



Empirically based dimensions of externalizing symptoms in children and adolescents: a multitrait-multisource approach

Ann-Kathrin Thöne¹ · Michaela Junghänel¹ · Anja Görtz-Dorten^{1,2} · Dieter Breuer¹ · Teresa del Giudice¹ · Charlotte Hanisch³ · Thomas Hennemann³ · Manfred Döpfner^{1,2,3}

Accepted: 10 June 2022 / Published online: 28 June 2022

© The Author(s) 2022

Abstract

The present study sought to refine knowledge about the structure underlying externalizing dimensions. From a “top-down” ICD/DSM-based perspective, externalizing symptoms can be categorized into attention-deficit/hyperactivity disorder (ADHD), oppositional defiant disorder (ODD), and conduct disorder (CD). From a “bottom-up” developmental theory-based perspective, disruptive behaviors can be meaningfully described as aggressive (AGG) and rule-breaking (RB) behaviors. We analyzed five large data sets comprising externalizing symptom ratings obtained with a screening instrument using different sources (parents, teachers, self-ratings) from different samples. Using confirmatory factor analyses, we evaluated several factor models (unidimensional; correlated factors; bifactor (S-1) models) derived from an ICD/DSM- and theory-based perspective. Our optimally fitting models were assessed for measurement invariance across all sources, sample settings, and sex. Following several model-based criteria (model fit indices; factor loadings; omega statistics; model parsimony), we discarded our models stepwise and concluded that both the ICD/DSM-based model with three correlated factors (ADHD, ODD, CD) and the developmental theory-based model with three correlated factors (ADHD, AGG, RB) displayed a statistically sound factor structure and allowed for straightforward interpretability. Furthermore, these two models demonstrated metric invariance across all five samples and across sample settings (community, clinical), as well as scalar invariance across sources and sex. While the dimensions AGG and RB may depict a more empirically coherent view than the categorical perspective of ODD and CD, at this point we cannot clearly determine whether one perspective really outperforms the other. Implications for model selection according to our model-based criteria and clinical research are discussed.

Keywords Bifactor model · Bifactor S-1 · ADHD · ODD · CD · Measurement invariance

Categorical and dimensional perspectives on mental disorders

Accurately classifying mental disorders remains a challenge for studying psychological symptoms and selecting appropriate treatment. Available classification systems, including the 5th edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5; American Psychiatric Association, 2013), the 10th version of the International Classification of Diseases (ICD-10; World Health Organization 1993), or the upcoming ICD-11 (World Health Organization, 2019), were developed following a “top-down” approach based on clinician consensus, and mental disorders were conceptualized as categorical concepts with a disorder being defined as either absent or present (reviewed by Achenbach 2020). In comparison, a “bottom-up” approach prioritizes empirical

Ann-Kathrin Thöne and Michaela Junghänel contributed equally to this work.

✉ Ann-Kathrin Thöne
ann-kathrin.thoene@uk-koeln.de

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 and University Hospital Cologne, University of Cologne, Cologne, Germany

³ Department of Special Education, Faculty of Human Sciences, University of Cologne, Cologne, Germany

data from which conceptualizations of mental abnormalities are derived (Achenbach, 2020). This bottom-up approach is consistent with a dimensional perspective and has spawned diagnostic instruments which are widely used in child psychopathology (e.g. the Achenbach System of Empirically Based Assessment [ASEBA]; Achenbach 1991; Achenbach & Rescorla, 2001). In addition, ongoing research efforts such as the Hierarchical Taxonomy of Psychopathology (HiTOP; Kotov et al., 2017) attempt to identify a more accurate and potentially more parsimonious representation of the underlying structure of psychopathology. This system, which is still a work in progress, specifies six hierarchical layers, ranging from super spectra on the highest level to symptoms on the lowest level. *Super spectra* are higher-order dimensions, assumed to influence all spectra on the layer below. The HiTOP specifies six *spectra* (e.g. antagonistic externalizing), an array of *subfactors* (e.g. antisocial behavior), *syndromes and disorders* (e.g. conduct disorder [CD]), which are used synonymously with DSM-5 diagnoses at this point to facilitate communication, *components* (e.g. maladaptive traits) and *symptoms* (e.g. physical aggression) on the lowest level. In recent years, attention has focused on the uppermost level of the hierarchy, leading to the search for a general factor of psychopathology, the so-called *p-factor* (Caspi et al., 2014; Lahey et al., 2012). Although the precise nature of the *p-factor* is not yet understood, it may reflect a broad liability to psychopathology (Caspi & Moffitt, 2018). Overall, existing categorical diagnostic systems are challenged by dimensional models of psychopathology, which may provide a more accurate and potentially more parsimonious representation of psychopathology.

Latent factor analysis as a method to examine the structure of psychopathology

Statistical methods such as latent factor analysis can be employed to identify a common underlying factor of mental disorders. In particular, hierarchical models and bifactor models, which have become very popular over the last decade (Reise, 2012), can help to identify common factors. Hierarchical models can have several levels: In psychological research, they are often limited to two layers, in which for each domain, one first-order factor is modeled, which are all influenced by a second-order factor (Eid et al., 2017). In a bifactor model, the general factor (*g-factor*) is modeled as another first-order factor, influencing all observed variables. Additionally, specific factors (*s-factors*) for the individual domains are modeled, which are assumed to influence the observed variables within their domain. As the bifactor model approach is based on the assumption that the *g-factor* causes all correlations between the *s-factors*, the

correlations between *g-* and *s-factors*, and between all the *s-factors*, are constrained to zero (Eid et al., 2017), although this latter requirement is often ignored (e.g. Caspi et al., 2014).

Bifactor models, however, come with a few serious problems, calling their application in psychological research into question. One issue concerns the accurate interpretation of the results of these models. As bifactor models are less restrictive than correlated factor models, they generally result in better global model fit indices, although this does not necessarily indicate that these models really do fit the data better, as bifactor models are prone to overfitting (Bonifay et al., 2017). While an array of additional bifactor-specific indices exists (Rodriguez et al., 2016), they are frequently not calculated and reported, and the choice of the best model is based solely on model fit. This often leads to a bifactor model being “undeservingly” chosen as the best model (Watts et al., 2019).

Some of the major issues inherent in traditional bifactor models cannot be identified when solely examining model fit, but rather require the additional assessment of bifactor-specific indices. Such issues pertain to frequently observed anomalous factor loadings (Burns et al., 2020a; Eid et al., 2017; Junghänel et al., 2020; Rodenacker et al., 2018; Thöne et al., 2021) and vanishing *s-factors* (Burns et al., 2020a). Anomalous results are any results that are not in line with the general structure of the model, and include negative or non-significant factor variances and/or factor loadings > 1 (Eid et al., 2017). Vanishing *s-factors* are defined by a large number of non-significant/negative factor loadings on their respective factor, suggesting that the factor in question might not exist and leaving the defined indicators for that factor to only measure the *g-factor* (Heinrich et al., 2021). This is problematic as it changes the meaning of the *g-factor* and the *s-factors*. The *g-factor* is no longer the general factor that was assumed to represent all domains equally well, and is instead now mainly defined by indicators from the vanished *s-factor*. As a further consequence of this, the meaning of the *g-factor* is study- and sample-specific and cannot be compared across studies (Burns et al., 2020a), as it does not represent a general psychopathology factor as intended. It has been suggested (Eid et al., 2017; Heinrich et al., 2021) that the mistaken assumption of interchangeable domains could be the reason for the frequently found anomalous results and vanishing *s-factors*. Although the assumption of a *g-factor* equally influencing all psychopathological symptom complexes is very parsimonious, easy to understand and indeed tempting, previous research has shown that it is at best questionable whether the *g-factor* in traditional bifactor models truly represents general psychopathology (Levin-Aspenson et al., 2021; Watts et al., 2019).

Apart from bifactor models, there are also alternative factor models to analyze multi-faceted constructs. A frequently used model is the first-order correlated factor model (CFM), which specifies multiple distinct facets of a construct that do not overlap (Eid, 2020). When high correlations between the facets emerge, suggesting commonality, the application of a more theory-driven version of a bifactor model – a bifactor S-1 model – could be the appropriate choice. In a bifactor S-1 model, one s-factor out of the specific existing domains is left unmodeled. The s-factors are statistically contrasted against this remaining s-factor, which is now defined as the general reference domain (Burns et al., 2020a). S-factors in a bifactor S-1 model represent the part of a domain that cannot be explained by the reference domain (Burns et al., 2020a). In this case, domains can be structurally different, correlations between s-factors can be meaningfully interpreted, and anomalous results disappear as a result (Heinrich et al., 2021). With this approach, assumed commonality between the s-factors is accounted for by modeling a general reference domain, the association between all factors can be assessed simultaneously, and interpretation can occur in a straightforward manner. The *a priori* definition of the general reference facet, which is chosen based on theoretical assumptions or the specific research question (Eid, 2020), allows for a comparison between studies, irrespective of which s-factors are included. Eid (2020) summarizes in which situation a CFM should be applied and when a bifactor S-1 model should be specified, as the two yield differential information and both can help to disentangle the association between different domains of psychopathology.

Latent factor structure of the externalizing spectrum

As the identification of a general psychopathology factor seems to be challenging at best, in this study, we focus on a more narrowly defined area of psychopathology – the externalizing spectrum – by applying the aforementioned factor analytic models to assess the structure of symptoms within that spectrum.

From an ICD/DSM-based perspective, externalizing symptoms can be categorized into the disorders ADHD, ODD, and CD, which have been shown to be strongly related to each other (Willcutt et al., 2012). Regarding ADHD and ODD, Willcutt et al., (2012) found that around 50% of children diagnosed with ADHD also met criteria for an ODD diagnosis. A study by Burns et al., (2020b), which applied a bifactor S-1 model in a community sample of Spanish children, supported the strong association between the ADHD domains and ODD, while at the same time emphasizing the importance of distinct domains. Specifying HI as the general

reference facet, the authors found the IN domain of ADHD and ODD to remain a stable component, which is strongly associated with, but still distinct from, ADHD HI. Junghänel et al., (2020) and Thöne et al., (2021) reported similar findings regarding the association of these domains and the stability of the ADHD IN and the ODD factor in clinical samples of German children, thus strengthening the results of Burns et al., (2020b) with respect to the latent factor structure of ADHD and ODD. Regarding ADHD and CD, around 20% of children diagnosed with ADHD also met criteria for a CD diagnosis (Willcutt et al., 2012). A study by Beauchaine et al., (2010) supported the association between ADHD and CD, with the authors reporting that the vast majority of adolescent boys with early-onset CD also meet the criteria for ADHD. While a highly heritable externalizing liability factor, which is expressed as temperamental trait impulsivity, is assumed to represent the predisposing vulnerability to both of these disorders, environmental factors are also assumed to play a major role in the emergence of early-onset CD (Beauchaine et al., 2017; Beauchaine & McNulty, 2013). Symptoms of ODD can be regarded as middlemen in the development from child impulsivity to CD problems, a pathway which is negatively reinforced by high-risk environments (Beauchaine, 2015).

