frontiers in Psychology

ORIGINAL RESEARCH
published: 24 July 2020
doi: 10.3389/fpsyg.2020.01840

Toward a Dimensional Assessment of Externalizing Disorders in Children: Reliability and Validity of a Semi-Structured Parent Interview

Ann-Kathrin Thöne[1], Anja Görtz-Dorten[1,2], Paula Altenberger[1], Christina Dose[1], Nina Geldermann[1], Christopher Hautmann[1], Lea Teresa Jendreizik[1], Anne-Katrin Treier[1], Elena von Wirth[1], Tobias Banaschewski[3], Daniel Brandeis[3,4,5,6], Sabina Millenet[3], Sarah Hohmann[3], Katja Becker[7,8], Johanna Ketter[7], Johannes Hebebrand[9], Jasmin Wenning[9], Martin Holtmann[10], Tanja Legenbauer[10], Michael Huss[11], Marcel Romanos[12], Thomas Jans[12], Julia Geissler[12], Luise Poustka[13], Henrik Uebel-von Sandersleben[13], Tobias Renner[14], Ute Dürrwächter[14] and Manfred Döpfner[1,2]*

[1] School of Child and Adolescent Cognitive Behavior Therapy (AKiP), Faculty of Medicine, University Hospital Cologne, University of Cologne, Cologne, Germany, [2] Department of Child and Adolescent Psychiatry, Psychosomatics, and Psychotherapy, Faculty of Medicine, University Hospital Cologne, University of Cologne, Cologne, Germany, [3] Department of Child and Adolescent Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany, [4] Department of Child and Adolescent Psychiatry and Psychotherapy, Psychiatric Hospital, University of Zurich, Zurich, Switzerland, [5] Zurich Center for Integrative Human Physiology, University of Zurich, Zurich, Switzerland, [6] Neuroscience Center Zurich, University and ETH Zürich, Zurich, Switzerland, [7] Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Medical Faculty, Philipps-University Marburg, Marburg, Germany, [8] Center for Mind, Brain and Behavior, University of Marburg and Justus Liebig University Giessen, Marburg, Germany, [9] Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, University Hospital Essen, University of Duisburg-Essen, Essen, Germany, [10] LWL-University Hospital for Child and Adolescent Psychiatry, Ruhr-University Bochum, Hamm, Germany, [11] Department of Child and Adolescent Psychiatry and Psychotherapy, University Medical Center, Johannes Gutenberg University Mainz, Mainz, Germany, [12] Center of Mental Health, Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, University Hospital Würzburg, Würzburg, Germany, [13] Department of Child and Adolescent Psychiatry and Psychotherapy, University Medical Center Göttingen, Göttingen, Germany, [14] Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Tübingen University Hospital, Tübingen, Germany*

OPEN ACCESS

Edited by:
Giancarlo Tamanza, Catholic University of the Sacred Heart, Italy

Reviewed by:
Robert Emery, University of Virginia, United States
Angélica Quiroga-Garza, University of Monterrey, Mexico

***Correspondence:**
Ann-Kathrin Thöne
ann-kathrin.thoene@uk-koeln.de

Specialty section:
This article was submitted to Psychology for Clinical Settings, a section of the journal Frontiers in Psychology

Received: *17 April 2020*
Accepted: *06 July 2020*
Published: *24 July 2020*

Citation:
Thöne A-K, Görtz-Dorten A, Altenberger P, Dose C, Geldermann N, Hautmann C, Jendreizik LT, Treier A-K, von Wirth E, Banaschewski T, Brandeis D, Millenet S, Hohmann S, Becker K, Ketter J, Hebebrand J, Wenning J, Holtmann M, Legenbauer T, Huss M, Romanos M, Jans T, Geissler J, Poustka L, Uebel-von Sandersleben H, Renner T, Dürrwächter U and Döpfner M (2020) Toward a Dimensional Assessment of Externalizing Disorders in Children: Reliability and Validity of a Semi-Structured Parent Interview. Front. Psychol. 11:1840. doi: 10.3389/fpsyg.2020.01840

Objective: This study assesses the reliability and validity of the DSM-5-based, semi-structured *Clinical Parent Interview for Externalizing Disorders in Children and Adolescents* (ILF-EXTERNAL).

Method: Participant data were drawn from the ongoing ESCAschool intervention study. The ILF-EXTERNAL was evaluated in a clinical sample of 474 children and adolescents (aged 6−12 years, 92 females) with symptoms of attention-deficit/hyperactivity disorder (ADHD). To obtain interrater reliability, the one-way random-effects, absolute agreement models of the intraclass correlation (ICC) for single ICC(1,1) and average measurements ICC(1,3) were computed between the interviewers and two independent raters for 45 randomly selected interviews involving ten interviewers. Overall agreement on DSM-5 diagnoses was assessed using Fleiss' kappa. Further analyses evaluated internal consistencies, item-total correlations as well as correlations between symptom severity and the degree of functional impairment. Additionally, parents completed the

German version of the *Child Behavior Checklist* (CBCL) and two DSM-5-based parent questionnaires for the assessment of ADHD symptoms and symptoms of disruptive behavior disorders (FBB-ADHS; FBB-SSV), which were used to evaluate convergent and divergent validity.

Results: ICC coefficients demonstrated very good to excellent interrater reliability on the item and scale level of the ILF-EXTERNAL [scale level: ICC(1,1) = 0.83−0.95; ICC(1,3) = 0.94−0.98]. Overall kappa agreement on DSM-5 diagnoses was substantial to almost perfect for most disorders ($0.38 \leq \kappa \leq 0.94$). With some exceptions, internal consistencies ($0.60 \leq \alpha \leq 0.86$) and item-total correlations ($0.21 \leq r_{it} \leq 0.71$) were generally satisfactory to good. Furthermore, higher symptom severity was associated with a higher degree of functional impairment. The evaluation of convergent validity revealed positive results regarding clinical judgment and parent ratings (FBB-ADHS; FBB-SSV). Correlations between the ILF-EXTERNAL scales and the CBCL *Externalizing Problems* were moderate to high. Finally, the ILF-EXTERNAL scales were significantly more strongly associated with the CBCL *Externalizing Problems* than with the *Internalizing Problems*, indicating divergent validity.

Conclusion: In clinically referred, school-age children, the ILF-EXTERNAL demonstrates sound psychometric properties. The ILF-EXTERNAL is a promising clinical interview and contributes to high-quality diagnostics of externalizing disorders in children and adolescents.

Keywords: structured interview, ADHD, ODD, externalizing disorders, reliability, intraclass correlation coefficient, validity

INTRODUCTION

Structured clinical interviews are considered to be the gold standard for diagnosing mental disorders (Rettew et al., 2009; Hoyer and Knappe, 2012; Nordgaard et al., 2013). Accumulating evidence suggests that structured interviews lead to improved diagnostic accuracy and reliability (Frick et al., 2010; Segal and Williams, 2014; Leffler et al., 2015), which can in turn enhance the quality of treatment decision making (Galanter and Patel, 2005). In clinical research, structured interviews are especially used to screen participants for study inclusion or to evaluate psychotherapeutic outcomes (Hoyer and Knappe, 2012; Segal and Williams, 2014). Besides their use in research, such interviews have increasingly found their way into clinical practice as part of a comprehensive and standardized diagnostic process (Frick et al., 2010; Hoyer and Knappe, 2012; Segal and Williams, 2014). Moreover, clinicians in training can also benefit from these instruments, as they cover diagnostic criteria in a systematic manner (Frick et al., 2010; Segal and Williams, 2014; Leffler et al., 2015).

In terms of their degree of structure, clinical interviews can be classified into highly structured versus semi-structured. While the highly structured interviews require only a minimum of training, they leave little flexibility for the interviewer to explore and rate the patient's symptomatology. Typically, closed-ended questions form a dichotomous assessment, that is, a clinical symptom is either present or absent (Frick et al., 2010; Segal and Williams, 2014; Leffler et al., 2015). Examples of highly structured clinical interviews for assessing children and adolescents include the *NIMH Diagnostic Interview Schedule for Children Version IV* (NIMH DISC-IV; Shaffer et al., 2000), the *Children's Interview for Psychiatric Syndromes* which encompasses separate child (ChIPS; Weller et al., 1999b) and parent versions (P-ChIPS; Weller et al., 1999a), and the *Mini-International Neuropsychiatric Interview for Children and Adolescents* (Mini-KID; Sheehan et al., 2010). Most of these structured interviews have yet to be revised and validated for DSM-5 (American Psychiatric Association, 2013). As reviewed by Leffler et al. (2015), the current versions of these interviews show high interrater reliability (IRR) and good validity in community and clinical samples. However, such findings may also be attributable to the high degree of structure, and the inherent limited scope for interviewers to form their own clinical judgment (Leffler et al., 2015).

By comparison, semi-structured interviews allow the interviewer to inquire about symptoms, make informed judgments, and score responses in a more flexible manner (e.g., Likert-type scales). While this scoring format requires more extensive training, it can follow a dimensional approach by taking into account the severity of symptoms (Frick et al., 2010; Döpfner and Petermann, 2012; Leffler et al., 2015). Therefore, different interviewers may form disparate judgments, which can in turn result in lower IRR compared to their highly structured counterparts. One of the most prominent semi-structured clinical interviews is the *Schedule for Affective Disorders and Schizophrenia for School Aged Children* (K-SADS; Kaufman et al., 1997) which mainly aims at an early diagnosis of affective

disorders but also includes sections on other common mental disorders. Both the parents and their child can be interviewed at the same time. Different editions of the K-SADS exist and the instrument has been evaluated in a variety of populations, with overall good psychometric evidence. Available diagnostic interrater agreement on DSM-IV or DSM-5 externalizing disorders ranges from moderate to almost perfect agreement for several cross-cultural K-SADS adaptations with generally higher agreement in clinical samples (Kim et al., 2004; Ghanizadeh et al., 2006; Ulloa et al., 2006; de la Peña et al., 2018; Nishiyama et al., 2020) than in the community population (Birmaher et al., 2009; Chen et al., 2017). Kariuki et al. (2018) evaluated the attention-deficit/hyperactivity disorder (ADHD) module of the K-SADS in a large community sample and obtained moderate to substantial intraclass correlation (ICC) coefficients for the subdimensions of ADHD. With regard to convergent validity, small to moderate correlations were found between clinical diagnoses and the broadband parent-rated *Child Behavior Checklist* (CBCL) questionnaire (Kim et al., 2004; Birmaher et al., 2009; Brasil and Bordin, 2010; Chen et al., 2017). Furthermore, correlations between clinical diagnoses and the corresponding scales of the CBCL were generally higher than divergent correlations (Birmaher et al., 2009; Chen et al., 2017). Another semi-structured interview is the *Child and Adolescent Psychiatric Assessment* (CAPA; Angold and Costello, 2000) which covers the full range of common mental disorders. With a duration of up to 120 min, it can be very time-consuming to administer (Leffler et al., 2015). In a test-retest study of the CAPA, ICC coefficients for DSM-III-R symptom scale scores ranged from 0.50 for oppositional defiant disorder (ODD) to 0.98 for substance abuse / dependence in self-reports of clinically referred children and adolescents (Angold and Costello, 1995). Furthermore, the CAPA interview has shown good construct validity in relation to ten formulated criteria (Angold and Costello, 2000).

Overall, semi-structured clinical interviews provide a valuable tool for diagnosing mental disorders in children and adolescents (Nordgaard et al., 2013). However, given the evolving conceptualizations of psychopathology, there is a current need for clinical interviews to meet these changing requirements. One such evolving conceptualization considers whether diagnostic domains are best characterized as discrete categories (such as in the DSM-5) or whether they should follow a dimensional approach (Coghill and Sonuga-Barke, 2012; Döpfner and Petermann, 2012). Consequently, state-of-the-art assessment instruments should have the flexibility to allow both a categorical assessment and follow a dimensional approach which allows for varying degrees of severity and functional impairment. To the best of our knowledge, there is no diagnostic system available which meets all of the following criteria: a DSM-5-based, semi-structured clinical interview for externalizing disorders which follows both a categorical and dimensional approach by assessing symptom severity and functional impairment on a Likert scale and includes parallel parent forms with the exact same diagnostic scales.

