DOCUMENT SUMMARY

This research article evaluates the reliability and validity of the **ILF-EXTERNAL**, a new DSM-5-based, semi-structured clinical parent interview for diagnosing externalizing disorders in children. The study, conducted on a clinical sample of school-age children with **ADHD** symptoms, found that the interview has sound psychometric properties. The results demonstrate very good to excellent interrater reliability, satisfactory internal consistency, and strong convergent and divergent validity, establishing the **ILF-EXTERNAL** as a promising tool for the dimensional and categorical assessment of disorders like **ADHD**, **ODD**, and **CD**.

FILENAME

thone_2020_clinical_research_article_externalizing_disorders_assessment_validation

METADATA

Category: CLINICAL

Type: research_article

Relevance: Core

Update Frequency: Static

Tags: #externalizing-disorders #adhd #odd #cd #dimensional-assessment #ilf-external #dsm-5 #clinical-interview #psychometrics #reliability #validity

Related Docs: N/A

Supersedes: N/A

FORMATTED CONTENT

Thöne_2020_Toward_a_Dimensional_Assessment_of_Externalizing_Disorders_in_Children_Reliability_and_Validity_of_a_Semi-Structured_Parent_Interview

Ann-Kathrin Thöne, Anja Görtz-Dorten, Paula Altenberger, Christina Dose, Nina Geldermann, Christopher Hautmann, Lea Teresa Jendreizik, Anne-Katrin Treier, Elena von Wirth, Tobias Banaschewski, Daniel Brandeis, Sabina Millenet, Sarah Hohmann, Katja Becker, Johanna Ketter, Johannes Hebebrand, Jasmin Wenning, Martin Holtmann, Tanja Legenbauer, Michael Huss, Marcel Romanos, Thomas Jans, Julia Geissler, Luise Poustka, Henrik Uebel-von Sandersleben, Tobias Renner, Ute Dürrwächter and Manfred Döpfner

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 ≤ r_it ≤ 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.

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 (p) 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 $\rho=0.80$ is no less than $\rho_0=0.65$ 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).

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 ≤ 0.40; fair = 0.41-0.59; good = 0.60-0.74; and excellent ≥ 0.75. However, other authors proposed more stringent guidelines: poor ≤ 0.50; moderate = 0.51–0.75; good =0.76-0.90; and excellent ≥ 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 ≤ 0.20; fair=0.21- 0.40; moderate =0.41-0.60; substantial =0.61-0.80; almost perfect agreement >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 ≤ r_it ≤ 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 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**: k=0.38), through moderate (**DMDD**: k=0.55), substantial (**ADHD combined type**: k=0.71; **ADHD inattentive type**: κ = 0.71; **any ADHD**: k=0.74) to almost perfect agreement (**ODD**: k=0.82; **conduct disorder**: k=0.94; with its specifier **limited prosocial emotions**: k=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).

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 (α=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 (α=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 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 (k=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 (k=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 ≤ κ ≤ 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** (κ=0.90) and **ADHD inattentive type** (k=0.93) and substantial agreement for **ADHD hyperactive-impulsive type** (k=0.65). Interestingly, high diagnostic agreement on diagnosing **ADHD combined type** (k=0.86) and **ADHD inattentive type** (k=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 ≤ κ ≤ 0.80; Ghanizadeh et al., 2006; de la Peña et al., 2018) **DMDD** (κ=0.53; de la Peña et al., 2018), and **CD** (0.78 ≤ κ ≤ 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 (k=0.29; de la Peña et al., 2018) we found very high diagnostic agreement on this specifier (k=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.

Tables

Table 1: Sample characteristics.