From a developmental theory-based perspective, disruptive behaviors can be meaningfully described as aggressive (AGG; e.g. having tantrums, arguing, threatening) and rule-breaking (RB; e.g. lying, stealing, skipping school) dimensions, instead of categorizing these behaviors into an ODD or CD diagnosis (Burt, 2012; Burt et al., 2015). In line with this, these dimensions have been shown to differ in terms of developmental course and etiology (Harden et al., 2015). Moreover, the factor analytic literature supports the notion that RB and AGG constitute separable, though positively correlated ($r = .55$), dimensions (Burt, 2012). Critically, however, there is a large variability around this mean correlation (range: $r = .28 - .73$), which may be attributable to informant discrepancies (Burt, 2012). Finally, Achenbach's internationally widely recognized ASEBA instruments also assess disruptive behaviors using the empirically derived AGG and RB syndrome scales (Achenbach & Rescorla, 2001).

Multi-informant assessment & measurement invariance of the externalizing spectrum

In clinical child assessment, integrating multiple informants' reports (e.g. children, parents, teachers) is considered a key component of best practices in evidence-based assessment, since it is unlikely that one single informant

is sufficiently privy to a child's situation-specific behavior, such as at home or school (Achenbach, 2020; De Los Reyes et al., 2013; Los Reyes et al., 2015; Dirks et al., 2012). In fact, an early meta-analysis demonstrated moderate correspondence ($\bar{r} = 0.60$) between similar informants in the same context (e.g. pairs of parents), but low correspondence ($\bar{r} = 0.28$) between different types of informants (e.g. parents vs. teachers), and the lowest correspondence ($\bar{r} = 0.22$) between self and other informant reports (Achenbach et al., 1987). Subsequent meta-analyses have demonstrated similar cross-informant discrepancies in child psychopathology (De Los Reyes et al., 2015; Los Reyes et al., 2019). While these informant discrepancies had been theorized to reflect some kind of invalidity or rater bias (De Los Reyes, 2011), there is growing recognition that such discrepancies may rather reflect how children's behavior varies meaningfully across contexts (De Los Reyes et al., 2013; Dirks et al., 2012).

Regarding measurement invariance of the externalizing spectrum, some studies demonstrated that their factor model was generalizable (i.e. *invariant*) across clinical and community samples (Rodenacker et al., 2016). Another study (Junghänel et al., 2020), which assessed the associations among ADHD and ODD symptomatology in a clinical sample, found individual domains of these diagnoses to be more independent from each other than was reported in a study assessing these symptoms in a community sample (Burns et al., 2020a). Accordingly, this suggests that the relation between disorders within the externalizing domain might be influenced by the sample setting (clinical vs. community). Moreover, some studies found measurement invariance across mothers and fathers, i.e., across informants who

rated their child's behavior in the same context (Burns et al., 2014). However, other research noted informant-specific variance across parent and teacher ratings, highlighting meaningful cross-situational variability in child behavior (Burns et al., 2020a; Thöne et al., 2021; Vitoratou et al., 2019). Studies including a self-report and systematically comparing this to teacher and parent reports are still rare.

Finally, there are well-characterized sex differences in the expression of externalizing symptoms, with much higher rates in boys than in girls at a ratio of 3:1 during childhood (Beauchaine et al., 2009; Copeland et al., 2011; reviewed by Martel 2013). Despite these differences in prevalence rates, several studies (Lee et al., 2016; Rodenacker et al., 2016, 2018) demonstrated measurement invariance across sex (but see King et al., 2018 for an exception), indicating that the structure of the externalizing spectrum itself is invariant across sex.

Summary of the current state of research

Overall, three general findings emerge from the literature, namely, that (a) existing categorical (i.e. ICD/DSM-based) diagnostic systems are challenged by empirically based models of psychopathology, which may provide a more accurate and potentially more parsimonious representation of psychopathology; (b) traditional bifactor models have often mistakenly been chosen as the “winning” model, despite serious statistical and interpretational difficulties; (c) integrating multiple informants’ reports is a key component of best practices in evidence-based assessment of child

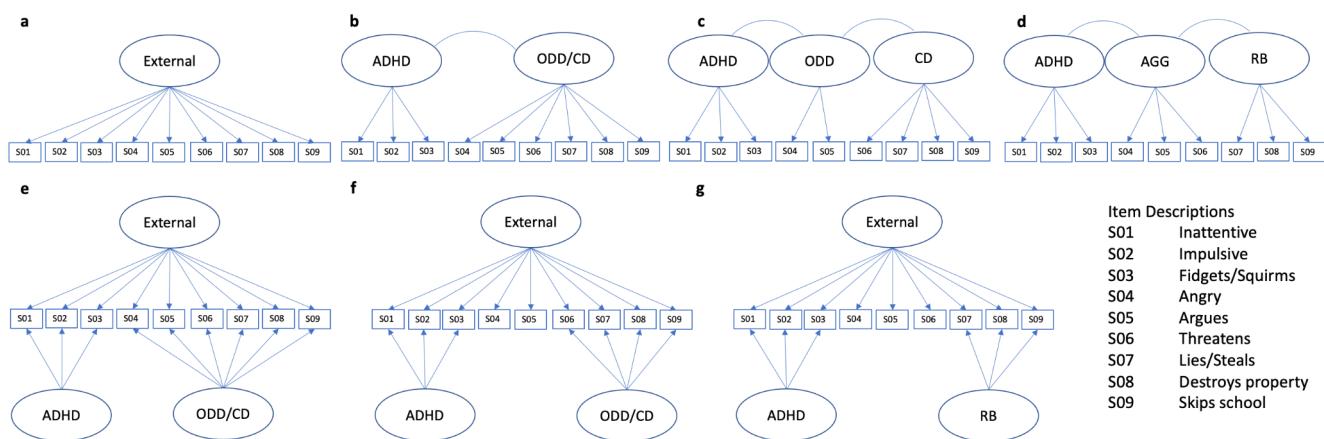


Fig. 1 Factor Models of Externalizing Symptoms. Note. Figure 1a Unidimensional model of externalizing symptomatology (Uni), b Factor model with two correlated factors (CFM-2; ADHD and ODD/CD), c Factor model with three correlated factors (CFM-3; ADHD, ODD, and CD), d Factor model with three correlated factors (CFM-3*; ADHD, AGG, and RB), e Bifactor model of externalizing symptomatology (BI Ext), f Bifactor S-1 model of externalizing symptomatology (BI Ext S-1) with ODD as a general reference factor, g Bifactor S-1 model of externalizing symptomatology (BI Ext S-1*) with AGG as a general reference factor. Item numbers are displayed in the boxes; for clarity of presentation, residuals are not shown. ADHD = attention-deficit/hyperactivity disorder, ODD = oppositional defiant disorder, CD = conduct disorder, AGG = aggression, RB = rule-breaking

psychopathology, since different informants will most likely provide diverging estimates of children's symptom severity.

Aim of the study

With this study, we aim to refine the knowledge about the structure underlying externalizing dimensions using symptoms from a standardized screening questionnaire. For this purpose, we will specify and compare several latent factor

models (see Fig. 1) according to an ICD/DSM-based perspective and a developmental theory-based perspective categorical and select the optimally fitting model according to four criteria: (1) model fit indices; (2) factor loadings; (3) omega statistics; and (4) model parsimony. Additionally, we aim to assess our optimally fitting model for measurement invariance across different sources (teacher, parent, self-ratings), sample settings (clinical, community), and sex (males, females).

Table 1 Descriptive Information for all Dimensions Across all Samples

Dimension	Number of items	α	Item-total correlation	$M (SD)$	Valid n
FBB-P-Clinical					
ADHD	3	0.81	0.63–0.71	1.24 (0.88)	742
ODD	2	0.80	.66 ^a	1.27 (0.96)	742
CD	4	0.59	0.18–0.56	0.27 (0.42)	739
ODD/CD	6	0.74	0.18–0.69	0.61 (0.053)	738
AGG	3	0.76	0.50–0.72	0.93 (0.76)	741
RB	3	0.51	0.21–0.43	0.28 (0.45)	739
External	9	0.83	0.13–0.74	0.81 (0.57)	738
SBB-Clinical					
ADHD	3	0.70	0.50–0.53	1.03 (0.72)	387
ODD	2	0.68	.52 ^a	0.99 (0.82)	387
CD	4	0.58	0.30–0.50	0.22 (0.36)	385
ODD/CD	6	0.68	0.24–0.56	0.48 (0.43)	385
AGG	3	0.63	0.37–0.56	0.70 (0.61)	386
RB	3	0.50	0.30–0.42	0.25 (0.40)	386
External	9	0.78	0.22–0.60	0.66 (0.47)	385
FBB-P-Community					
ADHD	3	0.78	0.57–0.65	0.58 (0.61)	2113
ODD	2	0.73	.58 ^a	0.57 (0.065)	2113
CD	4	0.55	0.30–0.42	0.07 (0.19)	2113
ODD/CD	6	0.67	0.28–0.64	0.23 (0.30)	2113
AGG	3	0.62	0.32–0.61	0.39 (0.47)	2113
RB	3	0.48	0.29–0.35	0.04 (0.15)	2113
External	9	0.80	0.24–0.68	0.35 (0.36)	2113
SBB-Community					
ADHD	3	0.70	0.50–0.54	0.59 (0.56)	960
ODD	2	0.63	.47 ^a	0.55 (0.65)	960
CD	4	0.57	0.26–0.49	0.09 (0.21)	960
ODD/CD	6	0.61	0.23–0.53	0.24 (0.30)	960
AGG	3	0.55	0.28–0.50	0.38 (0.47)	960
RB	3	0.48	0.25–0.40	0.10 (0.24)	960
External	9	0.76	0.19–0.59	0.36 (0.34)	960
FBB-T-Special Needs School					
ADHD	3	0.83	0.64–0.73	1.69 (0.94)	750
ODD	2	0.82	.70 ^a	1.59 (1.02)	751
CD	4	0.66	0.23–0.52	0.58 (0.63)	749
ODD/CD	6	0.80	0.20–0.76	0.92 (0.69)	744
AGG	3	0.84	0.67–0.78	1.34 (0.94)	750
RB	3	0.53	0.22–0.47	0.50 (0.61)	749
External	9	0.86	0.16–0.79	1.18 (0.70)	740

Note. n =sample size, α =Cronbach's alpha, M =mean, SD =standard deviation, ADHD=attention-deficit/hyperactivity disorder, ODD=oppositional defiant disorder, CD=conduct disorder, External=ADHD/ODD/CD dimensions combined, AGG=aggression, RB=rule-breaking, FBB-P-Clinical=clinical sample of children and adolescents based on proxy-ratings completed by parents, SBB-Clinical=clinical sample of adolescents based on self-ratings, FBB-P-Community=community sample of children and adolescents based on proxy-ratings completed by parents, SBB-Community=community sample of adolescents based on proxy-ratings completed by parents, FBB-T-Special Needs School=pupils attending special needs schools based on proxy-ratings completed by teachers

