings (r = .28–.32) [158].
- Treatment outcome: significant effects of non-pharmacological interventions were found using the DPICS-R [120] and the MTA observational tool [161] (but not in [159]); no significant treatment effects were found using the GIPCI-R [129, 160], the CRS [166], or the PAICS [165].

Peer–Child Interaction Observations

Five studies applied a specific observational tool for observing peer–child interactions (Table 7).

Test of Playfulness (ToP) The ToP [173] is an observer-rated scale to assess the construct of playfulness, consisting of 29 items. (1) In non-ADHD samples, evidence of acceptable IRR and TRR (ICC = .67) for the ToP was found [174, 175]. (2) Children with ADHD scored significantly lower on the overall playfulness measure in the laboratory [169] as well as in the naturalistic setting [168]. Sensitivity/specificity was not examined. (3) The convergent validity was not reported for the ToP. (4) A significantly improved ToP overall score was reported in children with ADHD after the completion of an intense play-based intervention compared to the pre-intervention baseline [172].

Discussion

This review sought to comprehensively cover the current state of systematic direct observational tools that are used in the study of ADHD. In total, 135 research findings from 29 different systematic observational tools, published between 1990 and 2016, were summarized in tables. We systematically delineated the reliability characteristics and the evidence of clinical validity for 16 observational instruments from the naturalistic setting, and for 13 instruments from the laboratory setting. A summary thereof is provided in Table 1.

Naturalistic Versus Laboratory Settings

We found considerably more research on systematic observational tools from naturalistic contexts (n = 83) than from standardized laboratory settings (n = 52). This imbalance might likely be attributed to the advantageous ecological validity of classroom observations.

In total, 55 out of 83 (66%) naturalistic observation studies and 30 out of 52 (58%) laboratory observation studies reported IRR. Enhanced objectivity and comparability in the laboratory minimizes the problem of low inter-rater agreement, which was a more particular problem of naturalistic observations.

TRR has been examined more frequently for laboratory tools (7 out of 13) than for naturalistic observational tools (4 out of 16) and coefficients were in a slightly higher range in analogue laboratory settings (e.g., test session: $r = .61–.86$) than in naturalistic settings (e.g., classroom: $r = .25–.77$). Playground, lunchroom, and parent–child interactions were the least stable situations for observation.

Classification rates seemed to be slightly higher for classroom observational tools than for laboratory observational tools. In general, group-level differences were more frequently analyzed than classification rates.

Significant treatment effects were found with both naturalistic and laboratory observational tools.

Which Tools to Use

Classroom

Based on the reviewed reliability and validity information, classroom observations should be preferred over other types of naturalistic observations. The BOSS, the COC, the COCADD, the DOF, and the SKAMP provide tools that are based on a number of independent studies and some psychometric validation. Nevertheless, each system has its advantages and disadvantages. Generalizability and dependability analyses have provided important information on the reliability of the BOSS and the DOF. Moreover, the DOF is the only tool that provides norms. The SKAMP has revealed good scores for TRR—although measured on the same day—[58], but low IRR. Even though an age-related decline in observable ADHD behavior may be assumed [91], the COCADD provided evidence for lasting observable behavioral differences and significant improvement with medication for adolescent patients with ADHD [89, 90].

Test Session

In the laboratory, more structured situations for observations, such as test sessions, were proven to discriminate better between ADHD and controls than independent play observations. Moreover, non-pharmacological interventions did not consistently cause a change in observed play behavior. Therefore, test session observations should be favored for studying ADHD behavior. The RAS and the TOF provide adequate tools for this purpose. The TOF has the advantage of providing norms. The RAS, however, is based on more evidence than the TOF. The RAS can be applied to observe behavior during academic seatwork and during CPTs. RAS variables were suggested to provide even better

Table 7 Peer–child interaction observation studies of children with ADHD ($n = 5$)

Author(s)	Age	N	Diagnosis	Control group	Intervention		Duration (min)	Observational instrument	Results
Group discrimination studies									
Leipold and Bundy [168]	5–14	50	ADHD	TD	–	DO RS	30–45	ToP 34 items on different components of playfulness (IRR: not reported)	Playfulness score: ADHD < TD*
Cordier et al. [169]	5–11	238	ADHD	TD	–	VT RS	20	ToP 29 items on different components of playfulness (IRR: not reported)	Playfulness score: ADHD < TD*
Cordier et al. [170]	5–11	238	ADHD	TD	–	VT RS	20	ToP 29 items on different components of playfulness (IRR: not reported)	Two of 29 items: ADHD < playmates*
Cordier et al. [171]	5–11	105	ADHD subtypes	None	–	VT RS	20	ToP 29 items on different components of playfulness (IRR: not reported)	Engaged play, playful mischief, clowning: ADHD-IN < ADHD-C/ADHD-HI*
Non-pharmacological intervention studies									
Wilkes et al. [172]	5–11	30	ADHD	TD	Play-based intervention	VT RS	2×20	ToP 29 items on different components of playfulness (IRR: not reported)	Play-based intervention improved* overall ToP score

Studies are sorted by year of publication within each section

DO direct observation, *improved** improved significantly ($p < .05$), *RS* rating-scale, *TD* typically developing, *ToP* Test of Playfulness, *VT* videotape discrimination between ADHD and controls than actual task performance [6, 133, 176]. Nonetheless, problems of the RAS lie in the low IRR coefficients that have been occasionally reported (e.g., [140, 142, 143]) and in the fact that group differences were not uniformly found in the same RAS variables. Furthermore, adolescents with ADHD could not be distinguished from clinically referred participants [152], which calls into question the specificity of the RAS behaviors for ADHD. However, effect sizes to detect stimulant-induced change were larger for RAS measures than for parent and teacher ratings [148].

Laboratory Interactions

Parent–child interactions were found to be rather unstable and results were mixed regarding the sensitivity to change of parent–child interaction observational tools. The use of these tools as treatment outcome measures is compounded by the possible difficulty of disentangling effects on parenting from effects on child behavior. This interdependency may also be responsible for the low stability. Furthermore, it must be kept in mind that parent-adolescent interactions seem to provide a measure of ODD symptoms rather than of ADHD [162, 163]. Nonetheless, for adolescents, the CRS and the PAICS are likely to be useful, while for younger children, all reviewed tools seem to have comparable utility. Interactions with a non-familiar adult could provide a more controllable alternative for highly unstable parent–child interaction observations, as the DB-DOS experimenter contexts reached better reliability coefficients (IRR, TRR, Cronbach's alpha) than the DB-DOS parent context [158] (see also [177, 178]). The evidence base for peer–child observations with the ToP in ADHD is not sufficiently established.

General Methodological Issues and Suggestions for Future Research

The present review revealed several issues to be resolved in future research. First, all studies need to formally assess and report IRR and to provide adequate training for observers. In particular, more consistent reporting of IRR should be aimed at for the SKAMP, in view of the frequent use of this observational scale in medication trials. For time-sampling procedures, kappa coefficients should be preferred over percentage agreement for the analysis of IRR. In particular, the RAS lacks reports of kappa coefficients.