Characteristic	Total Sample (N=474)	IRR Subsample (N=45)
Age, M (SD)	8.9 (1.5)	8.8 (1.5)

Sex, n (%)		
Female	92 (19.4)	7 (15.6)
Male	382 (80.6)	38 (84.4)
ADHD Diagnosis, n (%)		
Combined type	285 (60.1)	28 (62.2)
Inattentive type	125 (26.4)	14 (31.1)
Hyperactive-impulsive type	32 (6.8)	2 (4.4)
No ADHD diagnosis	32 (6.8)	1 (2.2)
Comorbid Diagnoses, n (%)		
ODD	146 (30.8)	13 (28.9)
CD	32 (6.8)	6 (13.3)
DMDD	12 (2.5)	2 (4.4)

Specific learning disorder	137 (28.9)	12 (26.7)
Tic disorder	49 (10.3)	4 (8.9)
Anxiety disorder	98 (20.7)	10 (22.2)
Depressive disorder	22 (4.6)	1 (2.2)

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

Scale	**k**	**M (SD) Total Sample**	**M (SD) Screening Negatives**	**α**	**Range of r_it**
ADHD					
Inattention	9	1.95 (0.48)	1.13 (0.37)	0.74	0.28-0.58
Hyperactivity-Impulsivity	9	1.55 (0.70)	0.69 (0.48)	0.81	0.49-0.66
ADHD Symptoms	18	1.75 (0.49)	0.91 (0.32)	0.83	0.35-0.63
ADHD Functional Impairment	5	1.83 (0.57)	1.04 (0.50)	0.62	0.28-0.49
ODD/CD					
ODD Symptoms	8	0.80 (0.61)	0.38 (0.32)	0.82	0.51-0.66

CD Symptoms - short version	5	0.40 (0.43)	0.11 (0.19)	0.60	0.21-0.49
ODD/CD Symptoms - short version	13	0.65 (0.47)	0.28 (0.24)	0.82	0.40-0.64
Disruptive Mood Dysregulation	5	0.58 (0.60)	0.24 (0.30)	0.76	0.48-0.63
Limited Prosocial Emotions	11	0.44 (0.42)	0.17 (0.23)	0.79	0.31-0.65
ODD/CD Functional Impairment	5	0.88 (0.75)	0.35 (0.41)	0.86	0.60-0.71

Table 3: Interrater reliability of the ILF-EXTERNAL scales.

Scale	**ICC(1,1)**	**95% CI**	**ICC(1,3)**	**95% CI**	**Pairwise % Agreement**
ADHD					
Inattention	0.83	[0.72, 0.90]	0.94	[0.89, 0.96]	83.0
Hyperactivity-Impulsivity	0.95	[0.92, 0.97]	0.98	[0.97, 0.99]	91.9
ADHD Symptoms	0.93	[0.88, 0.96]	0.97	[0.96, 0.99]	89.6

ADHD Functional Impairment	0.89	[0.82, 0.94]	0.96	[0.93, 0.98]	84.4
ODD/CD					
ODD Symptoms	0.94	[0.90, 0.97]	0.98	[0.96, 0.99]	91.1
CD Symptoms - short version	0.95	[0.91, 0.97]	0.98	[0.97, 0.99]	94.1
ODD/CD Symptoms - short version	0.94	[0.90, 0.97]	0.98	[0.96, 0.99]	92.6
Disruptive Mood Dysregulation	0.88	[0.80, 0.93]	0.96	[0.92, 0.98]	85.9
Limited Prosocial Emotions	0.93	[0.88, 0.96]	0.97	[0.96, 0.99]	89.6
ODD/CD Functional Impairment	0.92	[0.86, 0.95]	0.97	[0.95, 0.99]	88.9

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

Diagnosis	n (%)	Fleiss' Kappa	95% CI	Pairwise % Agreement
Any ADHD	44	0.74	[0.00,	100.0

	(97.8)		1.00]	
ADHD combined type	28 (62.2)	0.71	[0.52, 0.90]	82.2
ADHD inattentive type	14 (31.1)	0.71	[0.51, 0.92]	84.4
ADHD hyperactive-impulsive type	2 (4.4)	0.38	[-0.17, 0.93]	95.6
ODD	13 (28.9)	0.82	[0.66, 0.99]	91.1
CD	6 (13.3)	0.94	[0.82, 1.00]	97.8
Limited prosocial emotions	8 (17.8)	0.82	[0.63, 1.00]	93.3
DMDD	2 (4.4)	0.55	[0.05, 1.00]	95.6