^a = only one item-total correlation as ODD dimension consists of two items only

Materials and methods

Measures

The German *Symptom Checklist for Screening of Mental Disorders in Children and Adolescents* is a standardized questionnaire based on the ICD-10 and DSM-5 from the German *Diagnostic System of Mental Disorders in Children and Adolescents based on the ICD-10 and DSM-5* (DISYPS-III; Döpfner & Götz-Dorten 2017). The symptom checklist comprises a proxy rating form that can be completed by parents or teachers (FBB-SCREEN) and a self-report form that can be completed by adolescents aged 11 years or older (SBB-SCREEN). All items are rated 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. Item scores of 2 and higher are interpreted as clinically relevant and considered to fulfill diagnostic criteria. The FBB/SBB-SCREEN consists of 50 items for a wide range of symptom criteria which form seven symptom scales and three broadband scales (Externalizing Symptoms, Internalizing Symptoms, Contact Behavior). In addition, five items assess functioning and psychological strain associated with these symptoms. Scale scores are computed by averaging the associated item scores. Psychometric analyses have supported the basic factorial configuration and reliability (i.e. internal consistency, item-total correlations) of the FBB/SBB-SCREEN scale scores in a large representative national sample (Döpfner & Götz-Dorten, 2017). In the present study, we included three items assessing ADHD core symptoms (S01 *Inattentive*, S02 *Impulsive*, S03 *Fidgets/Squirms*), two items assessing ODD core symptoms (S04 *Angry*, S05 *Argues*), and four items assessing CD core symptoms (S06 *Threatens*, S07 *Lies/Steals*, S08 *Destroys property*, S09 *Skips school*), resulting in a total of nine items assessing externalizing symptomatology. Following an empirically based perspective, we also formed an AGG scale (S04 *Angry*, S05 *Argues*, S06 *Threatens*) and an RB scale (S07 *Lies/Steals*, S08 *Destroys property*, S09 *Skips school*). This decision was supported by the fact that the Achenbach instruments, which are most often used to assess AGG and RB, also take a similar approach (Achenbach, 1991; Achenbach & Rescorla, 2001). Besides descriptive information (mean scores, standard deviations) for each sample, Cronbach's alpha was calculated, with values of >0.70 indicating acceptable internal consistency. Furthermore, the corrected item-total correlations were computed, with values of >0.30 considered acceptable. Table 1 summarizes statistical information regarding reliability for each sample.

Participants and procedure

For this study, we used data from a clinical sample (FBB-P-Clinical; SBB-Clinical), a community sample (FBB-P-Community; SBB-Community), and teacher ratings from children attending a school for special education (FBB-T-Special Needs School). The German *Symptom Checklist for Screening of Mental Disorders in Children and Adolescents* was completed by either parents or teachers (FBB-P/FBB-T) or as a self-report (SBB).

FBB-P-Clinical Participants were parents of 743 children and adolescents aged 4–18 years ($M=11.64$, $SD=3.66$, 62% males) who were referred to the outpatient unit of the Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy of the University Hospital of Cologne, Germany.

SBB-Clinical A total of 387 adolescents aged 11–18 years ($M=14.51$, $SD=2.07$, 53% males) were referred to the outpatient unit of the Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy of the University Hospital of Cologne, Germany. Of these 387 participants, we further obtained a parent report in 375 of cases, which was included in the FBB-P-Clinical sample reported above.

FBB-P-Community Participants were recruited from three communities in Western Germany (Cologne, Bedburg, Siegburg). A total of 10,000 households with children aged between 3;0 and 17;11 years were selected randomly using a stratified sampling design. Of these, 2,113 parents returned the FBB-SCREEN of their children ($M=10.14$, $SD=4.44$, 50% males). Previous analyses demonstrated this approach to be nonbiased and representative (Döpfner et al., 2008).

SBB-Community A total of 960 adolescents aged 11–18 years ($M=14.26$, $SD=1.82$, 50% males) returned the SBB-SCREEN. We further obtained a parent report in 917 of these cases, which was included in the FBB-P-Community sample.

FBB-T-Special needs School A total of 755 children and adolescents aged 6–18 years ($M=11.17$, $SD=3.01$, 85% males) were recruited from six schools for special education in Western Germany as part of a larger research project (PEARL; Hennemann et al., 2020). Of these schools, five schools followed a German program for emotional and social skills development and one school additionally followed a program for language and learning. The current sample only included teacher ratings from children

participating in the program for emotional and social skills development.

Statistical analyses

Factor models

As a starting point for analyzing the factor structure of externalizing symptomatology, we constructed a unidimensional factor model (see Fig. 1a), given the high comorbidity and shared liability of externalizing symptoms (Willcutt et al., 2012). Then, we followed the ICD/DSM-based nomenclature and constructed a correlated factors model with three factors (CFM-3: ADHD, ODD, and CD; see Fig. 1c), and, given the diagnostic overlap between ODD and CD, also a correlated factors model with two factors (CFM-2: ADHD and ODD/CD; see Fig. 1b). In addition, we constructed a traditional bifactor model with two s-factors (BI Ext; see Fig. 1e), which simultaneously addressed general and specific facets of externalizing symptomatology. Furthermore, we constructed a bifactor S-1 model (BI Ext S-1; see Fig. 1f) using ODD symptoms as the general reference factor. When choosing a reference factor, data-driven as well as theory-driven approaches can be applied. We based the selection of the ODD reference factor on theoretical considerations (cf. Beauchaine, 2015). More specifically, ODD symptoms may potentially link the two disorder domains of ADHD and CD as middlemen and that is, after all, what we aim to investigate, namely, the links between the disorder domains. Accordingly, our reasoning for selecting the reference factor is based on a combination of theoretical considerations (cf. Beauchaine, 2015) and a particular research interest (cf. Eid, 2020). Following an developmental theory-based approach, we tested a correlated factors model with three factors (CFM-3*: ADHD, AGG, and RB; see Fig. 1d). Additionally, we tested an alternative bifactor S-1 model (BI Ext S-1*; see Fig. 1 g), in which we specified AGG symptoms as the general reference factor, analogous to the BI Ext S-1 model. Of note, formally, any domain could have been chosen as the reference factor and – assuming the absence of abnormal factor loadings – relationships between the domains could be reliably assessed in either case (Eid, 2020).¹

¹ Note that we also tested second-order factor models given the hierarchical structure of the externalizing spectrum (e.g. Kotov et al., 2017). In this model, the second-order factor reflects the operation of a higher-order construct that explains the intercorrelations (i.e., shared variance) among the first-order constructs. We constructed an ICD/DSM-based second-order factor model with ADHD, ODD, and CD as first-order factors. We further tested an empirically based second-order factor model with ADHD, AGG, and RB as first-order factors. The two models could not be identified, indicating that the data did not fit the

Model estimation

All latent construct analyses were conducted using Mplus version 8.4 (Muthén & Muthén, 2017). The weighted least squares means and variance adjusted estimator (WLSMV; delta parameterization) was chosen because it has been specifically proposed for ordinal data (Li, 2016). Missing data were handled using the recommended default Mplus missing strategy for WLSMV (Muthén & Muthén, 2017). Missing data were generally low across samples (see Table 1 for valid *n*). Covariance coverage was greater than 99.1% in all samples.

Evaluating model fit

Global model goodness of fit was assessed using χ^2 test of exact model fit, the comparative fit index (CFI), the Tucker-Lewis index (TLI), the root mean square error of approximation (RMSEA) with 90% confidence intervals, and the standardized root mean square residual (SRMR). Considering that the χ^2 test is highly influenced by sample size (Bentler & Bonett, 1980), we relied more on the CFI, TLI, RMSEA, and SRMR. Model fit was regarded as good when CFI and TLI values were ≥ 0.95 (Hu & Bentler, 1999) and RMSEA and SRMR were ≤ 0.05 (Browne & Cudeck, 1992; Hooper et al., 2008). RMSEA and SRMR values < 0.08 were regarded as acceptable (Browne & Cudeck, 1992; Hooper et al., 2008). In addition, the Bayesian information criterion (BIC) was calculated using maximum likelihood estimation with robust standard errors to compare our non-nested models with each other (see also Burnham & Anderson 2004). The BIC provides a way to rank a set of models from “best” to “worst” based on model uncertainty. In particular, the model with the lowest BIC contains the least uncertainty and therefore serves as the most preferred model. When comparing models, a BIC difference of 10 corresponds to the odds being 150:1 that the model with the smaller BIC value is the better-fitting model and is considered “very strong” evidence for the model with the smaller BIC value (Raftery, 1995). Besides, we computed difference tests based on log-likelihood values and scaling correction factors obtained with the maximum likelihood estimator with robust standard errors. The test statistic (TRd) was used as described on the Mplus website (Mplus, 2022) and calculated as follows: $TRd = -2*(L_0-L_1)/cd$, where L_0 is the log-likelihood for the null model, L_1 is the log-likelihood for the alternative model, and cd can be calculated as follows: $cd = (p_0*c_0-p_1*c_1)/(p_0-p_1)$, where p_0 is the number of parameters estimated in the null model, p_1 is the number of parameters estimated in the alternative model,

models well. Reasons for this non-identification are provided in the Discussion section.

c_0 is the scaling correction factor for the null model, and c_1 is the scaling correction factor for the alternative model. The test statistic TR_d follows a chi-square distribution with $df=p_1-p_0$. The corresponding p -values were obtained using this chi-square calculator: <https://www.socscistatistics.com/pvalues/chidistribution.aspx>.

On a factor level, we computed the explained common variance (ECV), a measure of unidimensionality, for our bifactor models. The ECV refers to the proportion of all

common variance explained by that factor (Reise, 2012). Furthermore, we computed omega statistics as model-based reliability indices of the g-factor and the s-factors (Rodriguez et al., 2016). In a bifactor model, coefficient omega (ω) reflects the amount of reliable variance attributable to the g-factor and the s-factors taken together (McDonald, 1999; Revelle & Zinbarg, 2009). In turn, omega hierarchical (ω_H) (McDonald, 1999; Zinbarg et al., 2005) assesses the amount of reliable variance attributable only to the g-factor, whereas