Second, the TRR should be assessed more consistently. Crucially, stability of behavior should be investigated within ADHD groups separately, because evidence strongly suggests increased behavioral variability in ADHD [179, 180]. In addition, naturalistic settings are particularly vulnerable to the impact of uncontrollable contextual factors. Influences such as the time of day, the academic subject, the teaching method, or even the time in the school year create potential biases to the reliability of observed behavior [39, 46, 88].

There is a particular lack of reports on the convergent validity for play observations and parent–child interaction observations. Otherwise, agreements with parent and teacher reports of ADHD symptoms are typically small to moderate and classification rates hardly exceeded 80%. Therefore, we conclude that none of the reviewed observational instruments can be applied as a stand-alone diagnostic procedure. Analyses on the incremental validity of observational tools revealed negligible contributions to the prediction of ADHD or functional impairment over and above parent and teacher reports [51, 139]. However, a problem of circularity compounds the predictive validity of observations because the diagnostic criteria of ADHD are primarily based on parent and teacher interviews [138], and not on an objective, absolute quantification of observable behavior. Therefore, it may be rather challenging to obtain objective measures that reach comparable clinical validity in ADHD diagnosis to parent and teacher ratings. Moderate degrees of agreement suggest that behavioral observations target unique aspects of problematic behavior that are not covered by parent and teacher reports alone. Observational data might therefore aid the interpretation of inconsistencies between different sources of ADHD ratings. However, systematic behavioral observations cannot act as a substitute for parent and teacher ratings.

Although the evidence indicates that observational methods are not appropriate for diagnosing ADHD when applied as a stand-alone approach, this does not preclude their potential value for assessing treatment outcome [11]. Significant improvements in observed behavior after treatment have been reported for most tools (see Table 1). It is debatable whether this is sufficient to assume treatment sensitivity for these methods (see [181]). Clearly, study designs that lack observations in an untreated control group (e.g., [84, 172]) or observation of pre-treatment behavior (e.g., [83]) should be avoided. Dependability studies suggest that pre-post designs with one observation each may not be sufficiently reliable to monitor treatment effects on classroom behavior [94, 95]. Normalization rates and the clinical significance of change should be more consistently reported and norms should be established. Furthermore, common instruments used for the evaluation of behavioral treatments (i.e., BOSS) should be validated with regard to pharmacological effects and vice versa (i.e., SKAMP). Observational methods have earned the reputation of being the gold standard for treatment evaluation [14]. Based on the present review, however, this presumption might have to be reconsidered. Issues of low temporal stability, considerable behavioral variability, unknown TRR of most instruments, and the lack of normative data impede the thorough evaluation of the sensitivity to change of these methods.

Observee reactivity poses a further problem for behavioral observations [182–184]. This phenomenon occurs if behavior is altered due to the awareness of being observed [182, 184]. Efforts should therefore be made to conduct observations as unobtrusively as possible [183] and studies should consistently specify how the presence of observers or video cameras was explained to the participants. First-grade students were not found to show reactivity to observers in the classroom [185]. Similar investigations would be necessary to ascertain the impact of observer reactivity in ADHD samples and adolescents.

In general, adolescent ADHD patients seem to be underrepresented in studies using observational methods (13 out of 135 studies [10%]). Validation of instruments with a specific focus on this age group would be highly desirable.

The confounding effect of comorbidity was not sufficiently addressed in many of the reviewed studies. Comorbid disruptive behavior disorders augmented the levels of observed dependent measures in some cases (e.g., [36, 111, 162]), but not all (e.g., [107, 158]). Not only the influence of comorbid disorders, but also the differentiation between different psychopathological groups needs to be more systematically analyzed in the future. Behaviors such as classroom aggression or noncompliance clearly overlap with symptoms of other externalizing behavior disorders such as ODD or CD. Moreover, students with learning disabilities were also found to exhibit elevated levels of off-task behavior and disruptive behavior in the classroom (for review see [186]), and children with ASD displayed high amounts of out-of-seat behavior [187]. Hence, the specificity of such behaviors for ADHD is questionable and needs to be taken into account when interpreting observational data.

Based on these considerations, we recommend considering the following critical factors for planning and conducting systematic behavioral observations:

(1) Is satisfactory IRR established through observer training?
(2) Is IRR formally assessed and reported? Is the kappa coefficient indicated if a standardized sampling procedure is applied?
(3) Are observers blind with regard to the subject status and the treatment condition?
(4) Is observee reactivity controlled for as effectively as possible?
(5) Is a sufficient amount and duration of observation episodes assured? Are dependability studies taken into account (see [94, 95])?
(6) Are situational influences controlled for (e.g., time of day, school subject, teacher)?
(7) Is an adequate control group included that is observed with the same intensity and frequency as the experimental group?
(8) Is the clinical significance of behavioral change (i.e., normalization) evaluated?
(9) Are other measures of ADHD symptoms included? To what extent do these reflect the observational findings?

Summary

This review evaluated the clinical utility of observational methods in the research in children and adolescents with ADHD. Twenty-nine instruments for observing classroom behavior (11 tools), naturalistic social interactions (5 tools), independent play (2 tools), test session behavior (4 tools), parent–child interactions (6 tools), or peer interactions (1 tool) were reviewed. Tools for classroom and test session observations showed the most promising psychometric properties. The RAS and the TOF may be recommended for test session observations. The BOSS, the COC, the COCADD, the DOF, and the SKAMP seem to be reasonable choices for the study of ADHD classroom behavior. However, the psychometric properties of all of these instruments need more systematic validation.

Future research should intensify the investigation of the discriminative validity of observational measures with regard to different comorbid groups, other psychiatric disorders (e.g., learning disorder, ASD), and clinically referred groups. The incremental validity of observations over other diagnostic methods should be assessed more consistently and efforts should be made to obtain normative data for observational instruments. Furthermore, many observational instruments lack a report of TRR and/or dependability. Treatment-related changes in observed behavior should be cross-validated with other instruments and the concurrent validity between different observational tools should be established.