To meet these criteria, we developed a comprehensive set of clinical parent and patient interviews *Diagnostic System of Mental Disorders in Children and Adolescents – Interview* (DISYPS-ILF; Görtz-Dorten et al., in press), which are part of the German *Diagnostic System of Mental Disorders in Children and Adolescents based on the ICD-10 and DSM-5* (DISYPS-III; Döpfner and Görtz-Dorten, 2017). Of these interviews, the *Clinical Parent Interview for Externalizing Disorders in Children and Adolescents* (ILF-EXTERNAL) covers diagnostic criteria according to the DSM-5 for the following externalizing disorders: ADHD, with the subtypes "combined type," "predominantly inattentive type," and "predominantly hyperactive-impulsive type;" ODD; conduct disorder (CD), with the specifier limited prosocial emotions; and disruptive mood dysregulation disorder (DMDD); for further details, see **Materials and Methods**. Besides this categorical assessment, the ILF-EXTERNAL also allows clinical symptoms to be viewed from a dimensional standpoint. A further distinguishing and novel characteristic of the ILF-EXTERNAL is that both the interview and the rating scales for parents, teachers, and patients correspond to the same diagnostic system DISYPS-III (Döpfner and Görtz-Dorten, 2017) and therefore have the exact same diagnostic scales. This allows a specific comparison of ratings of parents, teachers, and patients with clinical judgments. In addition, we sought to psychometrically evaluate this interview in a clinical sample of children with externalizing problems, as this group of children represents the prospective target group for clinical assessment using the ILF-EXTERNAL.

Currently, the ILF-EXTERNAL is being conducted in the multicenter consortium ESCAlife (ESCAlife: Evidence-Based Stepped Care of ADHD along the Lifespan). The purpose of this consortium is to evaluate adaptive interventions for patients diagnosed with ADHD, including 3—6-year-old preschool children (ESCApreschool; Becker et al., 2020), 6—12-year-old school children (ESCAschool; Döpfner et al., 2017), and 12—17-year-old adolescents (ESCAadol; Geissler et al., 2018).

The overall aim of this study is to present the newly developed clinical parent interview ILF-EXTERNAL and its psychometric properties, including (1) descriptive statistics for all scales, (2) internal consistencies and item-total correlations, (3) IRR on the item and scale level, (4) overall agreement on DSM-5 diagnoses, (5) associations between symptom severity and the degree of functional impairment, and (6) convergent and divergent validity in a clinical sample of school-age children with ADHD symptoms.

MATERIALS AND METHODS

Measures

During the ESCAschool study, the below-mentioned measures were collected at several main assessment points (Döpfner et al., 2017). In the present study, measures at baseline (i.e., before any intervention) were analyzed.

Clinical Parent Interview for Externalizing Disorders in Children and Adolescents (ILF-EXTERNAL)

The clinical parent interview ILF-EXTERNAL (Görtz-Dorten et al., in press) is part of the DISYPS-III (Döpfner and Görtz-Dorten, 2017). The ILF-EXTERNAL comprises a set of items,

each of which explores a DSM-5 symptom criterion. Following a semi-structured approach, clinicians give their own judgment by rating each item on a 4-point Likert scale ranging from 0 (age-typical / not at all), to 3 (very much), with higher scores indicating higher symptom severity. To aid clinical judgment, a short description of the symptom severity is provided for each score. Also, an example sentence of a child's behavior representing a rating of 3 is given for each item. Item scores of 2 and higher are interpreted as clinically relevant and considered to fulfill the DSM-5 symptom criteria. The ILF-EXTERNAL consists of 18 items assessing ADHD symptoms which can be aggregated into two scales, *Inattention* (nine items) and *Hyperactivity-Impulsivity* (nine items). Together, these 18 items form the *ADHD Symptoms* scale. Additionally, five items assess functioning and psychological strain associated with ADHD symptoms and form the *ADHD Functional Impairment* scale. Moreover, the ILF-EXTERNAL consists of 36 items assessing oppositional and disruptive symptoms which are aggregated to the following scales: *ODD Symptoms* (eight items) and *CD Symptoms* (15 items), which together form the scale *ODD/CD Symptoms* (23 items). Further items form the scales *Disruptive Mood Dysregulation* (five items, three of which are also part of the *ODD Symptoms* scale) and *Limited Prosocial Emotions* (11 items). In addition, five items assess functioning and psychological strain associated with ODD and CD symptoms and form the *ODD/CD Functional Impairment* scale (see **Supplementary Table 1** in the online **Supplementary Material** for a more detailed description of the items forming each scale). Scale scores are computed by averaging the associated item scores. In the present study, the items assessing aggressive and antisocial symptoms from the age of 11 (B06 to B15) were excluded from further analyses due to an obvious floor effect, resulting in the shortened scales *CD Symptoms – short version* (five items) and *ODD/CD Symptoms – short version* (13 items).

Child Behavior Checklist for Ages 6–18 (CBCL/6-18R)

To examine convergent and divergent validity, information from the German CBCL/6-18R was used (Arbeitsgruppe Deutsche Child Behavior Checklist, 1998; Döpfner et al., 2014). Originally developed by the Achenbach group (Achenbach, 1991; Achenbach and Rescorla, 2001), the German CBCL/6-18R is a broadband questionnaire comprising 120 items developed to assess behavioral and emotional problems in children and adolescents. Parents rate their child's behavior on a 3-point scale (0 = "not true," 1 = "somewhat or sometimes true," and 2 = "very true or often true"). The items form eight syndrome scales and three broadband scales (*Externalizing Problems, Internalizing Problems, Total Problems*). The German CBCL/6-18R has demonstrated at least satisfactory internal consistencies for the eight syndrome scales with slightly higher values in a large clinical sample than in a community sample (Döpfner et al., 2014). Exceptions are the scales *Thought Problems* ($\alpha < 0.70$ in both samples) and *Somatic Complaints* ($\alpha = 0.65$ in the community sample). Internal consistencies were good for the *Externalizing Problems* and *Internalizing Problems* ($\alpha > 0.80$) and excellent for the *Total Problems* ($\alpha > 0.90$) in both samples. In cross-cultural analyses, Rescorla et al. (2007) found that parents' ratings were similar across 31 societies including Germany, indicating the multicultural robustness of the CBCL. Furthermore, the configural invariance of the 8-syndrome structure of the CBCL was confirmed in large cross-cultural studies including Germany (Ivanova et al., 2007, 2019). In the present study, the raw scale scores of the eight syndrome scales and the *Internalizing Problems* and *Externalizing Problems* were used.

Symptom Checklist for Attention-Deficit/ Hyperactivity Disorder (FBB-ADHS)

The German *Symptom Checklist for Attention-deficit/hyperactivity Disorder* (FBB-ADHS) is part of the DISYPS-III (Döpfner and Görtz-Dorten, 2017). This questionnaire consists of 27 items which form identical scales to those in the ILF-EXTERNAL and an additional six items assessing the child's competencies. All items are rated on a 4-point Likert scale ranging from 0 ("not at all") to 3 ("very much"). Psychometric evaluations support the reliability and validity of the FBB-ADHS (Döpfner et al., 2008; Erhart et al., 2008). The present analyses included the scales *Inattention*, *Hyperactivity-Impulsivity*, *ADHD Symptoms*, and *ADHD Functional Impairment*.

Symptom Checklist for Disruptive Behavior Disorders (FBB-SSV)

The German *Symptom Checklist for Disruptive Behavior Disorders* (FBB-SSV) is also part of the DISYPS-III (Döpfner and Görtz-Dorten, 2017). The structure and assessment are the same as outlined for the FBB-ADHS. The FBB-SSV includes 46 items which also form identical scales to those in the ILF-EXTERNAL and an additional 12 items assessing the child's competencies. Psychometric evaluations of the FBB-SSV revealed positive results regarding reliability and validity (Görtz-Dorten et al., 2014). The scales *ODD Symptoms, CD Symptoms, ODD/CD Symptoms, Disruptive Mood Dysregulation, Limited Prosocial Emotions*, and *ODD/CD Functional Impairment* were used in the present study. For the sake of consistency with the scales *CD Symptoms – short version* and *ODD/CD Symptoms – short version* of the ILF-EXTERNAL, the items assessing aggressive and antisocial symptoms from the age of 11 (B06 to B15) from the FBB-SSV were also excluded from further analyses.

Participants and Procedure

Data collection was based on the ongoing ESCAschool intervention study (target $N = 521$), which is part of the research consortium ESCAlife and involves nine study centers located in Germany (Cologne, Essen, Göttingen, Hamm, Mainz, Mannheim, Marburg, Tübingen, Würzburg). The ESCAschool study investigates an evidenced-based, individualized, stepwise-intensifying treatment program based on behavioral and pharmacological interventions for children diagnosed with ADHD. For further details on the procedures, including inclusion and exclusion criteria, please refer to Döpfner et al. (2017). In the present study, the ILF-EXTERNAL was evaluated using ESCAschool baseline data from 474 children (age range

6−12 years, M = 8.9, SD = 1.5, 92 females). The assessment of the ILF-EXTERNAL baseline data is part of a screening to check the participants' eligibility for the ESCAschool study. The screening was conducted at two successive appointments no longer than 8 weeks apart. During the screening, the ILF-EXTERNAL was administered to the parents and was either video- or audio-recorded. About one third of the children (32.5%) were receiving ADHD medication prior to the study. In these cases, parents were asked to describe their child's behavior with and without medication, resulting in two ratings for each item. For the present analyses, the children's symptomatology without medication was analyzed. Besides children diagnosed with ADHD, the present sample also included children who did not meet criteria for an ADHD diagnosis (i.e., so-called screening negatives of the ESCAschool study). These screening negatives (n = 32, including 9 females) were characterized by subclinical ADHD symptomatology, which allowed us to capture the full spectrum of ADHD symptoms. Descriptive statistics (M, SD) for all ILF-EXTERNAL scales considering only the screening negatives are reported in **Table 2**. As can be seen, although these children did not fulfill inclusion criteria for the ESCAschool treatment study, they nevertheless exhibited symptoms of externalizing behavior problems. Clinical diagnoses of ADHD and comorbid externalizing disorders were based on the outcome of the ILF-EXTERNAL. To assess comorbid symptoms, all clinicians applied a clinical diagnostic checklist (DCL-SCREEN) from the DISYPS-III (Döpfner and Görtz-Dorten, 2017). All parents and children gave their assent and written informed consent, and each participating study site received ethical approval (Döpfner et al., 2017). Participant data are presented in **Table 1**.

TABLE 1 | Sample characteristics.

	Subsample for the analysis of interrater reliability (n = 45)	Total sample (N = 474)
Age: mean *(SD)*	9.2 (1.6)	8.9 (1.5)
Male: *n* (%)	34 (75.6)	382 (80.6)
Diagnosis *n* (%)		
No ADHD diagnosis	1 (2.2)	32 (6.8)
ADHD – combined type	28 (62.2)	208 (43.9)
ADHD – predominantly inattentive type	14 (31.1)	184 (38.8)
ADHD – predominantly hyperactive-impulsive type	2 (4.4)	50 (10.5)
Comorbidities *n* (%)		*n* = 454−465
Internalizing disorders:		
– Anxiety	2 (4.4)	29 (6.4)
– Depression	2 (4.4)	15 (3.1)
Externalizing disorders:		
– Oppositional defiant disorder	23 (51.1)	166 (36.6)
– Disruptive mood dysregulation disorder	6 (13.3)	40 (8.8)
– Conduct disorder	5 (11.1)	28 (6.2)
Other disorders:		
– Obsessive-compulsive disorder	1 (2.2)	2 (0.4)
– Tic disorder	4 (8.9)	24 (5.2)
– Autism spectrum disorder	0	2 (0.4)
Medication *n* (%)		*n* = 462
ADHD medication	17 (37.8)	150 (32.5)
Parents' primary language *n* (%)		*n* = 458
German	42 (93.3)	429 (93.7)
Highest parents' graduation *n* (%)		*n* = 455
Higher-track school	24 (53.3)	261 (57.4)
Vocational school	2 (4.4)	26 (5.7)
Medium-track school	14 (31.1)	123 (27.0)
Lower-track school	4 (8.9)	43 (9.5)

Clinical diagnoses of ADHD and comorbid externalizing disorders were based on the semi-structured clinical interview ILF-EXTERNAL conducted with the parents. Further comorbid symptoms were assessed using a clinical diagnostic checklist. ADHD, attention-deficit/hyperactivity disorder.