Table 2 Model Evaluation in all Samples

Evaluation Criterion	FBB-P-Clinical	SBB-Clinical	FBB-P-Community	SBB-Community	FBB-T-Special Needs School
(1) Model Fit Indices ^a	1. BI Ext S-1 2. BI Ext 3. BI Ext S-1* 4. CFM-3* 5. CFM-3 6. CFM-2 7. Uni	1. BI Ext S-1 2. BI Ext 3. BI Ext S-1* 4. CFM-3 5. CFM-3* 6. CFM-2 7. Uni	1. BI Ext 2. BI Ext S-1 3. CFM-3 4. BI Ext S-1* 5. CFM-3* 6. CFM-2 7. Uni	1. BI Ext 2. BI Ext S-1 3. CFM-3 4. BI-Ext S-1* 5. CFM-3* 6. CFM-2 7. Uni	1. BI Ext 2. BI Ext S-1 3. CFM-3* 4. CFM-2 5. CFM-3 6. Uni 7. Uni
Interim Conclusion	Uni model discarded due to worst model fit indices in all five samples.				
(2) Factor loadings	<i>BI Ext</i> : non-significant factor loadings (S04/S05)	<i>BI Ext</i> : non-significant (S05/S06) / negative factor loadings (S04)	<i>BI Ext</i> : non-significant (S05) / negative (S04)	<i>BI Ext</i> : non-significant (S05) / negative (S04)	<i>BI Ext</i> / <i>BI Ext S-1</i> : All ODD/CD factor loadings anomalous (non-significant/negative/>1). <i>BI Ext S-1*</i> did not converge
Interim Conclusion	BI Ext model discarded due to negative/non-significant factor loadings in all five samples.				
Comparison of model fit CFM-2 vs. CFM-3 ^b	CFM-3 better than CFM-2 ($p<.001$)	CFM-3 not better than CFM-2 ($p=.101$)			
Interim Conclusion	CFM-2 excluded due to superiority of CFM-3 in four out of five samples.				
Comparison of model fit CFM-3 vs. BI Ext S-1 ^b	BI Ext S-1 better than CFM-3 ($p<.001$)	BI Ext S-1 better than CFM-3 ($p=.003$)	BI Ext S-1 better than CFM-3 ($p=.034$)	BI Ext S-1 better than CFM-3 ($p=.002$)	BI Ext S-1 better than CFM-3 ($p<.001$)
Comparison of model fit CFM-3* vs. BI Ext S-1* ^b	BI Ext S-1* not better than CFM-3* ($p=.668$)	BI Ext S-1* better than CFM-3* ($p=.044$)	BI Ext S-1* not better than CFM-3* ($p=.213$)	BI Ext S-1* better than CFM-3* ($p<.001$)	BI Ext S-1* better than CFM-3* ($p<.001$)
(3) Omega-Statistics	Unstable s-factors for all bifactor models	<i>BI Ext</i> / <i>BI Ext S-1</i> : s-factors not stable/interpretable due to factor loadings > 1. <i>BI Ext S-1*</i> did not converge			
(4) Parsimony	CFM-3/CFM-3* more parsimonious than BI Ext S-1/BI Ext S-1*	CFM-3/CFM-3* more parsimonious than BI Ext S-1/BI Ext S-1*	CFM-3/CFM-3* more parsimonious than BI Ext S-1/BI Ext S-1*	CFM-3/CFM-3* more parsimonious than BI Ext S-1/BI Ext S-1*	CFM-3/CFM-3* more parsimonious than BI Ext S-1
Chosen models	CFM-3/CFM-3*	CFM-3/CFM-3*	CFM-3/CFM-3*	CFM-3/CFM-3*	CFM-3/CFM-3*

Note. FBB-P-Clinical = clinical sample of children and adolescents based on proxy ratings completed by parents, SBB-Clinical = clinical sample of adolescents based on self-ratings, FBB-P-Community = community sample of children and adolescents based on proxy ratings completed by parents, SBB-Community = community sample of adolescents based on proxy ratings completed by parents, FBB-T-Special Needs School = pupils attending special needs schools based on proxy ratings completed by teachers, Uni = unidimensional model of externalizing symptomatology, CFM-2 = factor model with two correlated factors (ADHD, ODD/CD), CFM-3 = factor model with three correlated factors (ADHD, ODD, CD), CFM-3* = factor model with three correlated factors (ADHD, AGG, RB), BI Ext = bifactor model of externalizing symptomatology, BI Ext S-1 = bifactor S-1 model of externalizing symptomatology with ODD as a general reference factor, BI Ext S-1* = bifactor S-1 model of externalizing symptomatology with AGG as a general reference factor, ADHD = attention-deficit/hyperactivity disorder, ODD = oppositional defiant disorder, CD = conduct disorder, AGG = aggression, RB = rule-breaking

^a Descriptive order according to the comparative fit index CFI, in case of a tie in a second step the root mean square error of approximation (RMSEA), in case of a tie in a third step the root mean square residuals (SRMR)

^b $\alpha=0.05$

omega hierarchical subscale (ω_{HS}) assesses the amount of reliable variance attributable to each s-factor (Reise et al., 2013). Importantly, ω_{HS} allows us to evaluate how much systematic reliable variance exists beyond the reliable variance accounted for by the g-factor, and whether it seems reasonable to interpret the s-factors as stable, independent factors.

Measurement invariance

We assessed our prevailing factor model for measurement invariance (Vandenberg & Lance, 2000) with regard to sample setting (community vs. clinical sample), sex (males vs. females), and source (parents vs. self). Furthermore, we combined all datasets and additionally assessed our prevailing model for measurement invariance in a sample of teacher ratings for children attending a special needs school (FBB-T-Special Needs School). Configural invariance indicates whether the basic model configuration (i.e. the same items belonging to the same factors) is equivalent across groups. Metric invariance additionally requires item loadings to be equal across groups, whereas scalar invariance further requires item thresholds to be equal across groups. Since the χ^2 statistic is sensitive to sample size, we computed difference tests for goodness-of-fit indices (CFI, TLI, RMSEA, SRMR; Theta parameterization) to evaluate measurement invariance. The same cut-off values mentioned above are valid to assess configural invariance. Additionally, it has been suggested to consider a change in CFI of ≤ -0.01 (Cheung & Rensvold, 2002), a change in SRMR of $\leq +0.03$ for testing metric invariance and of $\leq +0.01$ for testing scalar invariance (Chen, 2007), as well as equal or better fit of TLI and RMSEA (Marsh et al., 2010) as indicative that the imposed constraints were invariant across groups.

Results

Descriptive information and scale reliability

Table 1 shows descriptive information of the dimensions ADHD, ODD, CD, ODD/CD, AGG, RB, and combined external symptomatology. The internal consistency (Cronbach's α) was highest for the ADHD dimension as well as for the combined external symptomatology in all samples (ADHD: $0.70 \leq \alpha \leq 0.84$; combined external symptomatology: $0.76 \leq \alpha \leq 0.86$). By comparison, Cronbach's α ranged between 0.63 and 0.82 for the ODD dimension, between 0.55 and 0.66 for the CD dimension, between 0.61 and 0.80 for the combined ODD/CD dimension, between 0.55 and 0.84 for the AGG dimension, and between 0.48 and 0.53 for the RB dimension. Table 1 also shows the item-total

correlation for each scale. The lowest value for the dimensions CD, ODD/CD, RB, and combined external symptomatology was always due to the item *Skips school* (S09), which showed a considerably lower item-total correlation compared to all other items.

Factor-analytic evaluation

We applied four criteria to find the best model for all samples (Table 2).

First, we calculated prominent goodness-of-fit indices (CFI, TLI, RMSEA, and SRMR) to evaluate which of the tested models demonstrated the best model fit in each sample (Table S1). The unidimensional model had the worst model fit in all five samples and was therefore considered inadequate (CFI or TLI < 0.95 or RMSEA or SRMR > 0.08). One exception concerned the SBB-Clinical Sample, in which model fit was considered adequate to good (CFI and TLI ≥ 0.95 , RMSEA and SRMR ≤ 0.08), while still showing the worst model fit of all five models tested. The CFM-2 showed a good model fit in the two samples SBB-Clinical and FBB-P-Community, an adequate model fit in the samples FBB-P-Clinical sample and an inadequate fit in the SBB-Community sample (CFI and TLI < 0.95) and the FBB-T-Special Needs School (RMSEA = 0.084). Correlations between the dimensions ADHD and ODD/CD were high and significant ($r = .74 - .81$, all $p < .001$). The CFM-3 showed good model fit in all samples, solely the RMSEA in the FBB-T-Special Needs School sample was slightly above the recommended cut-off of 0.080 (RMSEA = 0.087). All correlations in all samples were high and significant (ODD-ADHD: $r = .75 - .86$, CD-ADHD: $r = .55 - .75$, CD-ODD: $r = .69 - .97$, all $p < .001$). The CFM-3* model showed good model fit in all samples; only the TLI in the SBB-Community sample was slightly below the recommended cut-off of 0.950 (TLI = 0.944). Correlations in all samples were high and significant (ADHD-AGG: $r = .75 - .85$, ADHD-RB: $r = .53 - .68$, AGG-RB: $r = .74 - .85$). The traditional bifactor model BI Ext was (one of) the best models in terms of model fit in all samples (Table 2). Similarly, the BI Ext S-1 and the BI Ext S-1* models showed good model fit (CFI and TLI ≥ 0.95 , RMSEA and SRMR ≤ 0.05) in most samples. In the SBB-Community sample, the BI Ext S-1* was in an adequate range (TLI = 0.935; RMSEA = 0.066). In the FBB-T-Special Needs School sample, the model fit for the BI Ext S-1 was in an adequate range, with an RMSEA = 0.055. The BI Ext S-1* model was not identified in this sample. In the BI Ext S-1 and the BI Ext S-1* model, we did not find any significant residual correlations between ADHD and CD or between ADHD and RB, respectively. In addition to the goodness-of-fit indices, we calculated the sample-size

adjusted BIC (Table S3). In all samples except for the FBB-T-Special Needs School sample, we found the bifactor models and the CFM-3 / CFM-3* models to be among the best models, while the exact order of these three models differed between samples (Table 2). The unidimensional model was the worst model according to BIC values in all five samples. Therefore, we decided to discard the unidimensional model after applying the first criterion *model fit indices*.

Second, we took a closer look at the *factor loadings* and found for the BI Ext S-1 model in the FBB-T-Special Needs School sample anomalous factor loadings >1 for the item S07 *Lies/Steals* (Table S2). By definition, this is impossible, since factor loadings range between 0 and 1. Hence, it did not seem reasonable to interpret omega statistics in the BI Ext S-1 model within the FBB-T-Special Needs School sample. Since the BI Ext S-1* model did not converge in the FBB-T-Special Needs School sample, it did not seem reasonable to further interpret this model in this specific sample. Regarding the BI Ext model, we found anomalous factor loadings, such as negative and/or non-significant factor loadings and/or factor loadings >1 in all five samples. Such statistical anomalies complicate a meaningful interpretation in terms of content. Therefore, we decided to exclude the BI Ext model in our model selection process. We then performed likelihood ratio tests to compare the goodness of fit between the CFM-2 and the CFM-3 models across all samples (Table S6). We found that the CFM-3 model was superior to the CFM-2 model in all samples ($p < .001$), except for the FBB-T Special Needs School sample ($p = .101$). Therefore, we decided to exclude the CFM-2 model due to the superiority of the CFM-3 model in our selection process (Table 2). Then, we compared the CFM-3 and the BI Ext S-1 models using likelihood ratio tests (Table S6). We found that the BI Ext S-1 model was superior to the CFM-3 model in all five samples ($p \leq .001 \leq .034$). When comparing the alternative CFM-3* and the BI Ext S-1* models using likelihood ratio tests, we found that the BI Ext S-1* was superior to the CFM-3* model in the three samples SBB-Clinical ($p = .044$), SBB-Community ($p < .001$), and FBB-T Special Needs School ($p < .001$), but not in the FBB-P-Clinical ($p = .668$) and the FBB-P-Community ($p = .213$) samples (Table 2).

Third, we calculated *omega statistics* to evaluate the model-based reliability, or, in other words, the stability of the s-factors. In our samples, the omega statistics revealed that in the BI Ext S-1 / BI Ext S-1* models, none of the s-factors in any of the samples remained a stable independent component, as the ω_{HS} for all s-factors remained below the recommended cut-off of 0.50 (Reise et al., 2013). Omega statistics for the FBB-T-Special Needs School sample were not computed, as their interpretation would have been flawed due to anomalous factor loadings. The same

held true for the ECV in this sample. The ECV (Table S2) supported the finding that all s-factors in the BI Ext S-1, and the BI Ext S-1* model were weakly defined, as the ECV was small for all s-factors (ADHD: ECV = 0.11 – 0.16, CD: ECV = 0.11 – 0.25, RB: ECV = 0.09 – 0.17) compared to the g-factor (ECV = 0.63 – 0.79). Exploratory analyses for the BI Ext models also showed weak factors and low ECV values (Table S2), reinforcing our decision to exclude the BI Ext model from our selection process.