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

Scale	**ILF-EXTERNAL M (SD)**	**FBB M (SD)**	**r**	**t(df)**
ADHD				
Inattention	1.95 (0.48)	2.05	0.69***	-3.14(467)

		(0.57)		
Hyperactivity-Impulsivity	1.55 (0.70)	1.51 (0.74)	0.78***	1.05(466)
ADHD Symptoms	1.75 (0.49)	1.78 (0.56)	0.78***	-1.96(465)
ADHD Functional Impairment	1.83 (0.57)	2.00 (0.69)	0.63***	-5.19(461)
ODD/CD				
ODD Symptoms	0.80 (0.61)	0.96 (0.64)	0.73***	-5.17(464)
CD Symptoms - short version	0.40 (0.43)	0.42 (0.47)	0.57***	-1.13(464)
ODD/CD Symptoms - short version	0.65 (0.47)	0.76 (0.50)	0.74***	-4.68(463)
Disruptive Mood Dysregulation	0.58 (0.60)	0.73 (0.67)	0.70***	-4.73(464)
Limited Prosocial Emotions	0.44 (0.42)	0.56 (0.48)	0.63***	-5.16(464)
ODD/CD Functional Impairment	0.88 (0.75)	1.10 (0.79)	0.67***	-5.87(462)

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

ILF-EXTERNAL Scale	Anxious/ Depressed	Withdrawn/ Depressed	Somatic Complaints	Social Problems	Thought Problems	Attention Problems	Rul Bre Beh
Inattention	0.17***	0.18***	0.12**	0.28***	0.16***	0.39***	0.23
Hyperactivity-Impulsivity	0.12**	0.10*	0.04	0.29***	0.13**	0.38***	0.46
ADHD Symptoms	0.17***	0.16***	0.09*	0.33***	0.17***	0.44***	0.40
ODD Symptoms	0.23***	0.11*	0.12**	0.32***	0.18***	0.38***	0.54
CD Symptoms - short	0.08	0.03	0.10*	0.12**	0.12**	0.20***	0.48
ODD/CD Symptoms - short	0.20***	0.09*	0.13**	0.28***	0.18***	0.36***	0.61

References

- **Achenbach_1991_Manual_for_the_Child_Behavior_Checklist/4-18_and_1991_Profile**
- **Achenbach_&*Rescorla_2001_Manual_for_the_ASEBA_School-Age_Forms*&_Profiles**
- **American_Psychiatric_Association_2013_Diagnostic_and_Statistical_Manual_of_Mental_Disorders_5th_Edn**
- **Angold_&*Costello_1995_A_test-retest_reliability_study_of_child-reported_psychiatric_symptoms_and_diagnoses_using_the_child_and_adolescent_psychiatric_assessment*(CAPA-C)**