Subsample for the Analysis of Interrater Reliability

To obtain IRR, a subsample of 45 interviews of the ILF-EXTERNAL was chosen (for the characteristics of this subsample, see **Table 1**). More specifically, we empirically determined the required sample size as recommended by published guidelines on IRR studies (Kottner et al., 2011). We selected a method for sample size calculation for the ICC coefficient (Zou, 2012) which estimates the required sample (N) to achieve a reliability coefficient (ρ) that is not less than a prespecified value (ρ_0) with a prespecified assurance probability. The calculations revealed that a minimum of 42 interviews rated by two additional raters (k = 3) is required to ensure that the lower limit of a 95% one-sided confidence limit for ρ = 0.80 is no less than ρ_0 = 0.65 with 80% assurance probability based on the ICC one-way random-effects model (Zou, 2012). Subsequently, 45 interviews (five interviews from one clinician from each of the nine study sites) were randomly selected using the *select cases* function in SPSS. Inclusion criteria for the interview recordings were as follows: A video- or audio-recording had to be present for both parts of the interview, the recordings needed to have sufficient audio quality, the clinical assessment had to follow the ILF-EXTERNAL, and, if possible, both parts of the interview should be conducted by the same interviewer. If it was not possible to rate an interview recording due to violation of the inclusion criteria, another recording from the same interviewer was randomly selected. For one study site, there were only four recordings available from one interviewer; in this case, we therefore included one recording by another interviewer from the same study site. In short, the subsample to obtain IRR consists of 45 recordings of the ILF-EXTERNAL conducted by ten interviewers from nine study sites. Typically, interviewers conducted the first part of the interview, assessing ADHD symptoms, at the first appointment and the second part, assessing ODD/CD symptoms, at the second appointment. For the ADHD part, 37 (82.2%) interviews were conducted with the mother, three (6.7%) with the father, and five (11.1%) with both parents. Similarly, for the ODD/CD part, 40 (88.9%) interviews were conducted with the mother, three (6.7%) with the father, and two (4.4%) with both parents. Regarding the duration of the interviews, the ADHD part had a mean length of 42 min (SD = 19 min, range 15 to 88 min) and the ODD/CD part had a mean length of 35 min (SD = 20 min, range 5 to 98 min).

TABLE 2 | Scale characteristics, Cronbach's alpha (α) and range of item-total correlations of the ILF-EXTERNAL.

		Total sample (*N* = 474)				Screening negatives (*n* = 32)	
Scale	***k* (items)**	**α**	**Item-total *r* (range)**	**Mean (*SD*)**	***N***	**Mean *(SD)***	***n***
ADHD symptoms	18	0.84	0.21–0.62	1.80 (0.50)	474	1.09 (0.51)	32
– Inattention	9	0.71	0.29–0.49	1.95 (0.48)	474	1.35 (0.49)	32
– Hyperactivity-Impulsivity	9	0.87	0.50–0.68	1.64 (0.73)	474	0.84 (0.62)	32
ADHD functional impairment	5	0.62	0.24–0.48	1.61 (0.59)	472	1.03 (0.49)	31
ODD/CD symptoms—short version	13	0.84	0.25–0.62	0.85 (0.50)	450	0.55 (0.55)	23
– ODD symptoms	8	0.82	0.41–0.61	1.12 (0.63)	451	0.78 (0.68)	24
– CD symptoms—short version	5	0.60	0.22–0.47	0.40 (0.43)	450	0.22 (0.37)	23
Disruptive mood dysregulation	5	0.83	0.53–0.67	1.08 (0.73)	452	0.73 (0.65)	24
Limited prosocial emotions	11	0.77	0.23–0.58	0.50 (0.42)	444	0.37 (0.30)	22
ODD/CD functional impairment	5	0.86	0.61–0.71	0.93 (0.78)	442	0.49 (0.73)	21

Screening negatives are participants who have been screened for eligibility of the ESCAschool study and are characterized by subclinical ADHD symptoms. ADHD, attention-deficit/hyperactivity disorder; CD, conduct disorder; k, number of items which form a particular scale; ODD, oppositional defiant disorder.

Thirty-eight interviews were video-recorded and seven were audio-recorded. Regarding ADHD diagnosis, 28 children were diagnosed with ADHD combined type, 14 children with ADHD inattentive type, two children with ADHD hyperactive-impulsive type and only one child was below the cut-off for any ADHD diagnosis. Hence, this sample does not capture the full spectrum of ADHD symptomatology but rather represents a clinical sample of different ADHD subtypes. Sample characteristics are reported in **Table 1**.

Interview Training

All interviewers who were involved in recruiting patients for the ESCAschool study were trained psychologists or educationists with a Master's degree, PhD candidates, or in training to become a child and adolescent psychotherapist/psychiatrist. During the ESCAschool study, all interviewers received a standardized training on administering and scoring the ILF-EXTERNAL, including watching a practice video. All interviewers were encouraged to consult their supervisor if they experienced any difficulties regarding the assessment with the ILF-EXTERNAL. Furthermore, two independent raters were asked to rate a subsample of 45 recordings of the ILF-EXTERNAL to obtain IRR. Both independent raters were PhD students at the University of Cologne and were completing their training as child and adolescent psychotherapists. In addition to the ESCAschool training on the ILF-EXTERNAL outlined above, both raters participated in a 1-day workshop in which they discussed the administration of the ILF-EXTERNAL, including detailed information on the scoring of each item. Both raters were then asked to independently code three practice videos randomly selected from the ESCAschool study, after which they received elaborate feedback from their supervisor and discussed potential difficulties when rating the recordings. Both raters were instructed not to discuss the interviews with each other during the rating process.

Statistical Analysis

All statistical analyses were performed using SPSS Version 26 (SPSS Inc, Chicago, IL, United States) if not stated otherwise. A first check of the data revealed no considerable floor or ceiling effects of the ILF-EXTERNAL item frequencies (except for the items that had been excluded previously). If more than 10% of the items forming a particular scale were missing, this scale was not computed for the affected participant due to a possible bias of the results (Bennett, 2001). This listwise exclusion criterion was also applied to the scales of the parent questionnaire data. A summary of the valid cases for each analysis is provided in the respective tables.

Besides descriptive statistics (mean scores, standard deviations) for all ILF-EXTERNAL scales, Cronbach's alpha was computed, with values of >0.70 indicating acceptable internal consistency (Nunnally, 1978). Moreover, the corrected item-total correlations were calculated, with values of >0.30 considered acceptable (Field, 2018).

The ICC coefficient (Shrout and Fleiss, 1979; McGraw and Wong, 1996) was computed to assess IRR between the interviewers and both independent raters. The ICC is one of the most common metrics when assessing IRR of continuous data (LeBreton and Senter, 2008; Hallgren, 2012; Koo and Li, 2016). It should be noted that different formulas exist, each involving distinct assumptions about their calculations and therefore leading to different interpretations (Koo and Li, 2016). We computed the ICC one-way random-effects, absolute agreement model for single rater/measurements ICC(1,1) as well as for measures based on a mean-rating ICC(1,3) with their 95% confidence intervals (CIs). The ICC one-way model was chosen because the physical distance between study centers prevented the same interviewer from measuring all participants which would otherwise qualify for the two-way measurement models (Koo and Li, 2016). Furthermore, we believe that the single rater/measurements model ICC(1,1) is more appropriate than average measures, given that the clinical outcome of ILF-EXTERNAL should be based on one clinician and not on the average information obtained from multiple clinicians (Koo and Li, 2016). Nevertheless, we also present average measurements ICC(1,*k*) to ensure comparability of our results across studies. We also calculated the IRR for both independent raters for all scales of the ILF-EXTERNAL using the two-way

random-effects models for single ICC(2,1) and for average ICC(2,2) measurements (Shrout and Fleiss, 1979; McGraw and Wong, 1996). To interpret ICC coefficients, different benchmarks are commonly cited. Cicchetti (1994) provided the following guidelines for interpreting ICC coefficients: poor $\leq$ 0.40; fair = 0.41–0.59; good = 0.60–0.74; and excellent $\geq$ 0.75. However, other authors proposed more stringent guidelines: poor $\leq$ 0.50; moderate = 0.51–0.75; good = 0.76–0.90; and excellent $\geq$ 0.91 (Koo and Li, 2016). The results are therefore presented using both 0.75 and 0.91 as interpretations of "excellent" reliability. Additionally, to obtain a further estimate on the degree of agreement, pairwise percent agreement was calculated based on integer scale scores (Wirtz and Caspar, 2002) using MATLAB and Statistics Toolbox Release 2018b. It should be noted that percentages of agreement do not correct for agreements that would be expected by chance and therefore, may overestimate the degree of agreement (Wirtz and Caspar, 2002).

Overall agreement on DSM-5 diagnoses was assessed using Fleiss' kappa (Fleiss, 1971) which is a statistical measure to assess agreement between multiple raters (i.e., the interviewers and both raters) on categorical variables (i.e., the presence or absence of a disorder). While Fleiss' kappa is a chance-corrected measure, it is dependent on the base rate of each disorder. Especially when the base rate of a disorder is low ($n < 10$), corresponding kappa values should only be interpreted with caution. The presence or absence of a DSM-5-based disorder was derived from the raw interview item scores by symptom counts. For example, if at least four items from the *ODD Symptoms* scale were scored with 2 or higher, these scores were considered to fulfill the diagnosis of ODD. Following common research practice, other exclusion criteria (such as not making the diagnosis of ODD in the presence of DMDD or such as only specifying limited prosocial emotions in the presence of CD) were ignored (Angold and Costello, 1995; de la Peña et al., 2018). Fleiss' kappa was calculated between the interviewers and two raters. To interpret kappa values, Landis and Koch (1977) suggested the following benchmarks: slight $\leq$ 0.20; fair = 0.21–0.40; moderate = 0.41–0.60; substantial = 0.61–0.80; almost perfect agreement $\geq$ 0.81.

Pearson product-moment correlations were computed between the ILF-EXTERNAL scales *ADHD Symptoms*, *ODD/CD Symptoms – short version* and the corresponding scales *ADHD Functional Impairment*, *ODD/CD Functional Impairment* in order to describe the relationship between symptom severity and the degree of functional impairment. To test for significant differences between pairs of correlations, the *cocor* software package for the R programming language (R 3.6.2) was applied (Diedenhofen and Musch, 2015). More specifically, we compared the magnitude of two dependent correlation coefficients with overlapping variables (i.e., the correlations have one variable in common) based on Steiger (1980) modification of Dunn and Clark (1969) *z*-transformation.

Additionally, Pearson product-moment correlations were computed between all ILF-EXTERNAL scales and the corresponding scales in the parent forms (FBB-ADHS; FBB-SSV) in order to evaluate convergent validity between clinical judgment and parent ratings. Two-sided paired samples *t*-tests were used group comparisons between the average scores of the ILF-EXTERNAL scales and the corresponding scales of the parent forms. This analysis allowed us to investigate whether clinician-rated scale scores on the ILF-EXTERNAL differed significantly from ratings on the corresponding parent-rated scales.