Fourth, our final criterion *model parsimony* favored the CFM-3 / CFM-3* models over all bifactor S-1 models, as models with fewer restrictions are generally favorable.

Considering all four criteria, we discarded models step by step, which culminated in the same models capturing the data best in all five samples – the CFM-3 and the CFM-3* model (Table 2). Overall, model fits were adequate to good; solely the RMSEA (0.087) for the CFM-3 model in the FBB-T-Special Needs School sample and the TLI (0.944) for the CFM-3* model in the SBB-Community sample were slightly above the recommended cut-off. All factor loadings of the two models across all samples were significant, with low standard errors (Tables S4, S5). The factor loadings for items S01 – S08 were all high, ranging from 0.63 – 0.97 for the CFM-3 and from 0.66 to 0.99 for the CFM-3* model. The factor loadings for the item S09 *Skips school* were considerably lower in the CFM-3 (0.24 – 0.67) and CFM-3* (0.27 – 0.69) models than the other items in most samples but remained significant. We conclude that both the ICD/DSM-based CFM-3 (ADHD, ODD, CD) model and the developmental theory-based CFM-3* (ADHD, AGG, RB) model provided a sound view of externalizing dimensions.

Measurement invariance

Due to our different samples regarding multiple sources (parents, teachers, self-ratings) and sample settings (clinical, community, special needs school), we additionally tested for measurement invariance in our CFM-3 and CFM-3* models (Table 3). For this purpose, we excluded the FBB-T-Special Needs School sample in our measurement invariance analyses, as this sample differed with regard to source *AND* sample setting, whereas the other four samples only differed in one aspect (source *OR* sample setting). We tested for measurement invariance across sources by combining the FBB-P-Clinical + FBB-P-Community samples and the SBB-Clinical + SBB-Community samples (parents vs. self-ratings). Our results demonstrated that measurement invariance across sources could be assumed on a configural, metric as well as scalar level, as indicated by changes in model fit indices (Table 3), suggesting the same factor organization, loading patterns, and symptom thresholds across

sources. We then tested for measurement invariance across sample setting (clinical vs. community) by combining the samples FBB-P-Clinical + SBB-Clinical and FBB-P-Community + SBB Community. The results demonstrated that measurement invariance could be assumed on a configural and metric level. Scalar invariance was almost given, but as the change in CFI of -0.011 was slightly above the recommended cut-off of -0.01 , we decided to remain on the level of metric invariance. These results imply that the CFM-3 / CFM-3* models present the same factor organization and loading patterns, but there is some variation in the individual item thresholds across sample settings. Next, we assessed measurement invariance across all five datasets combined. Similarly to measurement invariance across sample settings, we found invariance on a configural and metric level, but not on a scalar level, as the change in CFI was -0.18 (>-0.01). Given the well characterized sex differences in the expression of externalizing symptoms (Beauchaine et al., 2009; Copeland et al., 2011; reviewed by Martel 2013), we additionally assessed measurement invariance regarding sex across all five samples and found that configural, metric, and scalar invariance for the CFM-3 / CFM-3* models were supported (Table 3). This finding indicates the same factor organization, loading patterns, and symptom thresholds across males and females.

Discussion

The aim of this study was to refine the knowledge about the structure underlying externalizing dimensions. For this purpose, we analyzed items from a screening instrument assessing these symptoms across five large samples, which differed with respect to sample setting (clinical, community, special needs school) and source (i.e. parents, teachers, self-ratings).

A first conclusion drawn from our analyses is that a separation into different dimensions appears to be justifiable, as the unidimensional model was rejected in all samples. The study by Beauchaine et al., (2010) supports this finding, as the authors stated that in addition to inherent impulsivity, which underlies ADHD, ODD, and CD and often leads to an early presence of ADHD symptomatology, the co-occurrence of ODD and CD additionally depends on environmental influences. As the CFM-3 model yielded better results regarding global model and likelihood ratio tests fit than did the CFM-2 model in four out of our five samples, the separation of ODD and CD was supported. In this ICD/DSM-based CFM-3 model, all items except for the item S09 *Skips school* showed high factor loadings on their respective first-order factor. Following an developmental theory-based perspective, we also specified a model with three correlated

factors (CFM-3*; ADHD, AGG, RB), thereby describing disruptive behaviors as aggressive and rule-breaking problems (Burt, 2012; Burt et al., 2015). This CFM-3* model demonstrated overall good model fit and high factor correlations across samples (AGG-RB: $r=.74$ – $.85$). These correlations are somewhat higher than those reported in the factor analytic literature (cf. Burt, 2012). One explanation for this differing finding might lie in the different informants considered (Burt, 2012).

Due to the high correlations between all first-order factors in all samples in the CFM-3 model, we specified a traditional bifactor model. For this model, we had to combine ODD and CD, as the model was otherwise not identified, with solely two indicators loading on the ODD factor. As the two-factor solution was also adequate, we did not expect this to be a major problem, however this has to be pointed out as a limitation as it impedes the comparison with the other models. The model fit for the traditional bifactor model was superior. This in itself should not be overinterpreted, as the fit of traditional bifactor models is generally superior to the fit of first-order correlated factor models given that more free parameters are estimated, making the model less restrictive (Bonifay et al., 2017). However, upon closer examination of the factor loadings, it became apparent that anomalous factor loadings, such as negative and non-significant factor loadings, were present in all five samples and we even observed an unreasonable factor loading >1 in the FBB-T-Special Needs School sample, which further exacerbated the interpretation. These very small or even negative factor loadings do not concern the ADHD dimension, but only the ODD/CD dimension in all samples. Such statistical anomalies related to traditional bifactor models are consistent with methodological concerns (Eid et al., 2017) and empirical studies demonstrating a variety of anomalous results associated with the application of traditional bifactor models to externalizing symptoms (Arias et al., 2018; Burns et al., 2020a; Rodenacker et al., 2018; Thöne et al., 2021). Anomalous loadings might be a result of the generally mistaken assumption of interchangeability of domains, which is a statistical prerequisite for accurately applying a bifactor model (Heinrich et al., 2021). However, domains in psychopathology are most likely structurally different (Eid et al., 2017; Heinrich et al., 2021). The structural differences between domains are apparent when examining the correlations between subdomains in the CFM-3 / CFM-3* models. For interchangeability, these correlations would have to be equal, which is not the case. Anomalous results, which we observed in the FBB-T-Special Needs School sample for the item *Lies/Steals* (S07), are a major problem for interpreting the respective model, since test statistics cannot reliably be interpreted. Common reasons for these anomalous results are the extraction of too many factors,

Table 3 Measurement Invariance Analyses

Model	Parents ($n=2836$) vs. Children/Adolescents ($n=1347$)						Clinical ($n=1130$) vs. Community ($n=3073$)						FBB-P-Clinical ($n=743$) vs. SBB-P-Clinical ($n=387$) vs. FBB-P-Community ($n=2113$) vs. SBB-Community ($n=960$) vs. FBB-T-Special Needs School ($n=755$)						Males ($n=2804$) vs. Females ($n=2154$)						
	$\chi^2(df)$	CFI	TLI	RMSEA	SRMR	$\chi^2(df)$	CFI	TLI	RMSEA	SRMR	$\chi^2(df)$	CFI	TLI	RMSEA	SRMR	$\chi^2(df)$	CFI	TLI	RMSEA	SRMR	$\chi^2(df)$	CFI	TLI	RMSEA	SRMR
Configural	221.252* (48)	0.993 (0.036, 0.047)	0.990 (0.036, 0.047)	0.041 (0.036, 0.047)	0.032 (48)	225.544* (48)	0.990 (0.037, 0.048)	0.985 (0.037, 0.048)	0.042 (0.037, 0.048)	0.032 (120)	400.715* (120)	0.990 (0.043, 0.054)	0.985 (0.043, 0.054)	0.049 (0.043, 0.054)	0.037 (48)	371.450* (48)	0.992 (0.047, 0.057)	0.989 (0.047, 0.057)	0.052 (0.047, 0.057)	0.030 (0.047, 0.057)	371.450* (48)	0.992 (0.047, 0.057)	0.989 (0.047, 0.057)	0.052 (0.047, 0.057)	0.030 (0.047, 0.057)
Metric	185.995* (54)	0.995 (0.029, 0.040)	0.993 (0.037, 0.046)	0.034 (0.037, 0.046)	0.033 (78)	278.236* (54)	0.987 (78)	0.983 (0.046, 0.055)	0.044 (0.046, 0.055)	0.045 (144)	469.701* (144)	0.988 (0.043, 0.053)	0.985 (0.043, 0.053)	0.048 (0.043, 0.053)	0.054 (54)	277.243* (54)	0.995 (0.041, 0.056)	0.993 (0.041, 0.056)	0.041 (0.041, 0.056)	0.032 (0.041, 0.056)	277.243* (54)	0.995 (0.041, 0.056)	0.993 (0.041, 0.056)	0.041 (0.041, 0.056)	0.032 (0.041, 0.056)
Scalar	361.142* (78)	0.989 (0.037, 0.046)	0.990 (0.037, 0.046)	0.042 (0.037, 0.046)	0.035 (78)	495.088* (78)	0.976 (78)	0.978 (0.046, 0.055)	0.050 (0.046, 0.055)	0.047 (240)	1074.683* (240)	0.970 (0.056, 0.063)	0.978 (0.056, 0.063)	0.059 (0.056, 0.063)	0.058 (78)	421.858* (78)	0.992 (0.038, 0.046)	0.993 (0.038, 0.046)	0.042 (0.038, 0.046)	0.034 (0.038, 0.046)	421.858* (78)	0.992 (0.038, 0.046)	0.993 (0.038, 0.046)	0.042 (0.038, 0.046)	0.034 (0.038, 0.046)
Configural	240.253* (48)	0.992 (0.038, 0.049)	0.989 (0.038, 0.049)	0.044 (0.038, 0.049)	0.033 (48)	250.309* (48)	0.988 (0.039, 0.050)	0.982 (0.039, 0.050)	0.045 (0.040, 0.050)	0.036 (120)	424.053* (120)	0.989 (0.045, 0.056)	0.984 (0.045, 0.056)	0.051 (0.045, 0.056)	0.040 (48)	308.478* (48)	0.994 (0.042, 0.052)	0.991 (0.042, 0.052)	0.047 (0.042, 0.052)	0.028 (0.042, 0.052)	308.478* (48)	0.994 (0.042, 0.052)	0.991 (0.042, 0.052)	0.047 (0.042, 0.052)	0.028 (0.042, 0.052)
Metric	204.615* (54)	0.994 (0.031, 0.042)	0.992 (0.031, 0.042)	0.036 (0.038, 0.046)	0.034 (78)	286.482* (54)	0.987 (78)	0.982 (0.047, 0.056)	0.045 (0.040, 0.056)	0.044 (144)	474.904* (144)	0.988 (0.043, 0.053)	0.985 (0.043, 0.053)	0.048 (0.043, 0.053)	0.052 (54)	230.638* (54)	0.996 (0.032, 0.041)	0.995 (0.032, 0.041)	0.036 (0.032, 0.041)	0.029 (0.032, 0.041)	230.638* (54)	0.996 (0.032, 0.041)	0.995 (0.032, 0.041)	0.036 (0.032, 0.041)	0.029 (0.032, 0.041)
Scalar	366.053* (78)	0.989 (0.038, 0.046)	0.990 (0.038, 0.046)	0.042 (0.038, 0.046)	0.037 (78)	510.191* (78)	0.975 (78)	0.977 (0.047, 0.056)	0.051 (0.047, 0.056)	0.047 (240)	1121.013* (240)	0.969 (0.057, 0.064)	0.976 (0.057, 0.064)	0.061 (0.057, 0.064)	0.057 (78)	371.034* (78)	0.993 (0.035, 0.043)	0.994 (0.035, 0.043)	0.039 (0.035, 0.043)	0.030 (0.035, 0.043)	371.034* (78)	0.993 (0.035, 0.043)	0.994 (0.035, 0.043)	0.039 (0.035, 0.043)	0.030 (0.035, 0.043)