Acknowledgements This research was supported by Grant 320030_149411 from the Swiss National Science Foundation.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Wittchen HU, Jacobi F, Rehm J et al (2011) The size and burden of mental disorders and other disorders of the brain in Europe 2010. Eur Neuropsychopharmacol 21:655–679
2. Pelham WE Jr, Fabiano GA, Massetti GM (2005) Evidence-based assessment of attention deficit hyperactivity disorder in children and adolescents. J Clin Child Adolesc Psychol 34:449–476

3. Brock SE, Clinton A (2007) Diagnosis of attention-deficit/hyperactivity disorder (AD/HD) in childhood: A review of the literature. Calif Sch Psychol 12:73–91
4. Handler MW, DuPaul GJ (2005) Assessment of ADHD: differences across psychology specialty areas. J Atten Disord 9:402–412
5. Platzman KA, Stoy MR, Brown RT, Coles CD, Smith IE, Falek A (1992) Review of observational methods in attention deficit hyperactivity disorder (ADHD): implications for diagnosis. Sch Psychol Q 7:155–177
6. Barkley RA (1991) The ecological validity of laboratory and analogue assessment methods of ADHD symptoms. J Abnorm Child Psychol 19:149–178
7. American Psychiatric Association (2013) Diagnostic and statistical manual of mental disorders, 5th edn. Author, London
8. Abikoff H, Courtney M, Pelham WE Jr, Koplewicz HS (1993) Teachers' ratings of disruptive behaviors: The influence of halo effects. J Abnorm Child Psychol 21:519–533
9. Harper GW, Ottinger DR (1992) The performance of hyperactive and control preschoolers on a new computerized measure of visual vigilance: the preschool vigilance task. J Child Psychol Psychiatry 33:1365–1372
10. Hosterman SJ, DuPaul GJ, Jitendra AK (2008) Teacher ratings of ADHD symptoms in ethnic minority students: Bias or behavioral difference? Sch Psychol Q 23:418–435
11. Rapport MD, Chung K-M, Shore G, Denney CB, Isaacs P (2000) Upgrading the science and technology of assessment and diagnosis: Laboratory and clinic-based assessment of children with ADHD. J Clin Child Psychol 29:555–568
12. Matier-Sharma K, Perachia N, Newcorn JH, Sharma V, Halperin JM (1995) Differential diagnosis of ADHD: Are objective measures of attention, impulsivity, and activity level helpful? Child Neuropsychol 1:118–127
13. Demaray MK, Schaefer K, Delong LK (2003) Attention-deficit/hyperactivity disorder (ADHD): A national survey of training and current assessment practices in the schools. Psychol Sch 40:583–597
14. Pelham WE, Fabiano GA, Waxmonsky JG et al (2016) Treatment sequencing for childhood ADHD: a multiple-randomization study of adaptive medication and behavioral interventions. J Clin Child Adolesc Psychol 45:1–20
15. Hintze JM, Volpe RJ, Shapiro ES (2002) Best practices in the systematic direct observation of student behavior. Best Pract Sch Psychol 4:993–1006
16. Hintze JM (2005) Psychometrics of direct observation. Sch Psychol Rev 34:507–519
17. Salvia J, Ysseldyke J, Witmer S (2012) Assessment: in special and inclusive education. Cengage Learning, Wadsworth
18. Thompson T, Symons FJ, Felce D (2000) Principles of behavioral observation: assumptions and strategies. In: Thompson T, Felce D, Symons FJ (eds) Behavioral observation: technology and applications in developmental disabilities. Paul H. Brookes, Baltimore, pp 3–16
19. Williams White S, Keonig K, Scahill L (2007) Social skills development in children with autism spectrum disorders : a review of the intervention research. J Autism Dev Disord 37:1858–1868
20. Lord C, Risi S, Lambrecht L, Cook EH, Leventhal BL, Dilavore PC, Pickles A, Rutter M (2000) The autism diagnostic observation schedule—generic : a standard measure of social and communication deficits associated with the spectrum of autism. J Autism Dev Disord 30:205–223
21. Dawson G, Toth K, Abbott R, Osterling J, Munson J, Estes A, Liaw J (2004) Early social attention impairments in autism: social orienting, joint attention, and attention to distress. Dev Psychol 40:271–283
22. Pasalich DS, Witkiewitz K, Mcmahon RJ, Pinderhughes EE, Group TCPPR. (2016) Indirect effects of the fast track intervention on conduct disorder symptoms and callous-unemotional traits: distinct pathways involving discipline and warmth. J Abnorm Child Psychol 44:587–597
23. Kempes M, Matthys W, de Vries H, van Engeland H (2005) Reactive and proactive aggression in children: a review of theory, findings and the relevance for child and adolescent psychiatry. Eur Child Adolesc Psychiatry 14:11–19
24. Webster-Stratton C, Reid MJ, Hammond M (2004) Treating children with early-onset conduct problems: Intervention outcomes for parent, child, and teacher training. J Clin Child Adolesc Psychol 33:105–124
25. Hirshfeld-Becker DR, Masek B, Henin A et al (2010) Cognitive behavioral therapy for 4- to 7-year-old children with anxiety disorders : a randomized clinical trial. J Consult Clin Psychol 78:498–510
26. Mian ND, Carter AS, Pine DS, Wakschlag LS, Briggs-Gowan MJ (2015) Development of a novel observational measure for anxiety in young children: the Anxiety Dimensional Observation Scale. J Child Psychol Psychiatry 56:1017–1025
27. McConaughy SH, Achenbach TM (2009) Manual for the ASEBA direct observation form. University of Vermont, Research Center for Children, Youth, & Families, Burlington
28. McConaughy SH, Achenbach TM (2004) Manual for the test observation form for age 2–18. University of Vermont, Research Center for Children, Youth, & Families, Burlington
29. Volpe RJ, DiPerna JC, Hintze JM, Shapiro ES (2005) Observing students in classroom settings: a review of seven coding schemes. Sch Psychol Rev 34:454–474
30. Murillo GL, Cortese S, Anderson D, Di Martino A, Castellanos FX (2015) Locomotor activity measures in the diagnosis of attention deficit hyperactivity disorder: meta-analyses and new findings. J Neurosci Methods 252:14–26
31. Hall CL, Valentine AZ, Groom MJ, Walker GM, Sayal K, Daley D, Hollis C (2016) The clinical utility of the continuous performance test and objective measures of activity for diagnosing and monitoring ADHD in children: a systematic review. Eur Child Adolesc Psychiatry 25:677–699
32. Green BC, Johnson KA, Bretherton L (2014) Pragmatic language difficulties in children with hyperactivity and attention problems: an integrated review. Int J Lang Commun Disord 49:15–29
33. Patros CHG, Alderson RM, Kasper LJ, Tarle SJ, Lea SE, Hudec KL (2015) Choice-impulsivity in children and adolescents with attention-deficit/hyperactivity disorder (ADHD): a meta-analytic review. Clin Psychol Rev 43:162–174
34. Lett NJ, Kamphaus RW (1997) Differential validity of the BASC student observation system and the BASC teacher rating scale. Can J Sch Psychol 13:1–14
35. Skansgaard EP, Burns GL (1998) Comparison of DSM-IV ADHD combined and predominantly inattention types: correspondence between teacher ratings and direct observations of inattentive, hyperactivity/impulsivity, slow cognitive tempo, oppositional defiant, and overt conduct disorder sympto. Child Fam Behav Ther 20:1–14
36. Abikoff HB, Jensen PS, Arnold LLE et al (2002) Observed classroom behavior of children with ADHD: Relationship to gender and comorbidity. J Abnorm Child Psychol 30:349–359
37. DuPaul GJ, Volpe RJ, Jitendra AK, Lutz JG, Lorah KS, Gruber R (2004) Elementary school students with AD/HD: predictors of academic achievement. J Sch Psychol 42:285–301
38. Antrop I, Buysse A, Roeyers H, Van Oost P (2005) Activity in children with ADHD during waiting situations in the classroom: a pilot study. Br J Educ Psychol 75:51–69