To further assess convergent and divergent validity, Pearson product-moment correlations were computed between the ILF-EXTERNAL scales and the eight syndrome scales as well as the *Externalizing Problems* and *Internalizing Problems* of the CBCL/6-18R. The R *cocor* package and Steiger's test (Steiger, 1980) were again applied to compare the magnitude of two dependent correlations. In particular, we determined whether the correlations of a particular ILF-EXTERNAL scale (e.g., *Inattention*) with the CBCL/6-18R broadband scales (*Externalizing Problems* and *Internalizing Problems*) differed significantly.

RESULTS

Scale Characteristics

Table 2 summarizes the mean scores, standard deviations, internal consistencies (Cronbach's alpha) and the ranges of the item-total correlations for all ILF-EXTERNAL scales. The lowest mean score was observed for the scale *CD Symptoms – short version* (M = 0.40, SD = 0.43), and the highest mean score for the scale *Inattention* (M = 1.95, SD = 0.48). As can be expected, given the clinical sample of children with ADHD symptoms, average scale scores were generally higher on the ADHD scales than on the ODD/CD scales. Cronbach's alpha coefficients for the ILF-EXTERNAL symptom scales were generally acceptable to good, with the exception of the scale *CD Symptoms – short version* (α = 0.60). The scales comprising *Functional Impairment* showed questionable internal consistency for ADHD (α = 0.62) and very good internal consistency for ODD/CD (α = 0.86). Item-total correlations were generally satisfactory ($0.21 \leq r_{it} \leq 0.71$) with some exceptions. The following items demonstrated item-total correlations below r_{it} = 0.30 (ADHD items: *A01 Careless, A06 Concentration, F05 Interferes with educational activities*. ODD/CD items: *B03 Cruel to animals, B05 Steals without confrontation, C04c Manipulates*). However, excluding any of these items did not noticeably change the Cronbach's alpha of the respective scales.

Interrater Reliability

Table 3 presents the IRR of the ILF-EXTERNAL scales, according to the ICC one-way random-effects, absolute agreement model for single ICC(1,1) and average measures ICC(1,3), their respective 95% confidence intervals, and pairwise percent agreement. Regarding the ILF-EXTERNAL symptom scales, all ICC(1,1) coefficients were greater than 0.75, indicating excellent IRR according to Cicchetti (1994) or, following a more stringent interpretation, good to excellent IRR (Koo and Li, 2016). Furthermore, all ICC(1,3) coefficients of the

TABLE 3 | Interrater reliability of the ILF-EXTERNAL scales.

Scale	ICC(1,1)	95% CI	ICC(1,3)	95% CI	Pairwise percent agreement	*n*
ADHD symptoms	0.91	0.87–0.95	0.97	0.95–0.98	88.1	45
– Inattention	0.83	0.74–0.90	0.94	0.89–0.96	85.2	45
– Hyperactivity-Impulsivity	0.95	0.91–0.97	0.98	0.97–0.99	82.2	45
ADHD functional impairment	0.89	0.82–0.94	0.96	0.93–0.98	80.7	39
ODD/CD symptoms–short version	0.94	0.90–0.96	0.98	0.97–0.99	91.1	45
– ODD symptoms	0.94	0.90–0.96	0.98	0.96–0.99	83.7	45
– CD symptoms–short version	0.90	0.85–0.94	0.97	0.94–0.98	88.2	44
Disruptive mood dysregulation	0.90	0.85–0.94	0.97	0.94–0.98	83.7	45
Limited prosocial emotions	0.93	0.89–0.96	0.98	0.96–0.99	86.7	41
ODD/CD functional impairment	0.92	0.86–0.96	0.97	0.95–0.99	85.2	31

ADHD, attention-deficit/hyperactivity disorder; CD, conduct disorder; CI, confidence interval; ICC, intraclass correlation; ICC(1,1), one-way random-effects, absolute agreement model for single rater/measurements; ICC(1,3), one-way random-effects, absolute agreement model based on a mean-rating; ODD, oppositional defiant disorder.

average measurement model were greater than 0.90, indicating excellent IRR of the ILF-EXTERNAL symptom scales (Cicchetti, 1994; Koo and Li, 2016). Regarding the ILF-EXTERNAL scales assessing functional impairment, both the *ADHD Functional Impairment* scale [ICC(1,1) = 0.89; ICC(1,3) = 0.96] and the *ODD/CD Functional Impairment* scale [ICC(1,1) = 0.92; ICC(1,3) = 0.97] demonstrated ICC values in the upper range, indicating very good to excellent IRR by the single and average measurement model (Cicchetti, 1994; Koo and Li, 2016). In addition, pairwise percent agreement was consistently higher than 80%, indicating high agreement between the interviewers and both raters. For the interested reader, results on the IRR on the item level are reported in the **Supplementary Table 1**. Furthermore, we calculated the IRR for both independent raters for all scales of the ILF-EXTERNAL using the two-way random-effects models for single ICC(2,1) and for average measurements ICC(2,2) (McGraw and Wong, 1996; Shrout and Fleiss, 1979). The results show that all ICC coefficients were 0.90 or greater using single and average measures, indicating excellent IRR of the ILF-EXTERNAL scales (Cicchetti, 1994; Koo and Li, 2016). The results are summarized in the **Supplementary Table 2**.

Agreement on DSM-5 Diagnoses

Table 4 presents overall agreement on DSM-5 diagnoses assessed using Fleiss' kappa values, their corresponding 95% confidence intervals, and pairwise percent agreement. Following the benchmarks of Landis and Koch (1977) diagnostic agreement ranged from fair (ADHD hyperactive-impulsive type: $\kappa = 0.38$), through moderate (DMDD: $\kappa = 0.55$), substantial (ADHD combined type: $\kappa = 0.71$; ADHD inattentive type: $\kappa = 0.71$; any ADHD: $\kappa = 0.74$) to almost perfect agreement (ODD: $\kappa = 0.82$; conduct disorder: $\kappa = 0.94$; with its specifier limited prosocial emotions: $\kappa = 0.82$). However, due to the low base rate of the diagnoses ADHD hyperactive-impulsive type ($n = 2$) and CD ($n = 6$) in the subsample, agreement on these two disorders should be interpreted with caution. In particular, pairwise percent agreement mainly seems to reflect agreement by chance. With regard to the remaining DSM-5 diagnoses, agreement could be estimated more reliably.

Correlations Between ILF-EXTERNAL Symptom Scales and Functional Impairment

Regarding the association between symptom severity and the degree of functional impairment, Pearson correlations revealed a moderate to large ($r = 0.50$) association between the scales *ADHD Symptoms* and *ADHD Functional Impairment*. In turn, there was a strong positive association between the scales *ODD/CD Symptoms – short version* and *ODD/CD Functional Impairment* ($r = 0.67$). Furthermore, the scale *ADHD Symptoms* correlated significantly more strongly with the scale on functional impairment associated with ADHD than with the *ODD/CD Functional Impairment* scale ($z = 3.92$, $p < 0.001$). Likewise, the scale *ODD/CD Symptoms* correlated significantly more strongly with the scale on ODD/CD-related functional impairment than with the *ADHD Functional Impairment* scale ($z = -5.41$, $p < 0.001$).

TABLE 4 | Agreement on DSM-5 diagnoses in the subsample for the analysis of interrater reliability.

DSM-5 Diagnosis	Fleiss' Kappa	95% CI	Pairwise percent agreement	Base rate *n*
Any ADHD	0.74	0.74–0.75	100	44
ADHD: combined type	0.71	0.70–0.71	86.7	29
ADHD: predominantly inattentive type	0.74	0.74–0.75	89.6	13
ADHD: predominantly hyperactive-impulsive type[a]	0.38	0.37–0.38	95.6	2
Oppositional defiant disorder	0.82	0.82–0.83	91.1	24
Disruptive mood dysregulation disorder	0.55	0.54–0.55	88.2	10
Conduct disorder[a]	0.94	0.94–0.95	98.5	6
– Specifier: Limited prosocial emotions	0.82	0.81–0.82	91.1	18

The presence of a disorder was derived from the raw interview item scores by symptom counts. [a]Diagnostic agreement should be interpreted with caution due to a low base rate ($n < 10$). Sample size $n = 45$: ADHD, attention-deficit/hyperactivity disorder.

TABLE 5 | Comparisons of the ILF-EXTERNAL scales and the corresponding parent forms (FBB-ADHS and FBB-SSV).

Scale	ILF-EXTERNAL: Mean (SD)	FBB (parents): Mean (SD)	*r*	Paired samples *t*-test	*p*	*n*
ADHD Symptoms	1.83 (0.48)	1.79 (0.56)	0.69***	$t(425) = 1.95$	0.051	426
– Inattention	1.98 (0.45)	2.03 (0.57)	0.58***	$t(420) = -2.35$	0.019	421
– Hyperactivity-Impulsivity	1.67 (0.71)	1.59 (0.73)	0.78***	$t(422) = 3.78$	<0.001	423
ADHD Functional Impairment	1.62 (0.57)	1.74 (0.69)	0.59***	$t(400) = -4.16$	<0.001	401
ODD/CD Symptoms—short version	0.84 (0.49)	0.97 (0.54)	0.74***	$t(411) = -6.84$	<0.001	412
– ODD Symptoms	1.12 (0.63)	1.38 (0.70)	0.72***	$t(404) = -10.21$	<0.001	405
– CD Symptoms—short version	0.40 (0.43)	0.42 (0.44)	0.65***	$t(407) = -1.18$	0.260	408
Disruptive Mood Dysregulation	1.06 (0.72)	1.21 (0.72)	0.67***	$t(411) = -4.72$	<0.001	412
Limited Prosocial Emotions	0.49 (0.41)	0.68 (0.54)	0.63***	$t(406) = -8.96$	<0.001	407
ODD/CD Functional Impairment	0.94 (0.78)	1.39 (0.82)	0.57***	$t(380) = -11.75$	<0.001	381

*ADHD, attention-deficit/hyperactivity disorder; CD, conduct disorder; ODD, oppositional defiant disorder; ILF-EXTERNAL, Clinical Parent Interview for Diagnosing Externalizing Disorders in Children and Adolescents; FBB, parent-rated symptom checklists for the assessment of ADHD symptoms and symptoms of disruptive behavior disorders; ***$p < 0.001$.*

Convergent and Divergent Validity

Table 5 compares the ILF-EXTERNAL scales and the corresponding scales of the parent forms (FBB-ADHS; FBB-SSV). Pearson correlations were moderate to high and significant ($0.57 \leq r \leq 0.78$, $p < 0.001$), indicating convergent validity between clinical judgment and parent ratings. Overall, ratings on most ILF-EXTERNAL scale scores differed significantly from ratings on the corresponding scales of the parent forms ($p < 0.05$), with the exception of the scales *ADHD Symptoms* ($p = 0.051$) and *CD Symptoms – short version* ($p = 0.260$). Furthermore, mean scale scores on the parent forms were higher than the corresponding clinical judgment (exceptions: *ADHD Symptoms* and *Hyperactivity-Impulsivity*).

In addition, **Table 6** summarizes Pearson correlations between the ILF-EXTERNAL scales and the eight CBCL/6-18R syndrome scales as well as the CBCL/6-18R broadband scales *Externalizing Problems* and *Internalizing Problems*. Overall, correlations between the ILF-EXTERNAL scales and the CBCL *Externalizing Problems* were moderate to high and significant ($0.33 \leq r \leq 0.69$, $p < 0.001$). As can be expected, the highest observed correlations of the CBCL *Externalizing Problems* were with the ILF-EXTERNAL scales *ODD Symptoms*, *CD Symptoms – short version*, and *ODD/CD Symptoms – short version* scales ($0.58 \leq r \leq 0.69$). Furthermore, the ILF-EXTERNAL *Inattention* scale was most strongly associated with the CBCL syndrome scale *Attention Problems* ($r = 0.39$). As can be expected, the ILF-EXTERNAL scales were more strongly associated with the CBCL *Externalizing Problems* than the *Internalizing Problems*. When comparing the correlation coefficients of both CBCL problem scales, we found that all ILF-EXTERNAL scales were significantly more strongly associated with the CBCL *Externalizing Problems* ($0.001 \leq p \leq 0.005$). Taken together, these results provide support for the convergent and divergent validity of the ILF-EXTERNAL.