Note. FBB-P-Clinical = clinical sample of children and adolescents based on proxy ratings completed by parents, SBB-Clinical = clinical sample of adolescents based on self-ratings, FBB-P-Community = community sample of children and adolescents based on proxy ratings completed by parents, SBB-P-Community = community sample of adolescents based on proxy ratings completed by parents, FBB-T-Special Needs School = pupils attending special needs schools based on proxy ratings completed by teachers. χ^2 = Chi-Square, df = degrees of freedom, CFI = comparative fit index, TLI = Tucker-Lewis index, RMSEA = root mean square error of approximation, CI = confidence interval, SRMR = standardized root mean square residuals, CFM-3 = factor model with three correlated factors (ADHD, ODD, and CD), CFM-3* = factor model with three correlated factors (ADHD, AGG, and RB), ADHD = attention-deficit/hyperactivity disorder, ODD = oppositional defiant disorder, CD = conduct disorder, AGG = aggression, RB = rule-breaking

n = sample size, * $p < .001$

a small sample size, small variability in an indicator or a misspecification of the model, potentially through adding or omitting paths and/or posing restrictions based on non-conclusive assumptions (Chen et al., 2001). As our sample size was quite large ($n=755$), and we only extracted three factors with at least three indicators per factor, which was both identical (i.e., same factorial configuration) or at least similar to the other samples, we do not believe this to be the reason for these anomalous results. Descriptive statistics regarding the variance and skewness of item S07 in the FBB-Special Needs School sample were unremarkable and similar to the other items, eliminating this consideration as a potential cause for these anomalous results. A potential misspecification of the model in this particular case cannot fully be excluded, however, we found item S07 to significantly and highly load on its respective factor in the CFM-3, suggesting that the allocation to the CD factor was justified. Excluding item S07 from the model led to an elimination of the statistical anomalies in both models (BI Ext / BI Ext S-1), however, based on descriptive statistics and as this item is a symptom criterion of CD in the DSM-5, we did not see enough grounds for such a radical decision. Therefore, despite encountering this problem in our bifactor models, we proceeded to analyze further important details. To analyze a bifactor model in a statistically sound manner, it is important to consider additional bifactor-specific indices beyond the global model fit indices (Rodriguez et al., 2016), especially since bifactor models tend to overfit. Important bifactor-specific indices include, but are not limited to, the omega statistics and ECV. In the FBB-T-Special Needs School sample, it was not possible to calculate the omega statistics and the ECV, as the anomalous factor loading >1 on item S07 would have led to a misinterpretation of the results. For the other four models, we computed omega statistics and ECV. As expected from the small and/or negative factor loadings on the ODD/CD dimension, this s-factor vanished in all four remaining samples, explaining little remaining variance beyond the variance already explained by the g-factor. The s-factor ADHD was more strongly defined than the ODD/CD one, but was not strong enough to be considered a reliable s-factor (Reise, 2012). The g-factor explains more variance than the s-factors do for almost all items in all samples, with very few sample-specific exceptions. The ECV for the g-factor in all samples lay between 0.62 for the SBB-Community sample and 0.77 for the FBB-P-Community sample. According to Rodriguez et al., (2016), an ECV for the g-factor >0.80 supports the idea of one-dimensionality. In our bifactor models, the ECV for the g-factor was below this cut-off for all samples, and additionally, we found the ECV especially for the ADHD s-factor to be quite high (0.33 – 0.44). Despite the superior model fit of the bifactor model and ECV values pointing at potential

multidimensionality, we decided to discard this model, as it appears that the remaining s-factor could not be interpreted in a stable, reliable manner. Anomalous factor loadings additionally exacerbate a straightforward interpretation of g- and s-factors. One likely reason for the unstable s-factors is that the g-factor already explains a huge proportion of the common variance, leaving little variance to be explained for the individual s-factors. The strength of the g-factor here is supported by the very high correlation between the factors ADHD and ODD/CD in the CFM-2, ranging from 0.74 to 0.81, and the ω_H values between 0.73 and 0.84.

The strong g-factor we found could suggest one-dimensionality, but as already discussed, the unidimensional model did not fit the data well and the ECV values for the s-factors in the traditional bifactor model suggest that the s-factors still account for a significant part of reliable variance after partialling out the influence of the g-factor. In order to keep the concept of s-factors that are all associated with each other, likely through some sort of common factor, we specified two bifactor S-1 models. As opposed to a traditional bifactor model, an S-1 model considers structural differences among domains and does not assume interchangeability (Heinrich et al., 2021). In the ICD/DSM-based BI Ext S-1 model, we were able to keep the separation between ODD and CD. As we were interested in the relation of ADHD and CD with ODD, and since we observed the highest factor loadings on the ODD items in our BI Ext models across all samples, we specified ODD as the general reference domain, with the ADHD and CD dimension being orthogonal (uncorrelated) to it. Theoretically supporting the choice of our reference factor, ODD symptoms have been suggested as middlemen connecting impulsivity (one factor of ADHD) and CD (Beauchaine, 2015). We therefore regarded it as a domain of special interest, which according to Eid (2020) is a valid base for this selection. For the ODD dimension, no s-factor was modeled. Analogously, the AGG dimension was specified as the reference factor in the developmental theory-based Bi Ext S-1* model. When applying a bifactor S-1 model, one has to dismiss the idea of identifying a common overarching factor; however, as pointed out by Heinrich et al., (2021), it frequently remains unclear what this factor even really stands for. Due to the *a priori* definition of the general reference factor, a bifactor S-1 model allows for a straightforward interpretation of all factors, including their relation to one another. This is supported by the fact that bifactor S-1 models generally avoid anomalous results (Heinrich et al., 2021). In our case, we found a strong reduction of anomalous factor loadings for the BI Ext S-1 compared to the BI Ext model, but the factor loading >1 in the FBB-T-Special Needs School sample on item S07 remained, which did not allow for a closer examination of this model in this specific sample. The BI Ext

S-1* model did not converge in the FBB-T-Special Needs School sample, further indicating problems in this specific sample. In all other samples, the significant positive factor loadings allowed for a straightforward interpretation of all factors for both bifactor S-1 models. However, the s-factors of the BI Ext S-1 model still accounted for little reliable variance after accounting for the variance explained by the general reference factor ODD. This shows that the two ODD items already explain so much variance in the ADHD and the CD domains that not enough variance remained for these s-factors to explain to be considered stable. The ECV values in this model were similar to the BI Ext model and ranged between 0.63 and 0.77, again remaining just under the recommended cut-off of 0.80 proposed by Rodriguez et al., (2016). A similar pattern emerged for the BI Ext S-1* model, although the differentiation between AGG and RB might depict a more coherent view than the ICD/DSM-based perspective of ODD and CD (cf. Burt, 2012).

When comparing our factor models, we found both the ICD/DSM-based CFM-3 model and the developmental theory-based CFM-3* model to fit the data best in all samples. This was also supported using likelihood ratio tests and calculating BIC values to compare model fit, as the CFM-3 / CFM-3* models were superior to the Uni model and the CFM-2 model in all samples (with the only exception involving the FBB-T-Special Needs School sample, where the CFM-2 was equally as good as the CFM-3 model). The specific factors in both the traditional bifactor model and in the two bifactor S-1 models explained too little variance to be interpreted in an appropriately useful manner. In addition, the CFM-3 / CFM-3* models explain the data in a more parsimonious way than all bifactor models. We therefore found both the CFM-3 / CFM-3* to be the models representing our data best, although the CFM-3* model with the AGG and RB dimensions may offer a more accurate representation of disruptive behavior problems (Burt, 2012; Burt et al., 2015). However, it has to be pointed out that the high correlations between the different factors, in combination with the strong g-factor in both bifactor models, suggest unmodeled commonality. These high correlations correspond to the frequently found comorbidities of ADHD, ODD, and CD (Beauchaine et al., 2010; Willcutt et al., 2012). Although proposing a general externalizing psychopathology factor (or a *p*-factor in general) using bifactor modeling may be tempting, theoretical and statistical considerations, previous work, and the present study show that bifactor models may be unsuitable for this. It remains an open question, how the associations between these three frequently co-occurring diagnoses can be modelled in the most ideal way, i.e. to capture all specific aspects of each symptom complex but at the same time avoid an excessive number of comorbid diagnoses.

In a final step, we evaluated our CFM-3 / CFM-3* models' measurement invariance. When testing for measurement invariance across all five samples, scalar invariance could not be supported. However, scalar invariance was confirmed across sources (parents vs. self-ratings), demonstrating that despite the frequently reported cross-informant discrepancies (De Los Reyes et al., 2015; Los Reyes et al., 2019), the structure of the externalizing spectrum itself is invariant across sources. Furthermore, there was only minor evidence of scalar non-invariance across sample setting (community vs. clinical sample), suggesting that the samples differ with regard to symptom severity on one or more variables. Finally, we found that scalar invariance was supported across sex (males vs. females). These findings are in line with previous studies (Lee et al., 2016; Rodenacker et al., 2016, 2018) and provide support for the validity of the symptom dimensions across males and females.

The present study may have implications for model selection when examining the associations between psychological dimensions. Bifactor models of psychopathology have become increasingly popular in recent years and are often applied to search for general factors in psychopathology (Caspi & Moffitt, 2018), despite the statistical and interpretational difficulties outlined above. While our traditional bifactor model demonstrated good model fit, we nevertheless decided to discard it after carefully evaluating additional statistical indices. This finding is in line with previous observations that researchers often mistakenly regard their bifactor model as superior and fail to take into account further statistical indices (Arias et al., 2018; Watts et al., 2019). Although our results are limited to the externalizing spectrum in children, our approach of discarding factor models step by step according to several criteria may be adopted in the *p*-factor literature as well. Here, similar statistical and interpretational difficulties become apparent, calling a g-factor of a traditional bifactor model into question as the proper candidate for a *p*-factor (Levin-Aspenson et al., 2021; Watts et al., 2019).