39. Antrop I, Roeyers H, De Baecke L (2005) Effects of time of day on classroom behaviour in children with ADHD. Sch Psychol Int 26:29–43
40. DuPaul GJ, Jitendra AK, Tresco KE, Junod REV, Volpe RJ, Lutz JG (2006) Children with attention deficit hyperactivity disorder: are there gender differences in school functioning ? Sch Psychol Rev 35:292–308
41. Vile Junod RE, DuPaul GJ, Jitendra AK, Volpe RJ, Cleary KS (2006) Classroom observations of students with and without ADHD: Differences across types of engagement. J Sch Psychol 44:87–104
42. Lauth GW, Heubeck BG, Mackowiak K (2006) Observation of children with attention-deficit hyperactivity (ADHD) problems in three natural classroom contexts. Br J Educ Psychol 76:385–404
43. McConaughy SH, Ivanonva MY, Antshel K, Eiraldi RB, Dumenci L (2009) Standardized observational assessment of attention deficit hyperactivity disorder combined and predominantly inattentive subtypes. II. Classroom observations. Sch Psych Rev 38:362–381
44. Hart KC, Massetti GM, Fabiano GA, Pariseau ME, Pelham WE Jr (2011) Impact of group size on classroom on-task behavior and work productivity in children with ADHD. J Emot Behav Disord 19:55–64
45. Imeraj L, Antrop I, Roeyers H, Deboutte D, Deschepper E, Bal S, Sonuga-Barke EJS (2013) The impact of idle time in the classroom: Differential effects on children with ADHD. J Atten Disord 20:71–81
46. Imeraj L, Antrop I, Sonuga-Barke EJS, 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. J Sch Psychol 51:487–498
47. Steiner NJ, Sheldrick RC, Frenette EC, Rene KM, Perrin EC (2014) Classroom behavior of participants with ADHD compared with peers: Influence of teaching format and grade level. J Appl Sch Psychol 30:209–222
48. Nolan EE, Gadow KD (1994) Relation between ratings and observations of stimulant drug response in hyperactive children. J Clin Child Psychol 23:78–90
49. Solanto MV, Abikoff H, Sonuga-Barke EJS, Schachar R, Logan GD, Wigal T, Hechtman L, Hinshaw SP, Turkel E (2001) The ecological validity of delay aversion and response inhibition as measures of impulsivity in AD/HD: a supplement to the NIMH multimodal treatment study of AD/HD. J Abnorm Child Psychol 29:215–228
50. Sonuga-Barke EJS, Coghill D, DeBacker M, Swanson J (2009) Measuring methylphenidate response in attention-deficit/hyperactvity disorder: how are laboratory classroom-based measures related to parent ratings? J Child Adolesc Psychopharmacol 19:691–698
51. McConaughy SH, Harder VS, Antshel KM, Gordon M, Eiraldi R, Dumenci L (2010) Incremental validity of test session and classroom observations in a multimethod assessment of attention deficit/hyperactivity disorder. J Clin Child Adolesc Psychol 39:650–666
52. Gadow KD, Nolan EE, Sverd J, Sprafkin J, Paolicelli L (1990) Methylphenidate in aggressive-hyperactive boys: I. Effects on peer aggression in public school settings. J Am Acad Child Adolesc Psychiatry 29:710–718
53. Carlson CL, Pelham WE Jr, Milich R, Dixon J (1992) Single and combined effects of methylphenidate and behavior therapy on classroom performance of children with attention-deficit hyperactivity disorder. J Abnorm Child Psychol 20:213–232
54. Pelham WE Jr, Carlson C, Sams SE, Vallano G, Dixon MJ, Hoza B (1993) Separate and combined effects of methylphenidate and behavior modification on boys with attention deficit-hyperactivity disorder in the classroom. J Consult Clin Psychol 61:506–515
55. Gadow KD, Nolan EE, Sprafkin J, Sverd J (1995) School observations of children with attention-deficit hyperactivity disorder and comorbid tic disorder: effects of methylphenidate treatment. J Dev Behav Pediatr 16:167–176
56. Klein RG, Abikoff H (1997) Behavior therapy and methylphenidate in the treatment of children with ADHD. J Atten Disord 2:89–114
57. Nolan EE, Gadow KD (1997) Children with ADHD and tic disorder and their classmates: behavioral normalization with methylphenidate. J Am Acad Child Adolesc Psychiatry 36:597–604
58. Wigal SB, Gupta S, Guinta D, Swanson JM (1998) Reliability and validity of the SKAMP rating scale in a laboratory school setting. Psychopharmacol Bull 34:47–53
59. Swanson JM, Wigal S, Greenhill LL et al (1998) Analog classroom assessment of Adderall in children with ADHD. J Am Acad Child Adolesc Psychiatry 37:519–526
60. Pelham WE Jr, Aronoff HR, Midlam JK, Shapiro CJ, Gnagy EM, Chronis AM, Onyango AN, 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
61. Pelham WE Jr, Gnagy EM, Chronis AM, Burrows-MacLean L, Fabiano GA, Onyango AN, Meichenbaum DL, Williams A, Aronoff HR, Steiner RL (1999) A comparison of morning-only and morning/late afternoon Adderall to morning-only, twice-daily, and three times-daily methylphenidate in children with attention-deficit/hyperactivity disorder. Pediatrics 104:1300–1311
62. Pelham WE Jr, Gnagy EM, Burrows-Maclean L et al (2001) Once-a-day Concerta methylphenidate versus three-times-daily methylphenidate in laboratory and natural settings. Pediatrics 107:1–15
63. Swanson JM, Gupta S, Williams L, Agler D, Lerner M, Wigal S (2002) Efficacy of a new pattern of delivery of methylphenidate for the treatment of ADHD: effects on activity level in the classroom and on the playground. J Am Acad Child Adolesc Psychiatry 41:1306–1314
64. Greenhill LL, Swanson JM, Steinhoff K, Fried J, Posner K, Lerner M, Wigal S, Clausen SB, Zhan Y, Tulloch S (2003) A pharmacokinetic/pharmacodynamic study comparing a single morning dose of Adderall to twice-daily dosing in children with ADHD. J Am Acad Child Adolesc Psychiatry 42:1234–1241
65. Lopez F, Silva R, Pestreich L, Muniz R (2003) Comparative efficacy of two once daily methylphenidate formulations and placebo in children with attention deficit hyperactivity disorder across the school day. Pediatr Drugs 5:545–555
66. McCracken JT, Biederman J, Greenhill LL et al (2003) Analog classroom assessment of a once-daily mixed amphetamine formulation, SLI381 (Adderall XR), in children with ADHD. J Am Acad Child Adolesc Psychiatry 42:673–683
67. Abikoff H, Hechtman L, Klein RG, Weiss G, Fleiss K, Etcovitch J, Cousins L, Greenfield B, Martin D, Pollack S (2004) Symptomatic improvement in children with ADHD treated with long-term methylphenidate and multimodal psychosocial treatment. J Am Acad Child Adolesc Psychiatry 43:802–811
68. Döpfner M, Gerber WD, Banaschewski T et al (2004) Comparative efficacy of once-a-day extended-release methylphenidate, two-times-daily immediate-release methylphenidate, and placebo in a laboratory school setting. Eur Child Adolesc Psychiatry 13:i93–i101
69. Swanson JM, Wigal SB, Wigal T et al (2004) A comparison of once-daily extended-release methylphenidate formulations in children with attention-deficit/hyperactivity disorder in the laboratory (the Comacs Study). Pediatrics 113:e206–e216