DISCUSSION

This study presents the DSM-5-based, semi-structured, clinical parent interview ILF-EXTERNAL and its psychometric properties in a clinical sample of school-age children with ADHD symptoms. The results suggest that the ILF-EXTERNAL is a promising and overall reliable and valid clinical interview for diagnosing externalizing disorders in children and adolescents.

Regarding scale reliability, Cronbach's alpha coefficients for the ILF-EXTERNAL scales were generally acceptable to good. Accordingly, those items which were aggregated to form a particular scale predominantly seem to measure a common construct. One exception is the *CD Symptoms – short version* scale ($\alpha = 0.60$). Similar internal consistency of the *CD Symptoms* scale was reported for the DISC version 2.3 ($\alpha = 0.59$, Frick et al., 2010). We believe that for the following reasons, this rather low internal consistency is unsurprising: First, we excluded the items B06 to B15, assessing aggressive and antisocial symptoms from the age of 11, which resulted in a shortened scale of only five items. Second, with a low mean score ($M = 0.40$; $SD = 0.43$), the scores of the remaining items of this shortened scale displayed a skewed distribution. Third, these symptoms represent a heterogeneous group of symptoms, which may have impaired the reliability of this scale (Frick et al., 2010). Similarly, the *ADHD Functional Impairment* scale demonstrated low internal consistency ($\alpha = 0.62$), which might also be explained by the heterogeneity of the items. However, the *ODD/CD Functional Impairment* scale showed very good internal consistency ($\alpha = 0.86$). In addition, item-total correlations were generally satisfactory with some exceptions. Although some items demonstrated item-total correlations below $r_{it} = 0.30$, excluding any of these items did not noticeably change the Cronbach's alpha of the respective scales.

Having calculated the ICC one-way random-effects model for single ICC(1,1) and average ICC(1,3) measurements, ICC coefficients demonstrated "very good" to "excellent" IRR for all scales (Cicchetti, 1994; Koo and Li, 2016). Most IRR studies on broadband clinical interviews assessing children and adolescents did not provide IRR results on the scale level. One previous study assessed externalizing symptoms in children and adolescents using a modified ADHD-ODD scale of the K-SADS (Jans et al., 2009). This modified scale was based on a dichotomous assessment of the DSM-IV-based ADHD and ODD criteria, leading to a sum score. Pearson correlations revealed a strong positive association ($r = 0.98$) between the sum

TABLE 6 | Correlations of the ILF-EXTERNAL scales and the Child Behavior Checklist (CBCL/6-18) syndrome scales.

	CBCL/6-18											
Scale	**Anxious / Depressed**	**Withdrawn / Depressed**	**Somatic complaints**	**Social problems**	**Thought problems**	**Attention problems**	**Rule-Breaking behavior**	**Aggressive behavior**	**Externalizing problems**	**Internalizing problems**	**Externalizing vs. Internalizing problems**	
											z	***p***
ADHD Symptoms	0.21***	0.02	0.12*	0.37***	0.26***	0.31***	0.38***	0.50***	0.49***	0.17***	6.65	<0.001
– Inattention	0.16***	0.16***	0.14**	0.29***	0.21***	0.39***	0.29***	0.32***	0.33***	0.19***	2.79	0.005
– Hyperactivity-Impulsivity	0.18***	−0.07	0.08	0.31***	0.21***	0.17***	0.32***	0.47***	0.44***	0.10*	6.90	<0.001
ODD/CD Symptoms–short version	0.22***	0.12*	0.15**	0.35***	0.21***	0.21***	0.59***	0.68***	0.69***	0.21***	11.17	<0.001
– ODD Symptoms	0.24***	0.13**	0.15**	0.35***	0.21***	0.21***	0.52***	0.64***	0.64***	0.23***	9.27	<0.001
– CD Symptoms–short version	0.11*	0.06	0.08	0.24***	0.14**	0.16**	0.56***	0.54***	0.58***	0.11*	10.03	<0.001
Disruptive Mood Dysregulation	0.26***	0.15**	0.13**	0.32***	0.21***	0.18***	0.39***	0.59***	0.55***	0.24***	6.69	<0.001
Limited Prosocial Emotions	0.20***	0.21***	0.10*	0.29***	0.21***	0.25***	0.44***	0.43***	0.46***	0.22***	4.99	<0.001

*Sample size n = 407; ADHD, attention-deficit/hyperactivity disorder; CD, conduct disorder; ODD, oppositional defiant disorder; ILF-EXTERNAL, Clinical Parent Interview for Diagnosing Externalizing Disorders in Children and Adolescents. *p < 0.05; **p < 0.01; and ***p < 0.001.*

scores of the interviewers and the sum scores from independent raters. However, it should be noted that ICC might be a more appropriate measure to assess IRR than Pearson correlations. While the Pearson correlation coefficient indicates the strength of the linear relationship between two variables, a high correlation may be observed even though agreement is poor (Bland and Altman, 1986; Gisev et al., 2013). Another study assessed IRR of the ADHD subdimensions in the K-SADS using ICC (Kariuki et al., 2018). The results indicated moderate to good IRR for the inattentive subtype (ICC = 0.76), hyperactive-impulsive subtype (ICC = 0.41), combined type (ICC = 0.77), and any ADHD type (ICC = 0.64). While the authors calculated the one-way random-effects model, it remains unclear whether they relied on single or average measurements, which limits the interpretation of their results. Although our ICC coefficients were consistently higher on all ADHD scales, a comparison with the aforementioned study must be treated with caution for the following reasons: First, the authors validated the ADHD subdomains in a community sample, while our results were based on clinically referred children. Second, the authors only obtained IRR estimates from 20 children, while we empirically calculated our required sample size and based our IRR results on twice as many children. Overall, our study demonstrates high IRR and addresses the aforementioned research gap, providing valuable information regarding the psychometric quality on the scale level. These findings were largely confirmed even on the single-item level (see **Supplementary Table 1**).

Diagnostic agreement between the interviewers and both independent raters was "substantial" to "almost perfect" for most disorders with the exceptions of ADHD hyperactive-impulsive type and DMDD (Landis and Koch, 1977). With regard to diagnosing ADHD and its subtypes, we found substantial agreement for any ADHD diagnosis, for ADHD combined type, and for ADHD inattentive type. However, these results should be discussed within the scope of the subsample. The composition of this subsample may have influenced agreement estimates, particularly because of the high base rate of ADHD diagnoses (i.e., 44/45 children). Although both independent raters were not aware of this high base rate, the sole fact that almost all children exhibited clinically relevant symptoms (i.e., scorings of 2 or 3 on each item) may have led to an uneven distribution of item scorings and thus, possible overestimation of agreement on ADHD diagnoses. For example, the "perfect" pairwise agreement of 100% for any ADHD diagnosis ($\kappa = 0.74$) rather seems to reflect an overestimation of agreement due to sampling issues. Concerning diagnostic agreement on ADHD hyperactive-impulsive type, we found rather low Fleiss' kappa agreement ($\kappa = 0.38$) but almost perfect pairwise agreement (95.6%). Although this finding might seem somewhat perplexing, it can be explained as follows: Considering that Fleiss' kappa is influenced by the base rate of observations (Wirtz and Caspar, 2002), the agreement on ADHD hyperactive-impulsive type seems to primarily reflect sampling issues due its very low base rate ($n = 2$) in our subsample. This low base rate, in turn, influences pairwise percent agreement which does not correct for agreement that would be expected by chance. For example, even if both raters agreed on *no* ADHD hyperactive-impulsive diagnosis

for all 45 participants, they still would have demonstrated agreement in 43/45 cases.

As a newly developed clinical interview with a semi-structured format, it is particularly essential to compare diagnostic interrater agreement of the ILF-EXTERNAL with that from other semi-structured interviews. The degree of agreement on any ADHD diagnosis was comparable with other findings in clinical samples using the K-SADS ($0.42 \leq \kappa \leq 0.92$; Kim et al., 2004; Ghanizadeh et al., 2006; Ulloa et al., 2006; de la Peña et al., 2018; Nishiyama et al., 2020). Furthermore, our results regarding diagnostic agreement on ADHD subtypes were also relatable to previous literature. Having calculated kappa agreement using the MINI-KID interview in a clinical sample, Sheehan et al. (2010) reported almost perfect agreement for ADHD combined type ($\kappa = 0.90$) and ADHD inattentive type ($\kappa = 0.93$) and substantial agreement for ADHD hyperactive-impulsive type ($\kappa = 0.65$). Interestingly, high diagnostic agreement on diagnosing ADHD combined type ($\kappa = 0.86$) and ADHD inattentive type ($\kappa = 0.78$) was also reported in a clinical sample of children with ADHD symptoms (Power et al., 2004).

With regard to comorbid externalizing disorders, the degree of diagnostic agreement was comparable with other findings in clinical samples using the K-SADS for ODD ($0.69 \leq \kappa \leq 0.80$; Ghanizadeh et al., 2006; de la Peña et al., 2018) DMDD ($\kappa = 0.53$; de la Peña et al., 2018), and CD ($0.78 \leq \kappa \leq 1.0$; Ghanizadeh et al., 2006; Ulloa et al., 2006; de la Peña et al., 2018). Although our results concerning CD should be interpreted with caution due to its low base rate in the subsample ($n = 6$), these results were also in line with previous studies reporting the highest agreement on this diagnosis (Ghanizadeh et al., 2006; Ulloa et al., 2006). We suggest that this finding may be attributable to the clinical presentation of CD symptoms, which are clear to observe and unambiguous to score. Agreement on the specifier limited prosocial emotions was classified if symptoms in at least two out of four categories were considered as clinically relevant. While previous research observed fair agreement ($\kappa = 0.29$; de la Peña et al., 2018) we found very high diagnostic agreement on this specifier ($\kappa = 0.82$), which again, may be attributable to our sample characteristics.

The ranges of diagnostic agreement reported in the literature might arise from differences in the administration of the interview (e.g., parents or children as primary informant), the respective study samples (e.g., children or adolescents), methodological issues (e.g., number of raters or amount of training received on administering the interview), or the sample population (community vs. clinical) and its characteristics (e.g., base rates of disorders). Notably, diagnostic agreement is often higher in clinical than in community samples (Chen et al., 2017). One basic criticism of clinical samples is that they typically only include patients with clear and severe symptoms. Consequently, the patients' symptoms can be easily recognized and scored, which may lead to overestimated reliability results, an effect which is also referred to as spectrum bias (Ranshoff and Feinstein, 1978).

Overall, while these reliability results and their corresponding coefficients yield important empirical findings, these labels do not indicate their practical or clinical relevance (Kottner et al., 2011). In other words, even though we obtained very good to excellent IRR and diagnostic agreement results, discrepancies between ratings nevertheless occurred, which warrant further discussion. We critically explored discrepancies between the interviewers and both raters and propose the following reasons for rater disagreement: (1) In terms of the administration of the ILF-EXTERNAL, we noted that some interviewers explored the frequency and intensity of each symptom more thoroughly than did others. This possible lack of clinical information may have affected the scorings of both independent raters. Moreover, (2) noise disturbances during the recordings may have affected the raters, and (3) information variance (i.e., the interviewers may have integrated information prior to the interview into their ratings, as well as 4) interpretation variance (i.e., different raters may have subjective ideas about weighting of symptoms) might have arisen (Hoyer and Knappe, 2012).