Our findings may also have implications regarding categorical and dimensional perspectives on mental disorders. From an ICD/DSM perspective, mental disorders are conceptualized as categorical concepts, although there is little evidence that the underlying structure of mental disorders is, in fact, categorical in nature (Achenbach, 2020). Instead, empirical research points towards a dimensional perspective (Achenbach, 2020). More specifically, the two dimensions AGG and RB may offer more nuanced indications of clinical significance than the ODD / CD diagnoses (Burt, 2012). For example, the AGG and RB dimensions may outperform CD diagnoses when predicting adult symptoms of antisocial personality disorder (Burt et al., 2011). The results from our research show that both correlated factor models from an

ICD/DSM-based and developmental theory-based perspective provide a sound view of externalizing dimensions. At this point, therefore, we cannot conclude whether one perspective is truly superior to the other. Below, we provide suggestions for answering these questions unambiguously in the future.

Limitations and future research

This study has several strengths and limitations. One major strength is that we systematically tested the structure underlying externalizing dimensions according to an ICD/DSM-based perspective and a developmental theory-based perspective across five large samples, which differed with respect to sample setting (clinical, community, special needs school) and source (i.e. parents, teachers, self-ratings). One limitation of this study is the small number of indicators comprising the prevailing factors. This is due to the fact that the items assessed in our analyses belong to a screening instrument, which is supposed to indicate the need for further detailed assessment of one or more dimensions. The low number of indicators particularly concerns our ODD dimension (two items), since a minimum of three items per factor is generally needed to prevent model specification issues (cf. Koran, 2020). Nevertheless, we would like to note that our main conclusions do not change when specifying the factor AGG (three items) instead of ODD. In this regard, we must also concede that second-order factor models could not be specified. Even though our correlated factor models provide a sound view of externalizing dimensions, the high factor correlations may suggest a common overarching factor. However, the bifactor models could not adequately capture this commonality while at the same time keeping clearly interpretable distinct factors. Future research should therefore capture the dimensions of externalizing psychopathology with a larger number of indicators per factor to allow for better internal differentiation between dimensions. Likewise, we recommend specifying higher-order factor models in an attempt to incorporate what is common to the aforementioned externalizing mental disorders, approaching a model similarly to the HiTOP (Kotov et al., 2017).

Another limitation of the present study is that no external criterion variables were analyzed. However, we would like to point out that the main purpose of our study was to demonstrate the appropriate choice of factor models following several model-based criteria. Moreover, we followed a multitrait-multisource approach and analyzed five samples from different settings, rated by different informants, an approach which provides additional value to our study.

Regarding future directions, it would be informative to include symptoms of more severe externalizing symptoms, such as antisocial behavior or substance abuse, into a larger externalizing spectrum (cf. Krueger et al., 2007). Moreover, we encourage future researchers to compare the performance of their bifactor models with alternative structures of psychopathology. We believe that correlated factor models should not be discarded due to a sole reliance on global model fit indices. Instead, the performance of bifactor models' g- and s-factors warrant critical inspection— for example by applying our four criteria of model quality. Following this approach may help to improve study quality, and may ultimately help to contribute to the reorganization of mental disorders and efforts to identify a more parsimonious structure of psychopathology.

Conclusions

To the best of our knowledge, the present study is the first to compare the structure of a screening instrument for symptoms of externalizing disorders according to an ICD/DSM-based perspective and a developmental theory-based perspective across five large samples, which differed with respect to sample setting (clinical, community, special needs school) and source (i.e. parents, teachers, self-ratings).

Following an ICD/DSM-based perspective and a developmental theory-based perspective, we specified and systematically evaluated several latent factor models according to four criteria: (1) model fit indices; (2) factor loadings; (3) omega statistics; and (4) model parsimony. We discarded a unidimensional model due to poor model fit, a model with two correlated factors (CFM-2; ADHD, ODD/CD) due to statistical superiority of the model with three correlated factors (CFM-3; ADHD, ODD, CD) and all versions of the bifactor (S-1) models due to statistical and interpretational difficulties. We found that both the ICD/DSM-based model with three correlated factors (CFM-3; ADHD, ODD, CD) and the developmental theory-based model with three correlated factors (CFM-3*; ADHD, AGG, RB) provide a sound view of externalizing dimensions. These two models demonstrate metric invariance across all five datasets combined and across sample settings (clinical, community), as well as scalar invariance across sources (parents, self-ratings) and sex. While the dimensions AGG and RB may depict a more empirically coherent view than the ICD/DSM-based perspective of ODD and CD, at this point, we cannot clearly determine whether one perspective really outperforms the other. Given current research efforts toward a dimensional classification of psychopathology and the search for a p-factor in general, our statistical criteria for selecting the

optimally fitting factor model have important implications for the appropriate analysis of latent factor models.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10862-022-09983-7>.

Acknowledgements We would like to gratefully thank all the participating patients and their families. Furthermore, we would like to thank all teachers and school principals of the special needs schools.

Authors' contributions All authors made substantial contributions to the conception or design of the work. Material preparation and data analyses were performed by **AT** and **MJ** and reviewed by **MD**. The first draft of the manuscript was written by **AT** and **MJ** and all authors commented on previous versions of the manuscript. All authors critically revised the manuscript for important intellectual content and approved the final manuscript. All authors agree 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 Not applicable.

Open Access funding enabled and organized by Projekt DEAL.

Data Availability (data transparency): Available upon reasonable request.

Code Availability (software application or custom code): Available upon reasonable request.

Declarations

Conflicts of interest/competing interest **AGD** 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 the diagnostic system DISYPS-III. All remaining authors declare no conflicts of interest.

Ethics approval All procedures performed in this study were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. All studies, including this one, conducted with data collected in the process of routine therapy at the School for Child and Adolescent Behavior Therapy (AKiP) were approved by the ethics committee of the University of Cologne.

Consent to participate All parents and teachers of the participating children provided their written informed consent for data collection. Adolescents/young adults of legal age provided this consent themselves.

Consent for publication All parents and teachers of the participating children provided their written informed consent for anonymous data analysis and publication. Adolescents/young adults of legal age provided this consent themselves.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the

source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4–18 and 1991 profile*. University of Vermont, Department of Psychiatry
- Achenbach, T. M. (2020). Bottom-Up and Top-Down Paradigms for Psychopathology: A Half-Century Odyssey. *Annual Review of Clinical Psychology*, 16, 1–24. <https://doi.org/10.1146/annurev-clinpsy-071119-115831>
- Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/Adolescent Behavioral and Emotional Problems: Implications of Cross-Informant Correlations for Situational Specificity. *Psychological Bulletin*, 101(2), 213–232. <https://doi.org/10.1037/0033-2909.101.2.213>
- Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles: An integrated system of multi-informant assessment*. ASEBA
- American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Association. <https://doi.org/10.1176/appi.books.9780890425596>
- Arias, V. B., Ponce, F. P., & Núñez, D. E. (2018). Bifactor Models of Attention-Deficit/Hyperactivity Disorder (ADHD): An Evaluation of Three Necessary but Underused Psychometric Indexes. *Assessment*, 25(7), 885–897. <https://doi.org/10.1177/1073191116679260>
- Beauchaine, T. P. (2015). Future Directions in Emotion Dysregulation and Youth Psychopathology. *Journal of Clinical Child and Adolescent Psychology*, 44(5), 875–896. <https://doi.org/10.1080/15374416.2015.1038827>
- Beauchaine, T. P., Hinshaw, S. P., & Pang, K. L. (2010). Comorbidity of Attention-Deficit/Hyperactivity Disorder and Early-Onset Conduct Disorder: Biological, Environmental, and Developmental Mechanisms. *Clinical Psychology: Science and Practice*, 17(4), 327–336. <https://doi.org/10.1111/j.1468-2850.2010.01224.x>
- Beauchaine, T. P., Klein, D. N., Crowell, S. E., Deridibe, C., & Gatzke-Kopp, L. (2009). Multifinality in the development of personality disorders: A Biology × Sex × Environment interaction model of antisocial and borderline traits. *Development and Psychopathology*, 21(3), 735–770. <https://doi.org/10.1017/S0954579409000418>
- Beauchaine, T. P., & McNulty, T. (2013). Comorbidities and continuities as ontogenetic processes: Toward a developmental spectrum model of externalizing psychopathology. *Development And Psychopathology*, 25(4pt2), 1505–1528. <https://doi.org/10.1017/S0954579413000746>
- Beauchaine, T. P., Zisner, A. R., & Sauder, C. L. (2017). Trait Impulsivity and the Externalizing Spectrum. *Annu Rev Clin Psychol*, 16(1), 343–368. <https://doi.org/10.1146/annurev-clinpsy-021815-093253>
- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88(3), 588–606. <https://doi.org/10.1037/0033-2909.88.3.588>
- Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three Concerns With Applying a Bifactor Model as a Structure of