- **Angold_&*Costello_2000_The_child_and_adolescent_psychiatric_assessment*(CAPA)**
- **Arbeitsgruppe_Deutsche_Child_Behavior_Checklist_1998_Elternfragebogen_über_das_Verhalten_von_Kindern_und_Jugendlichen**
- **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"**
- **Becker_Banaschewski_Brandeis_et_al_2020_Individualised_stepwise_adaptive_treatment_for_3-6-year-old_preschool_children_impaired_by_attention-deficit_/*hyperactivity_disorder*(ESCApreschool)**
- **Bennett_2001_How_can_I_deal_with_missing_data_in_my_study?**
- **Birmaher_Ehmann_Axelson_et_al_2009_Schedule_for_affective_disorders_and_schizophrenia_for_school-age_children_(K-SADS-PL)_for_the_assessment_of_preschool_children**
- **Bland_&_Altman_1986_Statistical_methods_for_assessing_agreement_between_two_methods_of_clinical_measurement**
- **Boyle_Duncan_Georgiades_et_al_2017_Classifying_child_and_adolescent_psychiatric_disorder_by_problem_checklists_and_standardized_interviews**
- **Brasil_&_Bordin_2010_Convergent_validity_of_K-SADS-PL_by_comparison_with_CBCL_in_a_Portuguese_speaking_outpatient_population**
- **Chen_Shen_&_Gau_2017_The_Mandarin_version_of_the_Kiddie-Schedule_for_Affective_Disorders_and_Schizophrenia-Epidemiological_version_for_DSM-5**
- **Cicchetti_1994_Guidlines_criteria_and_rules_of_thumb_for_evalauting_normed_and_standardized_assessment_instruments_in_psychology**
- **Coghill_&_Sonuga-Barke_2012_Annual_research_review_categories_versus_dimensions_in_the_classification_and_conceptualisation_of_child_and_adolescent_mental_disorders**
- **de_la_Peña_Villavicencio_Palacio_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**
- **Diedenhofen_&_Musch_2015_Cocor_a_comprehensive_solution_for_the_statistical_comparison_of_correlations**
- **Döpfner_Breuer_Wille_et_al_2008_How_often_do_children_meet_ICD-10/DSM-IV_criteria_of_attention_deficit-/hyperactivity_disorder_and_hyperkinetic_disorder?**
- **Döpfner_&*Görtz-Dorten_2017_Diagnostik-System_für_Psychische_Störungen_nach_ICD-10_und_DSM-5_für_Kinder_und_Jugendliche*–_III**
- **Döpfner_Hautmann_Dose_et_al_2017_ESCAschool_study_trial_protocol_of_an_adaptive_treatment_approach_for_school-age_children_with_ADHD**

- **Döpfner_&_Petermann_2012_Diagnostik_Psychischer_Störungen_im_Kindes-und_Jugendalter**
- **Döpfner_Plück_Kinnen_&_Arbeitsgruppe_Deutsche_Child_Behavior_Checklist_2014_CBCL_Handbuch-Schulalter**
- **Dunn_&_Clark_1969_Correlation_coefficients_measured_on_the_same_individuals**
- **Erhart_Döpfner_Ravens-Sieberer_&_the_Bella_study_group_2008_Psychometric_properties_of_two_ADHD_questionnaires**
- **Field_2018_Discovering_Statistics_Using_IBM_SPSS_Statistics_5th_Edn**
- **Fleiss_1971_Measuring_nominal_scale_agreement_among_many_raters**
- **Frick_Barry_&_Kamphaus_2010_Clinical_Assessment_of_Child_and_Adolescent_Personality_and_Behavior_3rd_Edn**
- **Galanter_&_Patel_2005_Medical_decision_making_a_selective_review_for_child_psychiatrists_and_psychologists**
- **Geissler_Jans_Banaschewski_et_al_2018_Individualised_short-term_therapy_for_adolescents_impaired_by_attention-deficit/hyperactivity_disorder_despite_previous_routine_care_treatment_(ESCAadol)**
- **Ghanizadeh_Mohammadi_&_Yazdanshenas_2006_Psychometric_properties_of_the_Farsi_translation_of_the_kiddie_schedule_for_affective_disorders_and_schizophrenia-present_and_lifetime_version**
- **Gisev_Bell_&_Chen_2013_Interrater_agreement_and_interrater_reliability_key_concepts_approaches_and_applications**
- **Görtz-Dorten_Ise_Hautmann_et_al_2014_Psychometric_properties_of_a_German_parent_rating_scale_for_oppositional_defiant_and_conduct_disorder_(FBB-SSV)**
- **Görtz-Dorten_Thöne_&_Döpfner_in_press_DISYPS-ILF_Interviewleitfäden_zum_Diagnostik-System_für_psychische_Störungen_für_Kinder-_und_Jugendliche**
- **Hallgren_2012_Computing_inter-rater_reliability_for_observational_data_an_overview_and_tutorial**
- **Hoyer_&_Knappe_2012_Psychotherapie_braucht_strukturierte_Diagnostik!**
- **Ivanova_Achenbach_Dumenci_et_al_2007_Testing_the_8-syndrome_structure_of_the_child_behavior_checklist_in_30_Societies**
- **Ivanova_Achenbach_Rescorla_et_al_2019_Testing_syndromes_of_psychopathology_in_parent_and_youth_ratings_across_societies**
- **Jans_Weyers_Schneider_et_al_2009_The_Kiddie-SADS_allows_a_dimensional_assessment_of_externalizing_symptoms_in_ADHD_children_and_adolescents**
- **Kariuki_Newton_Abubakar_et_al_2018_Evaluation_of_psychometric_properties_and_factorial_structure_of_ADHD_module_of_K-SADS-PL_in_children_from_rural_Kenya**