A further finding was that higher symptom severity was associated with a higher degree of functional impairment. This result highlights the importance of the current DSM practice of considering a clinical significance criterion (Spitzer and Wakefield, 1999) which requires symptoms to be associated with clinically significant psychological strain and functional impairment in social, occupational, or other areas of life to warrant a diagnosis (American Psychiatric Association, 2013). Results from a large meta-analysis confirmed the relationship between ADHD subtypes and multiple domains of functional impairment (Willcutt et al., 2012).

Regarding convergent and divergent validity, we found moderate to strong correlations between the ILF-EXTERNAL scales and the scales of the German CBCL/6-18R covering similar symptoms. Furthermore, the ILF-EXTERNAL scales were significantly more strongly associated with the CBCL *Externalizing Problems* than with the *Internalizing Problems*, indicating construct validity. These results are largely consistent with previous studies reporting small to moderate relations between the CBCL and clinical diagnoses from semi-structured interviews for the assessment of clinical symptoms in children and adolescents (Kim et al., 2004; Birmaher et al., 2009; Brasil and Bordin, 2010; Chen et al., 2017). Moreover, correlations of the ILF-EXTERNAL scales with the corresponding CBCL scales were generally higher than correlations with the non-corresponding CBCL scales. Similar findings have also been reported in the community population (Kim et al., 2004; Birmaher et al., 2009; Chen et al., 2017). However, limitations of these findings are that they often rely solely on broad diagnostic categories such as "ADHD" without specification of its subtypes (Birmaher et al., 2009; Chen et al., 2017), "any disruptive disorder" (Brasil and Bordin, 2010), or that their results are based on small (i.e., less than $N = 100$) sample sizes (Kim et al., 2004; Brasil and Bordin, 2010). We therefore extended these findings by reporting validity results on diagnostic scales in a larger sample. A further strength of our study is that we included parent forms (FBB-ADHS; FBB-SSV) which cover the same DSM-5 symptoms as the ILF-EXTERNAL. This distinguishing and novel characteristic allowed us to specifically compare ratings between parental and clinical judgments. While our results indicate moderate to substantial convergence between parent ratings and clinical judgments, we believe that this convergence

is not sufficiently strong to argue that raters could be seen as interchangeable. In contrast, Boyle et al. (2017) challenged that structured clinical interviews may be replaced by self-completed problem checklists as a time- and cost-effective alternative. One basic criticism was that *"the dependence on respondents in these interviews is similar to the dependence on respondents completing a checklist on their own except for the potential error introduced by interviewer characteristics and interviewer–respondent exchanges"* (Boyle et al., 2017, p. 2). While we agree with this view inasmuch as clinical interviews should provide additional value to questionnaire data such as problem checklists, close inspection of our results revealed the following: Although we found moderate to large correlations between clinician and parent ratings, comparisons of the absolute scale scores revealed significant differences between the ratings on several scales. This indicates that both perspectives are complementary and that both are necessary for an informed clinical diagnosis. On top of that, similar recommendations are made by the German interdisciplinary evidence- and consensus-based (S3) guidelines on the clinical assessment of ADHD (Association of Scientific Medical Societies in Germany AWMF, 2018).

In terms of limitations, one drawback of the present study is that parents were the only informants for both the interview and the questionnaires. Hence, no information was available from the children themselves. However, we believe that this limitation is surmountable given that parents are typically better informants regarding their children's externalizing behavior problems than their children.

Another significant aspect to consider is the composition of the subsample for the analysis of IRR. We concede that the high base rate of ADHD diagnoses may have influenced interrater agreement. As percentages agreement do not correct for agreements that would be expected by chance, they may overestimate the degree of agreement. In particular, the almost perfect percentages of agreement on some diagnoses rather seem to reflect an overestimation due to chance agreement and sampling issues.

While agreement between parent and teacher ratings on childhood diagnoses is typically quite low (Willcutt et al., 2012) studies investigating interrater agreement between interviewers using clinical interviews yield higher estimates. We concede that these higher agreement estimates may be explained as follows: (1) Intensive rater trainings on the administration and scoring of a clinical interview may lead to more homogenous ratings, and thus, higher rates of agreement. (2) Within research settings, it is common practice to classify agreement on diagnoses based on raw interview item scores by symptom counts. However, this approach may overestimate diagnostic agreement because additional criteria for an informed clinical diagnosis are not further considered. (3) As the interviews are video- or audio-recorded, the interviewers and raters have the exact same informants (e.g., parents) with the exact same information. This approach results in higher agreement estimates compared to other forms of reliability, e.g., test-retest reliability where the same informant is interviewed twice but may provide different information (Angold and Costello, 1995).

Finally, the factor structure of the ILF-EXTERNAL has not yet been validated. While this clinical interview comprises a set of items with each item exploring a DSM-5 symptom criterion, it remains unclear whether this DSM-5-based factor structure can be replicated empirically. For this reason, a follow-up study exploring the factor structure of the ILF-EXTERNAL using correlated factor models and bifactor models is planned. Nevertheless, it should be noted that the factor structure of the corresponding DISYPS parent forms, FBB-ADHS and FBB-SSV, has been confirmed (Erhart et al., 2008; Görtz-Dorten et al., 2014).

We suggest that future studies evaluating psychometric properties of structured clinical interviews should include ratings of symptom severity on the scale level as part of a dimensional approach. Ideally, specific aspects covering functioning and psychological strain could also be included.

CONCLUSION

The aim of this study was to assess the reliability and validity of a DSM-5-based, semi-structured parent interview for diagnosing externalizing disorders in children and adolescents. In clinically referred, school-age children, the ILF-EXTERNAL demonstrates sound psychometric properties in terms of IRR on the item and on the scale level, rater agreement on most DSM-5 diagnoses, internal consistency, and convergent and divergent validity. In line with current literature and the DSM practice to consider functional impairment as prerequisite for making a diagnosis, higher symptom severity was associated with a higher degree of functional impairment. Having developed a comprehensive set of clinical parent and patient interviews (DISYPS-ILF), we hope to contribute to a high-quality standard of diagnosing mental disorders in children and adolescents.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethical approval has been obtained for the study centre Cologne by the University of Cologne (ID 15–216), for the study centre Essen by the University Duisburg-Essen (ID 17–7404-BO), for the study centre Hamm by the Ruhr-University Bochum (ID 15–5564), for the study centre Göttingen by the University Medical Center Göttingen (ID 3/3/17), for the study centre Mainz by the Federal Medical Association of Rhineland-Palatinate (ID 837.237.17 [11071]), for the study centre Mannheim by the Ruprecht-Karls- University Heidelberg (ID 2015-646 N-MA), for the study centre Marburg by the Philipps-University Marburg (ID Studie 03/16), for the study centre Tübingen by the Eberhard-Karls-University Tübingen (ID 791/2015BO2), and for the study centre Würzburg by the

Julius-Maximilians-University Würzburg (ID 332/15_z). Written informed consent to participate in this study was provided by the participants legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

A-KTh developed the first draft of the manuscript, rated the data (Rater 1), analyzed the data, and was involved in the recruitment and data acquisition of the ESCAschool study site in Cologne and was co-author of the DISYPS-ILF. MD was principal investigator of the ESCAschool trial, developed the basic ESCAschool study design, was part of the ESCAlife-Consortium, was head of the Cologne recruiting center for the ESCA children trials (ESCApreschool, ESCAschool, and ESCAadol), developed the DISYPS-III system, and the DISYPS-ILF, and critically revised the manuscript. AG-D developed the DISYPS-III system and the DISYPS-ILF, was involved in rater training of the ESCAschool study, and critically revised the manuscript. PA was a scientific staff member at the Cologne study site, was involved in patient recruitment and data acquisition, and critically revised the manuscript. CD heads the telephone-assisted self-help trial performed in one treatment arm of the ESCAschool study and critically revised the manuscript. NG rated the data (Rater 2) and critically revised the manuscript. CH was involved in the development of the ESCAschool research proposal and organization of the study, and critically revised the manuscript. LJ and A-KTr coordinated ESCAschool, contributed to the management and organization of the study, and critically revised the manuscript. EW coordinated the Cologne study site, contributed to the implementation of the study, provided supervision for patient treatment, and critically revised the manuscript. TB coordinated the ESCAlife consortium and was the co-principal investigator in the ESCAschool study, and made substantial contributions to the conception and design of the ESCAschool research proposal, heads the Mannheim study site of ESCAschool, and critically revised the manuscript. DB made substantial contributions to the conception and design of the ESCAschool research proposal and the neuropsychological research battery, heads the sub-project ESCAbrain, and critically revised the manuscript. SM heads the Mannheim study site of ESCAschool, coordinates the ESCAlife consortium, made substantial contributions to the conception of the neuropsychological research battery and the neurofeedback protocol, and critically revised the manuscript. SH coordinated the ESCAlife consortium, made substantial contributions to the conception of the neuropsychological research battery and the neurofeedback protocol, and critically revised the manuscript. KB heads the Marburg study site of ESCAschool, was PI for the ESCApreschool study, is a member of the ESCAlife consortium, was a co-applicant of the ESCAlife research project, made substantial contributions to the conception and design of the ESCAlife study, and critically revised the manuscript. JK coordinated ESCAschool at the Marburg study site, contributed to the management and organization of the study, was involved in patient recruitment and data acquisition, and critically revised the manuscript. JH heads the Essen study site of ESCAschool, made substantial contributions to the conception and design of the ESCAschool study, and critically revised the manuscript. JW coordinated ESCAschool at the Essen study site, contributed to the management and organization of the study, and critically revised the manuscript. MHo heads the Hamm/Bochum study site of ESCAschool, was involved in the development of the ESCAschool research proposal, and critically revised the manuscript. TL heads the research department of the Hamm/Bochum study site of ESCAschool, contributed to the implementation of the study in Hamm, supervised the realization of the trial in Hamm, and critically revised the manuscript. MHu heads the Mainz study site of ESCAschool, was involved in the implementation of the ESCAlife projects, and critically revised the manuscript. MR was involved in the planning of the ESCAlife research projects and application for funding, was co-PI for the ESCAadol trial, was involved in the implementation of the ESCAlife projects at the Würzburg study site, and critically revised the manuscript. TJ was involved in the planning of the ESCAlife research projects and application for funding, was co-PI for the ESCAadol trial and was involved in the implementation of the ESCAlife projects at the Würzburg study site, and critically revised the manuscript. JG coordinated the Würzburg study site, contributed to the management and organization of the ESCAlife projects, and critically revised the manuscript. LP heads the Göttingen study site of ESCAschool, was involved in the implementation of the ESCAlife projects, and critically revised the manuscript. HU coordinated ESCAschool at the Göttingen study site, contributes to the management and organization of the study, was involved in patient recruitment and data acquisition, and critically revised the manuscript. TR heads the Tübingen study site of ESCAschool, contributed to the implementation of the ESCAschool study, and critically revised the manuscript. UD coordinates the Tübingen study site of ESCAschool, contributes to the management and organization of the study, was involved in patient recruitment and data acquisition, and critically revised the manuscript. All authors gave final approval of the last version of the manuscript and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

FUNDING

The ESCAschool study was funded by the German Federal Ministry of Education and Research (BMBF) Grant 01EE1408. This work was supported by the research consortium on ADHD, ESCA-Life, funded by the German Federal Ministry of Education and Research (FKZ 01EE1408E).