70. Biederman J, Wigal SB, Spencer TJ, McGough JJ, Mays DA (2006) A post hoc subgroup analysis of an 18-day randomized controlled trial comparing the tolerability and efficacy of mixed amphetamine salts extended release and atomoxetine in school-age girls with attention-deficit/hyperactivity disorder. Clin Ther 28:280–293
71. McGough JJ, Wigal SB, Abikoff H, Turnbow JM, Posner K, Moon E (2006) A randomized, double-blind, placebo-controlled, laboratory classroom assessment of methylphenidate transdermal system in children with ADHD. J Atten Disord 9:476–485
72. Silva RR, Muniz R, Pestreich LK, Brams M, Childress A, Lopez FA (2005) Efficacy of two long-acting methylphenidate formulations in children with attention- deficit/hyperactivity disorder in a laboratory classroom setting. J Child Adolesc Psychopharmacol 15:637–654
73. Silva RR, Muniz R, Pestreich L, Childress A, Brams M, Lopez FA, Wang J (2006) Efficacy and duration of effect of extended-release dexmethylphenidate versus placebo in schoolchildren with attention-deficit/hyperactivity disorder. J Child Adolesc Psychopharmacol 16:239–251
74. Brams M, Muniz R, Childress A, Giblin J, Mao A, Turnbow J, Borrello M, McCague K, Lopez FA, Silva R (2008) A randomized, double-blind, crossover study of once-daily dexmethylphenidate in children with attention-deficit hyperactivity disorder. CNS Drugs 22:693–704
75. Muniz R, Brams M, Mao A, McCague K, Pestreich L, Silva R (2008) Efficacy and safety of extended-release dexmethylphenidate compared with d,l-methylphenidate and placebo in the treatment of children with attention-deficit/hyperactivity disorder: a 12-hour laboratory classroom study. J Child Adolesc Psychopharmacol 18:248–256
76. Schulz E, Fleischhaker C, Hennighausen K et al (2010) A double-blind, randomized, placebo/active controlled crossover evaluation of the efficacy and safety of Ritalin LA in children with attention-deficit/hyperactivity disorder in a laboratory classroom setting. J Child Adolesc Psychopharmacol 20:377–385
77. Brams M, Turnbow J, Pestreich L, Giblin J, Childress A, McCague K, Muniz R (2012) A randomized, double-blind study of 30 versus 20 mg dexmethylphenidate extended-release in children with attention-deficit/hyperactivity disorder: late-day symptom control. J Clin Psychopharmacol 32:637–644
78. Wigal SB, Greenhill LL, Nordbrock E, Connor DF, Kollins SH, Adjei A, Childress A, Stehli A, Kupper RJ (2014) A randomized placebo-controlled double-blind study evaluating the time course of response to methylphenidate hydrochloride extended-release capsules in children with attention-deficit/hyperactivity disorder. J Child Adolesc Psychopharmacol 24:562–569
79. Williamson D, Murray DW, Damaraju CV, Ascher S, Starr HL (2014) Methylphenidate in children with ADHD with or without learning disability. J Atten Disord 18:95–104
80. Childress AC, Brams M, Cutler AJ, Kollins SH, Northcutt J, Padilla A, Turnbow JM (2015) The efficacy and safety of Evekeo, racemic amphetamine sulfate, for treatment of attention-deficit/hyperactivity disorder symptoms: a multicenter, dose-optimized, double-blind, randomized, placebo-controlled crossover laboratory classroom study. J Child Adolesc Psychopharmacol 25:402–414
81. Manos MJ, Caserta DA, Short EJ, Raleigh KL, Giuliano KC, Pucci NC, Frazier TW (2015) Evaluation of the duration of action and comparative effectiveness of lisdexamfetamine dimesylate and behavioral treatment in youth with ADHD in a quasi-naturalistic setting. J Atten Disord 19:578–590
82. DuPaul GJ, Ervin RA, Hook CL, McGoey KE (1998) Peer tutoring for children with attention deficit hyperactivity disorder: Effects on classroom behavior and academic performance. J Appl Behav Anal 31:579–592
83. Barkley RA, Shelton TL, Crosswait C, Moorehouse M, Fletcher K, Barrett S, Jenkins L, Metevia L (2000) Multi-method psycho-educational intervention for preschool children with disruptive behavior: preliminary results at post-treatment. J Child Psychol Psychiatry 41:319–332
84. Pfiffner LJ, Villodas M, Kaiser N, Rooney M, McBurnett K (2013) Educational outcomes of a collaborative school-home behavioral intervention for ADHD. Sch Psychol Q 28:25–36
85. Steiner NJ, Frenette EC, Rene KM, Brennan RT, Perrin EC (2014) Neurofeedback and cognitive attention training for children with attention-deficit hyperactivity disorder in schools. J Dev Behav Pediatr 35:18–27
86. Steiner NJ, Frenette EC, Rene KM, Brennan RT, Perrin EC (2014) In-school neurofeedback training for ADHD: sustained improvements from a randomized control trial. Pediatrics 133:483–492
87. Carroll A, Houghton S, Taylor M, West J, List-Kerz M (2006) Responses to interpersonal and physically provoking situations: the utility and application of an observation schedule for school-aged students with and without attention deficit/hyperactivity disorder. Educ Psychol 26:483–498
88. Evans SW, Allen J, Moore S, Strauss V (2005) Measuring symptoms and functioning of youth with ADHD in middle schools. J Abnorm Child Psychol 33:695–706
89. Evans SW, Pelham WE Jr (1991) Psychostimulant effects on academic and behavioral measures for ADHD junior high school students in a lecture format classroom. J Abnorm Child Psychol 19:537–552
90. Evans SW, Pelham WE Jr, Smith BH, Bukstein O, Gnagy EM, Greiner AR, Altenderfer L, Baron-Myak C (2001) Dose-response effects of methylphenidate on ecologically valid measures of academic performance and classroom behavior in adolescents with ADHD. Exp Clin Psychopharmacol 9:163–175
91. Abikoff H, Gittelman R, Klein DF (1980) Classroom observation code for hyperactive children: a replication of validity. J Consult Clin Psychol 48:555–565
92. Gadow KD, Sprafkin J, Nolan EE (1996) ADHD school observation code. Checkmate Plus, Stony Brook
93. Shapiro ES (2004) Academic skills problems workbook (rev.). Guilford Press, New York
94. Ferguson TD, Briesch AM, Volpe RJ, Daniels B (2012) The influence of observation length on the dependability of data. Sch Psychol Q 27:187–197
95. Volpe RJ, McConaughy SH, Hintze JM (2009) Generalizability of classroom behavior problem and on-task scores from the direct observation form. School Psych Rev 38:382–401
96. Atkins MS, Pelham WE Jr, Licht MH (1985) A comparison of objective classroom measures and teacher ratings of attention deficit disorder. J Abnorm Child Psychol 13:155–167
97. Atkins MS, Pelham WE Jr, Licht MH (1989) The differential validity of teacher ratings of inattention/overactivity and aggression. J Abnorm Child Psychol 17:423–435
98. Swanson JM (1992) School based assessments and interventions for ADD students. K.C. Publishing, Irvine
99. Murray DW, Bussing R, Fernandez M, Hou W, Wilson Garvan C, Swanson JM, Eyberg SM (2009) Psychometric properties of teacher SKAMP ratings from a community sample. Assessment 16:193–208
100. Erhardt D, Hinshaw SP (1994) Initial sociometric impressions of attention-deficit hyperactivity disorder and comparison boys: predictions from social behaviors and from nonbehavioral variables. J Consult Clin Psychol 62:833–842
101. Anderson CA, Hinshaw SP, Simmel C (1994) Mother-child interactions in ADHD and comparison boys: Relationships with overt and covert externalizing behavior. J Abnorm Child Psychol 22:247–265