- Psychopathology. *Clin Psychol Sci*, 5(1), 184–186. <https://doi.org/10.1177/2167702616657069>
- Browne, M. W., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociol Methods Res*, 21(2), 230–258
- Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociol Methods Res*, 33(2), 261–304. <https://doi.org/10.1177/0049124104268644>
- Burns, G. L., De Moura, M. A., Beauchaine, T. P., & McBurnett, K. (2014). Bifactor latent structure of ADHD/ODD symptoms: Predictions of dual-pathway/trait-impulsivity etiological models of ADHD. *Journal Of Child Psychology And Psychiatry*, 55(4), 393–401. <https://doi.org/10.1111/jcpp.12165>
- Burns, G. L., Geiser, C., Servera, M., Becker, S. P., & Beauchaine, T. P. (2020a). Application of the Bifactor S – 1 Model to Multisource Ratings of ADHD/ODD Symptoms: an Appropriate Bifactor Model for Symptom Ratings. *Journal Of Abnormal Child Psychology*, 48(7), 881–894. <https://doi.org/10.1007/s10802-019-00604>
- Burns, G. L., Geiser, C., Servera, M., Becker, S. P., & Beauchaine, T. P. (2020b). Promises and Pitfalls of Latent Variable Approaches to Understanding Promises and Pitfalls of Latent Variable Approaches to Understanding Psychopathology: Reply to Burke and Johnston, Eid, Junghänel and Colleagues, and Willoughby. *Journal Of Abnormal Child Psychology*, 48, 917–922. <https://doi.org/10.1007/s10802-020-00656>
- Burt, S. A. (2012). How do we optimally conceptualize the heterogeneity within antisocial behavior? An argument for aggressive versus non-aggressive behavioral dimensions. *Clinical Psychology Review*, 32(4), 263–279. <https://doi.org/10.1016/j.cpr.2012.02.006>
- Burt, S. A., Donnellan, M. B., Iacono, W. G., & McGue, M. (2011). Age-of-onset or Behavioral Sub-types? A Prospective Comparison of Two Approaches to Characterizing the Heterogeneity within Antisocial Behavior. *Journal Of Abnormal Child Psychology*, 39(5), 633–644. <https://doi.org/10.1007/s10802-011-9491-9>
- Burt, S. A., Rescorla, L. A., Achenbach, T. M., Ivanova, M. Y., Almqvist, F., Begovac, I. ... Verhulst, F. C. (2015). The association between aggressive and non-aggressive antisocial problems as measured with the Achenbach System of Empirically Based Assessment: A study of 27,861 parent-adolescent dyads from 25 societies. *Personality and Individual Differences*, 85(May), 86–92. <https://doi.org/10.1016/j.paid.2015.04.036>
- Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S. ... Moffitt, T. E. (2014). The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clin Psychol Sci*, 2(2), 119–137. <https://doi.org/10.1177/2167702613497473>
- Caspi, A., & Moffitt, T. E. (2018). All for one and one for all: Mental disorders in one dimension. *American Journal Of Psychiatry*, 175(9), 831–844. <https://doi.org/10.1176/appi.ajp.2018.17121383>
- Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research*, 29(4), 468–508
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct Equ Modeling*, 14(3), 464–504. <https://doi.org/10.1080/10705510701301834>
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Struct Equ Modeling*, 9(2), 233–255. <https://doi.org/10.1207/S15328007SEM0902>
- Copeland, W., Shanahan, L., Costello, E. J., & Angold, A. (2011). Cumulative prevalence of psychiatric disorders by young adulthood: A prospective cohort analysis from the great smoky mountains study. *Journal of the American Academy of Child and Adolescent Psychiatry*, 50(3), 252–261. <https://doi.org/10.1016/j.jaac.2010.12.014>
- De Los Reyes, A. (2011). Introduction to the special section: More than measurement error: Discovering meaning behind informant discrepancies in clinical assessments of children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, 40(1), 1–9. <https://doi.org/10.1080/15374416.2011.533405>
- De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D. E., & Rabinowitz, J. (2015). The Validity of the Multi-Informant Approach to Assessing Child and Adolescent Mental Health. *Psychological Bulletin*, 141(4), 858–900. <https://doi.org/10.1037/a0038498>
- De Los Reyes, A., Lerner, M. D., Keeley, L. M., Weber, R. J., Drabick, D. A. G., Rabinowitz, J., & Goodman, K. L. (2019). Improving Interpretability of Subjective Assessments About Psychological Phenomena: A Review and Cross-Cultural Meta-Analysis. *Review of General Psychology*, 23(3), 293–319. <https://doi.org/10.1177/1089268019837645>
- De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kundey, S. M. A. (2013). Principles Underlying the Use of Multiple Informants' Reports. *Annu Rev Clin Psychol*, 9, 123–149. <https://doi.org/10.1146/annurev-clinpsy-050212-185617>
- Dirks, M. A., De Los Reyes, A., Briggs-Gowan, M., Cella, D., & Wakschlag, L. S. (2012). Annual research review: Embracing not erasing contextual variability in children's behavior - Theory and utility in the selection and use of methods and informants in developmental psychopathology. *Journal Of Child Psychology And Psychiatry*, 53(5), 558–574. <https://doi.org/10.1111/j.1469-7610.2012.02537.x>
- Döpfner, M., & Götz-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)*. Hogrefe
- Döpfner, M., Götz-Dorten, A., Lehmkuhl, G., Breuer, D., & Goetz, H. (2008). *Diagnostik-System für psychische Störungen nach ICD-10 und DSM-IV für Kinder und Jugendliche – II [Diagnostic System of Mental Disorders in Children and Adolescents based on the ICD-10 and DSM-IV] (DISYPS-II)*. Hogrefe
- Eid, M. (2020). Multi-Faceted Constructs in Abnormal Psychology: Implications of the Bifactor S – 1 Model for Individual Clinical Assessment. *Journal Of Abnormal Child Psychology*, 48, 895–900. <https://doi.org/10.1007/s10802-020-00624-9>
- Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods*, 22(3), 541–562. <https://doi.org/10.1037/met0000083>
- Harden, K. P., Patterson, M. W., Briley, D. A., Engelhardt, L. E., Kretsch, N., Mann, F. D. ... Tucker-Drob, E. M. (2015). Developmental changes in genetic and environmental influences on rule-breaking and aggression: Age and pubertal development. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 56(12), 1370–1379. <https://doi.org/10.1111/jcpp.12419>
- Heinrich, M., Geiser, C., Zagorscak, P., Leonard Burns, G., Bohn, J., Becker, S. P. ... Knaevelsrud, C. (2021). On the Meaning of the General Factor of Psychopathology (“P-Factor”) in Symmetrical Bifactor Models: Recommendations for Future Research From the Bifactor-(S – 1) Perspective. *Assessment*, 1–12. <https://doi.org/10.1177/10731911211060298>
- Hennemann, T., Casale, G., Leidig, T., Fleskes, T., Döpfner, M., & Hanisch, C. (2020). Psychische Gesundheit von Schülerinnen und Schülern an Förderschulen mit dem Förderschwerpunkt (PEARL). *Zeitschrift Für Heilpädagogik*, 71, 44–57
- Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, 6(1), 53–60. <https://doi.org/10.21427/D79B73>

- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct Equ Modeling*, 6(1), 1–55.
- Junghänel, M., Rodenacker, K., Dose, C., & Döpfner, M. (2020). Applying the Bifactor S-1 Model to Ratings of ADHD/ODD Symptoms: A Commentary on Burns et al. (2019) and a Re-Analysis. *J. Abnorm. Child Psychol.*, 48, 901–905. <https://doi.org/10.1007/s10802-020-00637-4>
- King, K. M., Luk, J. W., Witkiewitz, K., Racz, S., McMahon, R. J., & Wu, J. (2018). Externalizing Behavior Across Childhood as Reported by Parents and Teachers: A Partial Measurement Invariance Model. *Assessment*, 25(6), 744–758. <https://doi.org/10.1177/1073191116660381>
- Koran, J. (2020). Indicators per Factor in Confirmatory Factor Analysis: More is not Always Better. *Structural Equation Modeling*, 27(5), 765–772. <https://doi.org/10.1080/10705511.2019.1706527>
- Kotov, R., Waszczuk, M. A., Krueger, R. F., Forbes, M. K., Watson, D., Clark, L. A. ... Zimmerman, S. E., M (2017). The hierarchical taxonomy of psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal Of Abnormal Psychology*, 126(4), 454–477. <https://doi.org/10.1037/abn0000258>
- Krueger, R. F., Markon, K. E., Patrick, C. J., Benning, S. D., & Kramer, M. D. (2007). Linking antisocial behavior, substance use, and personality: An integrative quantitative model of the adult externalizing spectrum. *Journal Of Abnormal Psychology*, 116(4), 645–666. <https://doi.org/10.1037/0021-843X.116.4.645>
- Lahey, B. B., Applegate, B., Hakes, J. K., Zald, D. H., Hariri, A. R., & Rathouz, P. J. (2012). Is There a general factor of prevalent psychopathology during adulthood? *Journal Of Abnormal Psychology*, 121(4), 971–977. <https://doi.org/10.1037/a0028355>
- Lee, S., Burns, G. L., Beauchaine, T. P., & Becker, S. P. (2016). Bifactor Latent Structure of Attention-Deficit/Hyperactivity Disorder (ADHD)/Oppositional Defiant Disorder (ODD) Symptoms and First-Order Latent Structure of Sluggish Cognitive Tempo Symptoms. *Psychological Assessment*, 28(8), 917–928. <https://doi.org/10.1037/pas0000232>
- Levin-Aspenson, H. F., Watson, D., Clark, L. A., & Zimmerman, M. (2021). What is the General Factor of Psychopathology? Consistency of the p Factor Across Samples. *Assessment*, 28(4), 1035–1049. <https://doi.org/10.1177/1073191120954921>
- Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, 48(3), 936–949. <https://doi.org/10.3758/s13428-015-0619-7>
- Marsh, H. W., Lüdtke, O., Muthén, B. O., Asparouhov, T., Morin, A. J., Trautwein, U., & Nagengast, B. (2010). A New Look at the Big Five Factor Structure Through Exploratory Structural Equation Modeling. *Psychological Assessment*, 22(3), 471–491. <https://doi.org/10.1037/a0019227>
- Martel, M. M. (2013). Sexual selection and sex differences in the prevalence of childhood externalizing and adolescent internalizing disorders. *Psychological Bulletin*, 139(6), 1221–1259. <https://doi.org/10.1037/a0032247>
- McDonald, R. P. (1999). *Test theory: A unified approach*. Erlbaum Mplus (2022). *Chi-Square Difference Testing Using the Satorra-Bentler Scaled Chi-Square*. <http://www.statmodel.com/chidiff.shtml>
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide* (8th ed.). Muthén & Muthén
- Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociol. Methodol.*, 111–163
- Reise, S. P. (2012). The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research*, 47(5), 667–696. <https://doi.org/10.1080/00273171.2012.715555>
- Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal Of Personality Assessment*, 95(2), 129–140. <https://doi.org/10.1080/00223891.2012.725437>
- Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, 74(1), 145–154
- Rodenacker, K., Hautmann, C., Götz-Dorten, A., & Döpfner, M. (2016). Bifactor models show a superior model fit: Examination of the factorial validity of parent-reported and self-reported symptoms of Attention-Deficit/Hyperactivity Disorders in children and adolescents. *Psychopathology*, 49(1), 31–39. <https://doi.org/10.1159/000442295>
- Rodenacker, K., Hautmann, C., Götz-Dorten, A., & Döpfner, M. (2018). Evidence for the trait-impulsivity etiological model in a clinical sample: Bifactor structure and its relation to impairment and environmental risk. *Journal Of Abnormal Child Psychology*, 46(4), 659–669. <https://doi.org/10.1007/s10802-017-0329-y>
- Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying Bifactor Statistical Indices in the Evaluation of Psychological Measures. *Journal Of Personality Assessment*, 98(3), 223–237. <https://doi.org/10.1080/00223891.2015.1089249>
- Thöne, A. K., Junghänel, M., Götz-Dorten, A., Dose, C., Hautmann, C., Jendreizik, L. T. ... Döpfner, M. (2021). Disentangling symptoms of externalizing disorders in children using multiple measures and informants. *Psychological Assessment*, 33(11), 1065. <https://doi.org/10.1037/pas0001053>
- Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organ Res Methods*, 3(1), 4–69. <https://doi.org/10.1177/109442810031002>
- Vitoratou, S., Garcia-Rosales, A., Banaschewski, T., Sonuga-Barke, E., Buitelaar, J., Oades, R. D. ... Chen, W. (2019). Is the endorsement of the Attention Deficit Hyperactivity Disorder symptom criteria ratings influenced by informant assessment, gender, age, and co-occurring disorders? A measurement invariance study. *Int J Methods Psychiatr*, 28(4), 1–14. <https://doi.org/10.1002/mpr.1794>
- Watts, A. L., Poore, H. E., & Waldman, I. D. (2019). Riskier Tests of the Validity of the Bifactor Model of Psychopathology. *Clin Psychol Sci*, 7(6), 1285–1303. <https://doi.org/10.1177/2167702619855035>
- Willcutt, E. G., Nigg, J. T., Pennington, B. F., Solanto, M. V., Rohde, L. A., Tannock, R. ... Lahey, B. B. (2012). Validity of DSM-IV attention deficit/hyperactivity disorder symptom dimensions and subtypes. *Journal Of Abnormal Psychology*, 121(4), 991–1010. <https://doi.org/10.1037/a0027347>
- World Health Organization. (1993). *The ICD-10 classification of mental and behavioural disorders: Diagnostic criteria for research*. World Health Organization
- World Health Organization. (2019). *International Statistical Classification of Diseases and Related Health Problems ICD-11* (11th ed.). World Health Organization
- Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's, α Revelle's β and McDonald's ω : Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70(1), 123–133. <https://doi.org/10.1007/s11336-003-0974-7>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.