ACKNOWLEDGMENTS

We thank all staff members who participated in the ESCA-school study in nine different study recruiting centers and who were involved in recruitment, study organization and management, blind rating, diagnostics and therapy, as well as the ESCAbrain study group and the CTU. These are (sorted by

recruiting center affiliation and in alphabetical order): Study center Cologne: Ruth Beckmann, Caroline Betcher, Jana Bretthauer, Corinna Broekmans, Yvonne Buntrock, Mareike Bürger, Dominique Deli, Elvan Dogan, Lisa Firmenich, Saskia Heintzmann, Katrin Floß, Anna Fronhofen, Claudia Ginsberg, Jana Harenkamp, Julia Kellner, Claudia Kinnen, Marie-Theres Klemp, Hedda Körner, Constanze Kitzig, Katrin Krugmann, Jessica Leite, Sarah Meier, Ronja Meingast, Judith Mühlenmeister, Lea Ostrowski, Imke Partzsch, Jeannine Pawliczek, Daniela Perri, Judith Ratayczak, Stephanie Schürmann, Ellen Settegast, Johanna Spicher, Susanne Stollewerk, Katharina Verhülsdonk, Laura Wähnke, Christine Wälde, Annabelle Warschburger, Friederike Waschau, Tanja Wolff Metternich-Kaizman, and Sara Zaplana Labarga. Study center Essen: We would like to gratefully thank all the participating patients and their families. Study center Göttingen: Fatmeh Al-Debi and Jan Schulz. Study center Hamm: Miriam Davids, Natalie Deux, Mareen Dökel, Silvia Eißing, Regina Herdering, Carina Huhn, Lara Kaffke, Inken Kirschbaum-Lesch, Franziska Martin, Nina Müller, Nicola Nolting, Daniela Pingel, Carina Schulz, Karen Schumann, Birthe Wagner, and Daniela Walkenhorst. Study center Mainz: We would also like to gratefully thank all the participating patients and their families. Study center Mannheim: Karina Abenova, Brigitta Gehrig, Monja Groh, Raphael Gutzweiler, Christine Igel, Anna Kaiser, Theresa Nickel, Angelina Samaras, Anne Schreiner, Anne Schröter, Marie-Therese Steiner, Nina Trautmann, Lisa Wagner, Matthias Winkler, and Mirjam Ziegler. Study center Marburg: Anette Becker, Viktoria Birkenstock, Doreen Blume, Anita Dehnert, Silas Deutsch, Sabine Finkenstein, Charlotte Finger, Claudia Freitag, Nicole Grau, Wiebke Haberhausen, Julia Häusser, Daria Kasperzack, Franziska Körfgen, Johannes Kresse, Jana Langhammer, Christopher Mann, Tanja Mingebach, Thomas Mooz, Jens Pfeiffer, Maren Rumpf, Alisa Samel, Bastian Schrott, Anna Schuh, Isabell Schulz, Thomas Stehr, Rebecca Stein, Anne von Stiphout, Linda Weber, Svenja-Katharina Weber, Anne-Kathrin Wermter, and Maximiliane Werther. Study center Tübingen: Yael Amling, Anna Hertle, Natalie Herrmann, Melinda Mross, Anja Pascher, Alena Roth, Priska Schneider, Anja Schöllhorn, and Ida Steinacker. Study center Würzburg: Nadja Becker, Christoph Biohlawek, Lea Brosig, Kathrin Delius, Adrian Dernbach, Anja Diehl, Simone Dorsch, Lisa Eidenschink, Stefanie Fekete, Zuzana Fouskova, Julia Gläser, Jana Halmagyi, Verena Hartlieb, Silke Hauck, Julia Häusler, Christin Heim, Sonja Hetterich, Michaela Linke, Annette Nowak, Corinna Otte, Isabel Paul, Katharina Peters, Hanna Schwarz, Carolin Steinmetz, Kristin Wehrmann, Johannes Weigl, and Lea Zwilling. Additional ESCAbrain staff members, who were involved in ESCA-school: Clinical Trials Unit Freiburg: Veronika Frommberger, Alexander Hellmer, Lydia Herbstritt, Maria Huber, Carolin Jenkner, Stefanie Lade, Patrick Müller, Barbara Schilling, Carla Schneider, and Ralf Tostmann. Further thanks to Sarah Mannion for English-language proofreading.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01840/full#supplementary-material

REFERENCES

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4–18 and 1991 Profile*. Burlington, VT: University of Vermont.

Achenbach, T. M., and Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles: An Integrated System of Multi-Informant Assessment*. Burlington, VT: ASEBA.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*, 5th Edn. Washington, DC: APA.

Angold, A., and Costello, E. J. (1995). A test—retest reliability study of child-reported psychiatric symptoms and diagnoses using the child and adolescent psychiatric assessment (CAPA-C). *Psychol. Med.* 25, 755–762. doi: 10.1017/S0033291700034991

Angold, A., and Costello, E. J. (2000). The child and adolescent psychiatric assessment (CAPA). *J. Am. Acad. Child Adolesc. Psychiatry* 39, 39–48. doi: 10.1097/00004583-200001000-00015

Arbeitsgruppe Deutsche Child Behavior Checklist (1998). "Elternfragebogen über das Verhalten von Kindern und Jugendlichen; deutsche Bearbeitung der Child Behavior Checklist (CBCL/4-18)," in *Einführung und Anleitung zur Handauswertung. 2. Auflage mit Deutschen Normen*, eds B. V. M. Döpfner, J. Plück, S. Bölte, K. Lenz, P. Melchers, and K. Heim (Köln: Arbeitsgruppe Kinder Jugend- und Familiendiagnostik [KJFD]).

Association of Scientific Medical Societies in Germany AWMF (2018). Interdisciplinary Evidence- and Consensus-Based (S3) Guideline "Attention Deficit/Hyperactivity Disorder (ADHD) in Children, Young People and Adults" [in German]. Available at: https://www.awmf.org/uploads/tx_szleitlinien/028-045l_S3_ADHS_2018-06.pdf (accessed July 15, 2020).

Becker, K., Banaschewski, T., Brandeis, D., Dose, C., Hautmann, C., Holtmann, M., et al. (2020). Individualised stepwise adaptive treatment for 3 – 6-year-old preschool children impaired by attention-deficit / hyperactivity disorder (ESCApreschool): study protocol of an adaptive intervention study including two randomised controlled trials within the Consortium ESCAlife. *Trials* 21:56. doi: 10.1186/s13063-019-3872-8

Bennett, D. A. (2001). How can I deal with missing data in my study? *Aust. N. Z. J. Public Health* 25, 464–469. doi: 10.1111/j.1467-842x.2001.tb00294.x

Birmaher, B., Ehmann, M., Axelson, D. A., Goldstein, B. I., Monk, K., Kalas, C., et al. (2009). Schedule for affective disorders and schizophrenia for school-age children (K-SADS-PL) for the assessment of preschool children - A preliminary psychometric study. *J. Psychiatr. Res.* 43, 680–686. doi: 10.1016/j.jpsychires.2008.10.003

Bland, J. M., and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 327, 307–310. doi: 10.1016/s0140-6736(86)90837-8

Boyle, M. H., Duncan, L., Georgiades, K., Bennett, K., Gonzalez, A., Van Lieshout, R. J., et al. (2017). Classifying child and adolescent psychiatric disorder by problem checklists and standardized interviews. *Int. J. Methods Psychiatr. Res.* 26:e1544. doi: 10.1002/mpr.1544

Brasil, H. H. A., and Bordin, I. A. (2010). Convergent validity of K-SADS-PL by comparison with CBCL in a Portuguese speaking outpatient population. *BMC Psychiatry* 10:83. doi: 10.1186/1471-244X-10-83

Chen, Y. L., Shen, L. J., and Gau, S. S. F. (2017). The Mandarin version of the Kiddie-Schedule for Affective Disorders and Schizophrenia-Epidemiological version for DSM–5 – A psychometric study. *J. Formos. Med. Assoc.* 116, 671–678. doi: 10.1016/j.jfma.2017.06.013

Cicchetti, D. V. (1994). Guidlines, criteria, and rules of thumb for evalauting normed and standardized assessment instruments in psychology. *Psychol. Assess.* 6, 284–290. doi: 10.1037/1040-3590.6.4.284

Coghill, D., and Sonuga-Barke, E. J. S. (2012). Annual research review: categories versus dimensions in the classification and conceptualisation of child and adolescent mental disorders - Implications of recent empirical study. *J. Child*

Psychol. Psychiatry Allied Discip. 53, 469–489. doi: 10.1111/j.1469-7610.2011.02511.x

de la Peña, F. R., Villavicencio, L. R., Palacio, J. D., Félix, F. J., Larraguibel, M., Viola, L., et al. (2018). Validity and reliability of the kiddie schedule for affective disorders and schizophrenia present and lifetime version DSM-5 (K-SADS-PL-5) Spanish version. *BMC Psychiatry* 18:193. doi: 10.1186/s12888-018-1773-0

Diedenhofen, B., and Musch, J. (2015). Cocor: a comprehensive solution for the statistical comparison of correlations. *PLoS One* 10:e0121945. doi: 10.1371/journal.pone.0121945

Döpfner, M., Breuer, D., Wille, N., Erhart, M., and Ravens-Sieberer, U. (2008). How often do children meet ICD-10/DSM-IV criteria of attention deficit-/hyperactivity disorder and hyperkinetic disorder? Parent-based prevalence rates in a national sample - Results of the BELLA study. *Eur. Child Adolesc. Psychiatry* 17(Suppl. 1), 59–70. doi: 10.1007/s00787-008-1007-y

Döpfner, M., and Görtz-Dorten, A. (2017). *Diagnostik-System für Psychische Störungen nach ICD-10 und DSM-5 für Kinder und Jugendliche – III [Diagnostic System of Mental Disorders in Children and Adolescents based on the ICD-10 and DSM-5] (DISYPS–III)*. Göttingen: Hogrefe.

Döpfner, M., Hautmann, C., Dose, C., Banaschewski, T., Becker, K., Brandeis, D., et al. (2017). ESCAschool study: trial protocol of an adaptive treatment approach for school-age children with ADHD including two randomised trials. *BMC Psychiatry* 17:269. doi: 10.1186/s12888-017-1433-9

Döpfner, M., and Petermann, F. (2012). *Diagnostik Psychischer Störungen im Kindes-und Jugendalter (Vol. 2)*. Göttingen: Hogrefe Verlag.

Döpfner, M., Plück, J., Kinnen, C., and Arbeitsgruppe Deutsche Child Behavior Checklist, (2014). *CBCL Handbuch-Schulalter. Manual zum Elternfragebogen über das Verhalten von Kindern und Jugendlichen, (CBCL/6-18R), zum Lehrerfragebogen über das Verhalten von Kindern und Jugendlichen (TRF/6-18R) und zum Fragebogen für Jugendliche (YSR/11-18R)*. Göttingen: Hogrefe.