102. Hinshaw SP, Simmel C, Heller TL (1995) Multimethod assessment of covert antisocial behavior in children: Laboratory observations, adult ratings, and child self-report. Psychol Assess 7:209–219
103. Hinshaw SP, Zupan BP, Simmel C, Nigg JT, Melnick S (1997) Peer status in boys with and without attention-deficit hyperactivity disorder: predictions from overt and covert antisocial behavior, social isolation, and authoritative parenting beliefs. Child Dev 68:880–896
104. DuPaul GJ, McGoey KE, Eckert TL, VanBrakle J (2001) Preschool children with attention-deficit/hyperactivity disorder: impairments in behavioral, social, and school functioning. J Am Acad Child Adolesc Psychiatry 40:508–515
105. Mikami AY, Hinshaw SP (2003) Buffers of peer rejection among girls with and without ADHD: The role of popularity with adults and goal-directed solitary play. J Abnorm Child Psychol 31:381–397
106. Riley C, DuPaul GJ, Pipan M, Kern L, Van Brakle J, Blum NJ (2008) Combined type versus ADHD predominantly hyperactive-impulsive type: is there a difference in functional impairment? J Dev Behav Pediatr 29:270–275
107. Pollack B, Hojnoski R, DuPaul GJ, Kern L (2015) Play behavior differences among preschoolers with ADHD: Impact of comorbid ODD and anxiety. J Psychopathol Behav Assess 38:66–75
108. Nigg JT, Hinshaw SP, Halperin JM (1996) Continuous performance test in boys with attention deficit hyperactivity disorder: Methylphenidate dose response and relations with observed behaviors. J Clin Child Psychol 25:330–340
109. Thomas LB, Shapiro ES, DuPaul GJ, Lutz JG, Kern L (2011) Predictors of social skills for preschool children at risk for ADHD: The relationship between direct and indirect measurements. J Psychoeduc Assess 29:114–124
110. Pelham WE Jr, Greenslade KE, Vodde-Hamilton M, Murphy DA, Greenstein JJ, Gnagy EM, Guthrie KJ, Hoover MD, Dahl RE (1990) Relative efficacy of long-acting stimulants on children with attention deficit-hyperactivity disorder: a comparison of standard methylphenidate, sustained-release methylphenidate, sustained-release dextroamphetamine, and pemoline. Pediatrics 86:226–237
111. Murphy DA, Pelham WE Jr, Lang AR (1992) Aggression in boys with attention deficit-hyperactivity disorder: Methylphenidate effects on naturalistically observed aggression, response to provocation, and social information processing. J Abnorm Child Psychol 20:451–466
112. Pelham WE Jr, Gnagy EM, Greiner AR et al (2000) Behavioural versus behavioural and pharmacological treatment in ADHD children attending a summer treatment program. J Abnorm Child Psychol 28:507–525
113. Pelham WE Jr, Hoza B, Pillow DR et al (2002) Effects of methylphenidate and expectancy on children with ADHD: behavior, academic performance, and attributions in a summer treatment program and regular classroom settings. J Consult Clin Psychol 70:320–335
114. Chacko A, Pelham WE Jr, Gnagy EM, Greiner A, Vallano G, Bukstein O, Rancurello M (2005) Stimulant medication effects in a summer treatment program among young children with attention-deficit/hyperactivity disorder. J Am Acad Child Adolesc Psychiatry 44:249–257
115. Pelham WE Jr, Burrows-MacLean L, Gnagy EM et al (2005) Transdermal methylphenidate, behavioral, and combined treatment for children with ADHD. Exp Clin Psychopharmacol 13:111–126
116. Pelham WE Jr, Manos MJ, Ezzell CE et al (2005) A dose-ranging study of a methylphenidate transdermal system in children with ADHD. J Am Acad Child Adolesc Psychiatry 44:522–529
117. Chronis AM, Fabiano GA, Gnagy EM, Onyango AN, Pelham WE Jr, Lopez-Williams A, Chacko A, Wymbs BT, Coles EK, Seymour KE (2004) An evaluation of the summer treatment program for children with attention-deficit/hyperactivity disorder using a treatment withdrawal design. Behav Ther 35:561–585
118. Fabiano GA, Pelham WE Jr, Manos MJ et al (2004) An evaluation of three time-out procedures for children with attention-deficit/hyperactivity disorder. Behav Ther 35:449–469
119. Mrug S, Hoza B, Pelham WE Jr, Gnagy EM, Greiner AR (2007) Behavior and peer status in children with ADHD: continuity and change. J Atten Disord 10:359–371
120. Webster-Stratton CH, Reid MJ, Beauchaine T (2011) Combining parent and child training for young children with ADHD. J Clin Child Adolesc Psychol 40:191–203
121. Walker HM, Severson HH, Feil EG (1995) Early screening project: a proven child find process. Sopris West, Longmont
122. Roberts MA (1990) A behavioral observation method for differentiating hyperactive and aggressive boys. J Abnorm Child Psychol 18:131–142
123. Paternite CE, Loney J, Roberts MA (1996) A preliminary validation of subtypes of DSM-IV attention-deficit/hyperactivity disorder. J Atten Disord 1:70–86
124. Byrne JM, DeWolfe NA, Bawden HN (1998) Assessment of attention-deficit hyperactivity disorder in preschoolers. Child Neuropsychol 4:49–66
125. Handen BL, McAuliffe S, Janosky J, Feldman H, Breaux AM (1998) A playroom observation procedure to assess children with mental retardation and ADHD. J Abnorm Child Psychol 26:269–277
126. DeWolfe NA, Byrne JM, Bawden HN (2000) Preschool inattention and impulsivity-haperactivity: development of a clinic-based assessment protocol. J Atten Disord 4:80–90
127. Handen BL, McAuliffe S, Janosky J, Feldman H, Breaux AM (1995) Methylphenidate in children with mental retardation and ADHD: Effects on independent play and academic functioning. J Dev Phys Disabil 7:91–103
128. Sonuga-Barke EJS, Daley D, Thompson M, Laver-Bradbury C, Weeks A (2001) Parent-based therapies for preschool attention-deficit/hyperactivity disorder: a randomized, controlled trial with a community sample. J Am Acad Child Adolesc Psychiatry 40:402–408
129. Daley D, O'Brien M (2013) A small-scale randomized controlled trial of the self-help version of the New Forest Parent Training Programme for children with ADHD symptoms. Eur Child Adolesc Psychiatry 22:543–552
130. Abikoff HB, Thompson M, Laver-Bradbury C, Long N, Forehand RL, Miller Brotman L, Klein RG, Reiss P, Huo L, Sonuga-Barke EJS (2015) Parent training for preschool ADHD: a randomized controlled trial of specialized and generic programs. J Child Psychol Psychiatry 56:618–631
131. Roberts MA, Ray RS, Roberts RJ (1984) A playroom observational procedure for assessing hyperactive boys. J Pediatr Psychol 9:177–191
132. Milich R, Loney J, Roberts MA (1986) Playroom observations of activity level and sustained attention: two-year stability. J Consult Clin Psychol 54:272–274
133. Barkley RA, DuPaul GJ, McMurray MB (1990) Comprehensive evaluation of attention deficit disorder with and without hyperactivity as defined by research criteria. J Consult Clin Psychol 58:775–789
134. Pliszka SR (1992) Comorbidity of attention-deficit hyperactivity disorder and overanxious disorder. J Am Acad Child Adolesc Psychiatry 31:197–203
135. Glutting JJ, Robins PM, de Lancey E (1997) Discriminant validity of test observations for children with attention deficit/hyperactivity. J Sch Psychol 35:391–401