- Kaufman_Birmaher_Brent_et_al_1997_Schedule_for_affective_disorders_and_schizophrenia_for_school-age_children-present_and_lifetime_version_(K-SADS-PL)
- Kim_Cheon_Kim_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)
- Koo_&_Li_2016_A_guideline_of_selecting_and_reporting_intraclass_correlation_coefficients_for_reliability_research
- Kottner_Audigé_Brorson_et_al_2011_Guidelines_for_reporting_reliability_and_agreement_studies_(GRRAS)_were_proposed
- Landis_&_Koch_1977_The_measurement_of_observer_agreement_for_categorical_data
- LeBreton_&_Senter_2008_Answers_to_20_Questions_about_interrater_reliability_and_interrater_agreement
- Leffler_Riebel_&_Hughes_2015_A_review_of_child_and_adolescent_diagnostic_interviews_for_clinical_practitioners
- McGraw_&_Wong_1996_Forming_inferences_about_some_intraclass_correlation_coefficients
- Nishiyama_Sumi_Watanabe_et_al_2020_The_Kiddie_schedule_for_affective_disorders_and_schizophrenia_present_and_lifetime_version_(K-SADS-PL)_for_DSM-5
- Nordgaard_Sass_&_Parnas_2013_The_psychiatric_interview_validity_structure_and_subjectivity
- Nunnally_1978_Psychometric_Theory_2nd_Edn
- Power_Costigan_Eiraldi_&_Leff_2004_Variations_in_anxiety_and_depression_as_a_function_of_ADHD_subtypes_defined_by_DSM-IV
- Ranshoff_&_Feinstein_1978_Problems_of_spectrum_bias_in_evaluating_the_efficacy_of_diagnostic_tests
- Rescorla_Achenbach_Ivanova_et_al_2007_Behavioral_and_emotional_problems_reported_by_parents_of_children_ages_6_to_16_in_31_societies
- Rettew_Lynch_Achenbach_et_al_2009_Meta-analyses_of_agreement_between_diagnoses_made_from_clinical_evaluations_and_standardized_diagnostic_interviews
- Segal_&_Williams_2014_Structured_and_semistructured_interviews_for_differential_diagnosis
- Shaffer_Fisher_Lucas_et_al_2000_NIMH_Diagnostic_Interview_Schedule_for_Children_Version_IV_(NIMH_DISC-IV)
- Sheehan_Sheehan_Shytle_et_al_2010_Reliability_and_validity_of_the_mini_international_neuropsychiatric_interview_for_children_and_adolescents_(MINI-KID)
- Shrout_&_Fleiss_1979_Intraclass_correlations_uses_in_assessing_rater_reliability
- Spitzer_&_Wakefield_1999_DSM-IV_diagnostic_criterion_for_clinical_significance
- Steiger_1980_Tests_for_comparing_elements_of_a_correlation_matrix

- **Ulloa_Ortiz_Higuera_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)**
- **Weller_Weller_Rooney_&*Fristad_1999a_Children's_Interview_for_Psychiatric_Syndromes*-Parent_version_(P-ChIPS)**
- **Weller_Weller_Rooney_&*Fristad_1999b_Children's_Interview_for_Psychiatric_Syndromes*(ChIPS)**
- **Willcutt_Nigg_Pennington_et_al_2012_Validity_of_DSM-IV_attention_deficit/hyperactivity_disorder_symptom_dimensions_and_subtypes**
- **Wirtz_&_Caspar_2002_Beurteilerübereinstimmung_und_Beurteilerreliabilität**
- **Zou_2012_Sample_size_formulas_for_estimating_intraclass_correlation_coefficients_with_precision_and_assurance**