Dunn, O. J., and Clark, V. (1969). Correlation coefficients measured on the same individuals. *J. Am. Stat. Assoc.* 64, 366–377. doi: 10.1080/01621459.1969.10500981

Erhart, M., Döpfner, M., Ravens-Sieberer, U., and the Bella study group. (2008). Psychometric properties of two ADHD questionnaires: Comparing the Conners' scale and the FBB-HKS in the general population of German children and adolescents – Results of the BELLA study. *Eur. Child Adolesc. Psychiatry* 17(Suppl. 1), 106–115. doi: 10.1007/s00787-008-1012-1

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics*, 5th Edn. Thousand Oaks, CA: SAGE.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76, 378–382. doi: 10.1037/h0031619

Frick, P. J., Barry, C. T., and Kamphaus, R. W. (2010). *Clinical Assessment of Child and Adolescent Personality and Behavior*, 3rd Edn. Heidelberg: Springer. doi: 10.1007/978-1-4419-0641-0_11

Galanter, C. A., and Patel, V. L. (2005). Medical decision making: a selective review for child psychiatrists and psychologists. *J. Child Psychol. Psychiatry Allied Discip.* 46, 675–689. doi: 10.1111/j.1469-7610.2005.01452.x

Geissler, J., Jans, T., Banaschewski, T., Becker, K., Renner, T., Brandeis, D., et al. (2018). Individualised short-term therapy for adolescents impaired by attention-deficit/hyperactivity disorder despite previous routine care treatment (ESCAadol)-Study protocol of a randomised controlled trial within the consortium ESCAlife. *Trials* 19:254. doi: 10.1186/s13063-018-2635-2

Ghanizadeh, A., Mohammadi, M. R., and Yazdanshenas, A. (2006). Psychometric properties of the Farsi translation of the kiddie schedule for affective disorders and schizophrenia-present and lifetime version. *BMC Psychiatry* 6:10. doi: 10.1186/1471-244X-6-10

Gisev, N., Bell, J. S., and Chen, T. F. (2013). Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Res. Soc. Adm. Pharm.* 9, 330–338. doi: 10.1016/j.sapharm.2012.04.004

Görtz-Dorten, A., Ise, E., Hautmann, C., Walter, D., and Döpfner, M. (2014). Psychometric properties of a German parent rating scale for oppositional defiant and conduct disorder (FBB-SSV) in clinical and community samples. *Child Psychiatry Hum. Dev.* 45, 388–397. doi: 10.1007/s10578-013-0409-3

Görtz-Dorten, A., Thöne, A.-K., and Döpfner, M. (in press). *DISYPS-ILF: Interviewleitfäden zum Diagnostik-System für psychische Störungen für Kinder- und Jugendliche [Manual submitted for publication]*. Göttingen: Hogrefe.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor. Quant. Methods Psychol.* 8, 23–34. doi: 10.20982/tqmp.08.1.p023

Hoyer, J., and Knappe, S. (2012). Psychotherapie braucht strukturierte Diagnostik! *Psychother. Dialog* 13, 2–5. doi: 10.1055/s-0031-1298922

Ivanova, M. Y., Achenbach, T. M., Dumenci, L., Rescorla, L. A., Almqvist, F., Weintraub, S., et al. (2007). Testing the 8-syndrome structure of the child behavior checklist in 30 Societies Masha. *J. Clin. Child Adolesc. Psychol.* 36, 405–417. doi: 10.1080/15374410701444363

Ivanova, M. Y., Achenbach, T. M., Rescorla, L. A., Guo, J., Althoff, R. R., Kan, K. J., et al. (2019). Testing syndromes of psychopathology in parent and youth ratings across societies. *J. Clin. Child Adolesc. Psychol.* 48, 596–609. doi: 10.1080/15374416.2017.1405352

Jans, T., Weyers, P., Schneider, M., Hohage, A., Werner, M., Pauli, P., et al. (2009). The Kiddie-SADS allows a dimensional assessment of externalizing symptoms in ADHD children and adolescents. *Atten. Defic. Hyperact. Disord.* 1, 215–222. doi: 10.1007/s12402-009-0013-3

Kariuki, S. M., Newton, C. R. J. C., Abubakar, A., Bitta, M. A., Odhiambo, R., and Phillips Owen, J. (2018). Evaluation of psychometric properties and factorial structure of ADHD module of K-SADS-PL in children from rural Kenya. *J. Atten. Disord.* doi: 10.1177/1087054717753064 [Epub ahead of print].

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., et al. (1997). Schedule for affective disorders and schizophrenia for school-age children-present and lifetime version (K-SADS-PL): initial reliability and validity data. *J. Am. Acad. Child Adolesc. Psychiatry* 36, 980–988. doi: 10.1097/00004583-199707000-00021

Kim, Y. S., Cheon, K. A., Kim, B. N., Chang, S. A., Yoo, H. J., Kim, J. W., et al. (2004). The reliability and validity of Kiddie-schedule for affective disorders and schizophrenia-present and lifetime version-korean version (K-SADS-PL-K). *Yonsei Med. J.* 45, 81–89. doi: 10.3349/ymj.2004.45.1.81

Koo, T. K., and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* 15, 155–163. doi: 10.1016/j.jcm.2016.02.012

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., et al. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J. Clin. Epidemiol.* 64, 96–106. doi: 10.1016/j.jclinepi.2010.03.002

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.1109/ICDMA.2010.328

LeBreton, J. M., and Senter, J. L. (2008). Answers to 20 Questions about interrater reliability and interrater agreement. *Organ. Res. Methods* 11, 815–852. doi: 10.1177/1094428106296642

Leffler, J. M., Riebel, J., and Hughes, H. M. (2015). A review of child and adolescent diagnostic interviews for clinical practitioners. *Assessment* 22, 690–703. doi: 10.1177/1073191114561253

McGraw, K. O., and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychol. Methods* 1, 30–46. doi: 10.1037/1082-989X.1.1.30

Nishiyama, T., Sumi, S., Watanabe, H., Suzuki, F., Kuru, Y., Shiino, T., et al. (2020). The Kiddie schedule for affective disorders and schizophrenia present and lifetime version (K-SADS-PL) for DSM-5: a validation for neurodevelopmental disorders in Japanese outpatients. *Compr. Psychiatry* 96, 152148. doi: 10.1016/j.comppsych.2019.152148

Nordgaard, J., Sass, L. A., and Parnas, J. (2013). The psychiatric interview: validity, structure, and subjectivity. *Eur. Arch. Psychiatry Clin. Neurosci.* 263, 353–364. doi: 10.1007/s00406-012-0366-z

Nunnally, J. C. (1978). *Psychometric Theory*, 2nd Edn. New York, NY: McGraw-Hill.

Power, T. J., Costigan, T. E., Eiraldi, R. B., and Leff, S. S. (2004). Variations in anxiety and depression as a function of ADHD subtypes defined by DSM-IV: Do subtype differences exist or not? *J. Abnorm. Child Psychol.* 32, 27–37. doi: 10.1023/B:JACP.0000007578.30863.93

Ranshoff, D. F., and Feinstein, A. R. (1978). Problems of spectrum bias in evaluating the efficacy of diagnostic tests. *N. Engl. J. Med.* 299, 926–930. doi: 10.1056/nejm197810262991705

Rescorla, L., Achenbach, T., Ivanova, M. Y., Dumenci, L., Almqvist, F., Bilenberg, N., et al. (2007). Behavioral and emotional problems reported by parents of children ages 6 to 16 in 31 societies. *J. Emot. Behav. Disord.* 15, 130–142. doi: 10.1177/10634266070150030101

Rettew, D. C., Lynch, A. D., Achenbach, T. M., Dumenci, L., and Ivanova, M. Y. (2009). Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *Int. J. Methods Psychiatr. Res.* 18, 169–184. doi: 10.1002/mpr

Segal, D. L., and Williams, K. N. (2014). "Structured and semistructured interviews for differential diagnosis: fundamental Issues, applications, and features," in *Adult Psychopathology and Diagnosis*, 7th Edn, eds D. C. Beidel, B. C. Frueh, and M. Hersen (New York, NY: Wiley), 103–129.

Shaffer, D., Fisher, P., Lucas, C. P., Dulcan, M. K., and Schwab-Stone, M. E. (2000). NIMH Diagnostic Interview Schedule for Children Version IV (NIMH DISC- IV): description, differences from previous versions, and reliability of some common diagnoses. *J. Am. Acad. Child Adolesc. Psychiatry* 39, 28–38. doi: 10.1097/00004583-200001000-00014

Sheehan, D. V., Sheehan, K. H., Shytle, R. D., Janavs, J., Bannon, Y., Rogers, J. E., et al. (2010). Reliability and validity of the mini international neuropsychiatric interview for children and adolescents (MINI-KID). *J. Clin. Psychiatry* 71, 313–326. doi: 10.4088/JCP.09m05305whi

Shrout, P. E., and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* 86, 420–428. doi: 10.1037/0033-2909.86.2.420

Spitzer, R. L., and Wakefield, J. C. (1999). DSM-IV diagnostic criterion for clinical significance: Does it help solve the false positives problem? *Am. J. Psychiatry* 156, 1856–1864. doi: 10.1176/ajp.156.12.1856

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychol. Bull.* 87, 245–251. doi: 10.1037//0033-2909.87.2.245

Ulloa, R. E., Ortiz, S., Higuera, F., Nogales, I., Fresán, A., Apiquian, R., et al. (2006). Interrater reliability of the Spanish version of schedule for affective disorders and schizophrenia for school-age children-present and lifetime version (K-SADS-PL). *Actas Esp. Psiquiatr.* 34, 36–40.

Weller, E. B., Weller, R. A., Rooney, M. T., and Fristad, M. A. (1999a). *Children's Interview for Psychiatric Syndromes –Parent version (P-ChIPS)*. Washington, DC: American Psychiatric Press, Inc.

Weller, E. B., Weller, R. A., Rooney, M. T., and Fristad, M. A. (1999b). *Children's Interview for Psychiatric Syndromes (ChIPS)*. Washington, DC: American Psychiatric Press, Inc.

Willcutt, E. G., Nigg, J. T., Pennington, B. F., Solanto, M. V., Rohde, L. A., Tannock, R., et al. (2012). Validity of DSM-IV attention deficit/hyperactivity disorder symptom dimensions and subtypes. *J. Abnorm. Psychol.* 121, 991–1010. doi: 10.1037/a0027347

Wirtz, M., and Caspar, F. (2002). Beurteilerübereinstimmung und Beurteilerreliabilität. [Inter-rater agreement and inter-rater reliability]. Göttingen: Hogrefe.

Zou, G. Y. (2012). Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Stat. Med.* 31, 3972–3981. doi: 10.1002/sim.5466

Conflict of Interest: A-KTh, AG-D, and MD were involved in the development of the DISYPS-ILF and will receive royalties after publication of this instrument from the publisher Hogrefe. AG-D and MD are AKiP supervisors and lecturers and received income as heads of the School for Child and Adolescent Behavior Therapy at the University of Cologne and royalties from treatment manuals, books and psychological tests published by Guilford, Hogrefe, Enke, Beltz, and Huber. MD received consulting income and research support from Lilly, Medice, Shire, Janssen Cilag, Novartis, and Vifor. AG-D is head of the AKiP Research and Evaluation department. TB served in an advisory or consultancy role for Lundbeck, Medice, Neurim Pharmaceuticals, Oberberg GmbH, Shire, and Infectopharm; and received conference support or speaker's fees from Lilly, Medice, and Shire; and received royalties from Hogrefe, Kohlhammer, CIP Medien, and Oxford University Press. DB served as an unpaid scientific advisor for an EU-funded neurofeedback trial unrelated to the present work. KB has been involved in research/clinical trials with Eli Lilly ($\leq$2011) and Shire (2010), was on the Advisory Board of Eli Lilly/Germany ($\leq$2014), a member of the Scientific Committee of Shire ($\leq$2012) and was paid for public speaking by Eli Lilly ($<$2011) and Shire. These activities do not bias the objectivity of this manuscript (in her opinion), but are mentioned for the sake of completeness. MHo served in an advisory role for Shire and Medice and received conference attendance support or was paid for public speaking by Medice, Shire and Neuroconn. He receives research support from the German Research Foundation and the German Ministry of Education and Research. He receives royalties as editor in chief of the German Journal for Child and Adolescent Psychiatry and for text books from Hogrefe. MHu served as a member of the advisory boards of Eli Lilly and Co., Engelhardt Arzneimittel, Janssen-Cilag, Medice, Novartis, Shire, and Steiner Arzneimittel within the past 5 years; served as a consultant to Engelhardt Arzneimittel, Medice, and Steiner Arzneimittel; received honoraria from Eli Lilly and Co., Engelhardt Arzneimittel, Janssen-Cilag, Medice, Novartis, and Shire; and received unrestricted grants for investigator-initiated trials from Eli Lilly and Co., Medice, Engelhardt Arzneimittel, and Steiner Arzneimittel. LP served in an advisory or consultancy role for Shire, Roche and Infectopharm; and received speaker's fees from Shire and royalties from Hogrefe, Kohlhammer, and Schattauer. HU served in an advisory or consultancy role for Medice; and received speaker's fees from Shire and Medice.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.