136. Mariani MA, Barkley RA (1997) Neuropsychological and academic functioning in preschool boys with attention deficit hyperactivity disorder. Dev Neuropsychol 13:111–129
137. Bauermeister JJ, Matos M, Reina G, Salas CC, Martínez JV, Cumba E, Barkley RA (2005) Comparison of the DSM-IV combined and inattentive types of ADHD in a school-based sample of Latino/Hispanic children. J Child Psychol Psychiatry 46:166–179
138. McConaughy SH, Ivanova MY, Antshel K, Eiraldi RB (2009) Standardized observational assessment of attention deficit hyperactivity disorder combined and predominantly inattentive subtypes. I. Test session observations. Sch Psych Rev 38:45–66
139. Willcutt EG, Hartung CM, Lahey BB, Loney J, Pelham WE Jr (1999) Utility of behavior ratings by examiners during assessments of preschool children with attention-deficit/hyperactivity disorder. J Abnorm Child Psychol 27:463–472
140. Barkley RA, DuPaul GJ, McMurray MB (1991) Attention deficit disorder with and without hyperactivity: Clinical response to three dose levels of methylphenidate. Pediatrics 87:519–531
141. Fischer M, Newby RF (1991) Assessment of stimulant response in ADHD children using a refined multimethod clinical protocol. J Clin Child Psychol 20:232–244
142. DuPaul GJ, Barkley RA, McMurray MB (1994) Response of children with ADHD to methylphenidate: interaction with internalizing symptoms. J Am Acad Child Adolesc Psychiatry 33:894–903
143. Ialongo NS, Lopez M, Horn WF, Pascoe JM, Greenberg G (1994) Effects of psychostimulant medication on self-perceptions of competence, control, and mood in children with attention deficit hyperactivity disorder. J Clin Child Psychol 23:161–173
144. Fischer M, Newby RF (1998) Use of the restricted academic task in ADHD dose-response relationships. J Learn Disabil 31:608–612
145. Grizenko N, Bhat M, Schwartz G, Ter-Stepanian M, Joober R (2006) Efficacy of methylphenidate in children with attention-deficit hyperactivity disorder and learning disabilities: a randomized crossover trial. J Psychiatry Neurosci 31:46–51
146. Gorman EB, Klorman R, Thatcher JE, Borgstedt AD (2006) Effects of methylphenidate on subtypes of attention-deficit/hyperactivity disorder. J Am Acad Child Adolesc Psychiatry 45:808–816
147. Karama S, Ben Amor L, Grizenko N, Ciampi A, Mbekou V, Ter-Stepanian M, Lageix P, Baron C, Schwartz G, Joober R (2009) Factor structure of the restricted Academic Situation Scale: implications for ADHD. J Atten Disord 12:442–448
148. Grizenko N, Pereira RMR, Joober R (2013) Sensitivity of scales to evaluate change in symptomatology with psychostimulants in different ADHD subtypes. J Can Acad Child Adolesc Psychiatry 22:153–158
149. Green CT, Long DL, Green D, Iosif A-M, Dixon JF, Miller MR, Fassbender C, Schweitzer JB (2012) Will working memory training generalize to improve off-task behavior in children with attention-deficit/hyperactivity disorder? Neurotherapeutics 9:639–648
150. Fischer M, Barkley RA, Edelbrock CS, Smallish L (1990) The adolescent outcome of hyperactive children diagnosed by research criteria: II. Academic, attentional, and neuropsychological status. J Consult Clin Psychol 58:580–588
151. Barkley RA, Anastopoulos AD, Guevremont DC, Fletcher KE (1991) Adolescents with ADHD: Patterns of behavioral adjustment, academic functioning, and treatment utilization. J Am Acad Child Adolesc Psychiatry 30:752–761
152. McGrath AM, Handwerk ML, Armstrong KJ, Lucas CP, Friman PC (2004) The validity of the ADHD section of the diagnostic interview schedule for children. Behav Modif 28:349–374
153. Barkley RA, Fischer M, Newby RF, Breen MJ (1988) Development of a multimethod clinical protocol for assessing stimulant drug response in children with attention deficit disorder. J Clin Child Psychol 17:14–24
154. Breen MJ (1989) Cognitive and behavioral differences in AdHD boys and girls. J Child Psychol Psychiatry 30:711–716
155. Hale JB, Fiorello CA, Brown LL (2005) Determining medication treatment effects using teacher ratings and classroom observations of children with ADHD: Does neuropsychological impairment matter? Educ Child Psychol 22:39–61
156. Hale JB, Reddy LA, Semrud-Clikeman M, Hain LA, Whitaker J, Morley J, Lawrence K, Smith A, Jones N (2011) Executive impairment determines ADHD medication response: implications for academic achievement. J Learn Disabil 44:196–212
157. Glutting JJ, Oakland T (1993) GATSB, Guide to the assessment of test session behavior for the WISC-III and the WIAT: manual. Psychological Corporation, San Antonio
158. Bunte TL, Laschen S, Schoemaker K, Hessen DJ, van der Heijden PGM, Matthys W (2013) Clinical usefulness of observational assessment in the diagnosis of DBD and ADHD in preschoolers. J Clin Child Adolesc Psychol 42:749–761
159. Wells KC, Chi TC, Hinshaw SP et al (2006) Treatment-related changes in objectively measured parenting behaviors in the multimodal treatment study of children with attention-deficit/hyperactivity disorder. J Consult Clin Psychol 74:649–657
160. Thompson M, Laver-Bradbury C, Ayres M et al (2009) A small-scale randomized controlled trial of the revised new forest parenting programme for preschoolers with attention deficit hyperactivity disorder. Eur Child Adolesc Psychiatry 18:605–616
161. Babinski DE, Waxmonsky JG, Pelham WE Jr (2014) Treating parents with attention-deficit/hyperactivity disorder: the effects of behavioral parent training and acute stimulant medication treatment on parent–child interactions. J Abnorm Child Psychol 42:1129–1140
162. Barkley RA, Fischer M, Edelbrock C, Smallish L (1991) The adolescent outcome of hyperactive children diagnosed by research criteria—III. Mother-child interactions, family conflicts and maternal psychopathology. J Child Psychol Psychiatry 32:233–255
163. Barkley RA, Anastopoulos AD, Guevremont DC, Fletcher KE (1992) Adolescents with attention deficit hyperactivity disorder: Mother-adolescent interactions, family beliefs and conflicts, and maternal psychopathology. J Abnorm Child Psychol 20:263–288
164. Edwards G, Barkley RA, Laneri M, Fletcher K, Metevia L (2001) Parent–adolescent conflict in teenagers with ADHD and ODD. J Abnorm Child Psychol 29:557–572
165. Barkley RA, Guevremont DC, Anastopoulos AD, Fletcher KE (1992) A comparison of three family therapy programs for treating family conflicts in adolescents with attention-deficit hyperactivity disorder. J Consult Clin Psychol 60:450–462
166. Barkley RA, Edwards G, Laneri M, Fletcher K, Metevia L (2001) The efficacy of problem-solving communication training alone, behavior management training alone, and their combination for parent–adolescent conflict in teenagers with ADHD and ODD. J Consult Clin Psychol 69:926–941
167. Christensen A, Heavey CL (1990) Gender and social structure in the demand/withdraw pattern of marital conflict. J Pers Soc Psychol 59:73–81
168. Leipold EE, Bundy AC (2000) Playfulness in children with attention deficit hyperactivity disorder. OTJR 20:61–79
169. Cordier R, Bundy AC, Hocking C, Einfeld S (2010) Empathy in the play of children with attention deficit hyperactivity disorder. OTJR 30:122–132
170. Cordier R, Bundy AC, Hocking C, Einfeld S (2010) Playing with a child with ADHD: A focus on the playmates. Scand J Occup Ther 17:191–199

171. Cordier R, Bundy AC, Hocking C, Einfeld S (2010) Comparison of the play of children with attention deficit hyperactivity disorder by subtypes. Aust Occup Ther J 57:137–145
172. Wilkes S, Cordier R, Bundy AC, Docking K, Munro N (2011) A play-based intervention for children with ADHD: A pilot study. Aust Occup Ther J 58:231–240
173. Bundy AC, Skard G (1997) Test of playfulness. Colorado State University, Fort Collins
174. Brentnall J, Bundy AC, Kay FCS (2008) The effect of the length of observation on test of playfulness scores. OTJR 28:133–140
175. Bundy AC, Nelson L, Bingaman K (2001) Validity and reliability of a test of playfulness. OTJR 21:277–292
176. Börger N, van der Meere J (2000) Visual behaviour of ADHD children during an attention test: an almost forgotten variable. J Child Psychol Psychiatry 41:525–532
177. Stroes A, Alberts E, Van Der Meere JJ (2003) Boys with ADHD in social interaction with a nonfamiliar adult: an observational study. J Am Acad Child Adolesc Psychiatry 42:295–302
178. Buitelaar JK, Swinkels SH, De Vries H, Van der Gaag RJ, Van Hooff JA (1994) An ethological study on behavioural differences between hyperactive, aggressive, combined hyperactive/aggressive and control children. J Child Psychol Psychiatry 35:1437–1446
179. Kofler MJ, Rapport MD, Alderson RM (2008) Quantifying ADHD classroom inattentiveness, its moderators, and variability: a meta-analytic review. J Child Psychol Psychiatry 49:59–69
180. Rapport MD, Kofler MJ, Alderson RM, Dupaul GJ (2009) Variability of attention processes in ADHD: Observations from the classroom. J Atten Disord 12:563–573
181. Kazdin AE (2005) Evidence-based assessment for children and adolescents: Issues in measurement development and clinical application. J Clin Child Adolesc Psychol 34:548–558
182. Steiner NJ, Sidhu T, Rene K, Tomasetti K, Frenette E, Brennan RT (2013) Development and testing of a direct observation code training protocol for elementary aged students with attention deficit/hyperactivity disorder. Educ Assess Eval Account 25:281–302
183. Nock MK, Kurtz SMS (2005) Direct behavioral observation in school settings: bringing science to practice. Cogn Behav Pract 12:359–370
184. Gardner F (2000) Methodological issues in the direct observation of parent–child interaction: Do observational findings reflect the natural behavior of participants? Clin Child Fam Psychol Rev 3:185–198
185. Dubey DR, Kent RN, O'Leary SG, Broderick JE, O'Leary KD (1977) Reactions of children and teachers to classroom observers: a series of controlled investigations. Behav Ther 8:887–897
186. Bender WN, Smith JK (1990) Classroom behavior of children and adolescents with learning disabilities: a meta-analysis. J Learn Disabil 23:298–305
187. Charlop MH, Schreibman L, Mason J, Vesey W (1983) Behavior-setting interactions of autistic children: a behavioral mapping approach to assessing classroom behaviors. Anal Interv Dev Disabil 3:359–373