


# Harmonizing bifactor models of psychopathology between distinct assessment instruments: Reliability, measurement invariance, and authenticity

Maurício Scopel Hoffmann<sup>1,2,3,4</sup>  | Tyler Maxwell Moore<sup>5</sup> | Luiza Kvitko Axelrud<sup>2,3</sup> | Nim Tottenham<sup>6</sup> | Luis Augusto Rohde<sup>2,7,8</sup> | Michael Peter Milham<sup>9,10</sup> | Theodore Daniel Satterthwaite<sup>5,11</sup> | Giovanni Abrahão Salum<sup>2,3,7,8,10</sup>

<sup>1</sup>Department of Neuropsychiatry, Universidade Federal de Santa Maria, Santa Maria, Brazil

<sup>2</sup>Graduate Program in Psychiatry and Behavioral Sciences, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

<sup>3</sup>Section on Negative Affect and Social Processes, Hospital de Clínicas de Porto Alegre, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

<sup>4</sup>Care Policy and Evaluation Centre, London School of Economics and Political Science, London, UK

<sup>5</sup>Department of Psychiatry, University of Pennsylvania, Philadelphia, Pennsylvania, USA

<sup>6</sup>Department of Psychology, Columbia University, New York, New York, USA

<sup>7</sup>National Institute of Developmental Psychiatry for Children and Adolescents (INCT-CNPq), São Paulo, Brazil

<sup>8</sup>Department of Psychiatry and Legal Medicine, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

<sup>9</sup>Nathan S. Kline Institute for Psychiatric Research, Orangeburg, New York, USA

<sup>10</sup>Center for the Developing Brain, Child Mind Institute, New York, New York, USA

<sup>11</sup>Lifespan Informatics and Neuroimaging Center, Philadelphia, Pennsylvania, USA

## Correspondence

Maurício Scopel Hoffmann, Universidade Federal de Santa Maria, Santa Maria, Brazil.  
Email: [mauricio.hoffmann@ufsm.br](mailto:mauricio.hoffmann@ufsm.br)

## Funding information

National Institute of Mental Health, Grant/Award Number: R01MH120482; National Institute of Development Psychiatric, Grant/Award Numbers: CNPq 465550/2014-2, FAPESP 2014/50917-0; Beijing Municipal Science and Technology Commission, Grant/Award Numbers: Z161100002616023, Z181100001518003; Development Program of Guangdong Province, Grant/Award Number: 2019B030335001

## Abstract

**Objectives:** Model configuration is important for mental health data harmonization. We provide a method to investigate the performance of different bifactor model configurations to harmonize different instruments.

**Methods:** We used data from six samples from the Reproducible Brain Charts initiative ( $N = 8,606$ , ages 5–22 years, 41.0% females). We harmonized items from two psychopathology instruments, Child Behavior Checklist (CBCL) and GOASSESS, based on semantic content. We estimated bifactor models using confirmatory factor analysis, and calculated their model fit, factor reliability, between-instrument invariance, and authenticity (i.e., the correlation and factor score difference between the harmonized and original models).

**Results:** Five out of 12 model configurations presented acceptable fit and were instrument-invariant. Correlations between the harmonized factor scores and the original full-item models were high for the p-factor ( $>0.89$ ) and small to moderate (0.12–0.81) for the specific factors. 6.3%–50.9% of participants presented factor score differences between harmonized and original models higher than 0.5 z-score.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. International Journal of Methods in Psychiatric Research published by John Wiley & Sons Ltd.

**Conclusions:** The CBCL-GOASSESS harmonization indicates that few models provide reliable specific factors and are instrument-invariant. Moreover, authenticity was high for the p-factor and moderate for specific factors. Future studies can use this framework to examine the impact of harmonizing instruments in psychiatric research.

**KEYWORDS**

CBCL, data integration, GOASSESS, harmonization, p-factor, questionnaire

## 1 | INTRODUCTION

Phenotypic data integration is essential to advance knowledge in psychiatry and psychology. Most of the recent advances in reproducible findings for the mental health sciences involve the aggregation of many datasets (Sullivan et al., 2018; Thompson et al., 2014). A pressing problem in that endeavor is that different datasets frequently use different assessment instruments (Mansolf et al., 2020; McElroy et al., 2021; Thompson et al., 2022). Understanding the best methods to integrate data from distinct datasets can result in increased reproducibility of research findings, accelerate scientific discovery, and, consequently, translate scientific findings into practice.

Most available harmonization strategies rely on item-wise semantic matching (McElroy et al., 2021). Using this method, researchers who aim to harmonize two questionnaires look for semantically common items and select an item pool that is similar enough to allow a combined analysis. We recently showed these methods' advantages over alternative data harmonization methods (Hoffmann, Moore, Axelrud, Tottenham, Pan, et al., 2022, submitted). However, when harmonizing such items, the researcher must choose a measurement model for putting items together. We can take the Child Behavior Checklist (CBCL) as an example. The CBCL is one of the most widely used parent-reported assessment of emotional and behavioral problems in youth, containing 120 items (Achenbach & Rescorla, 2001). Using this instrument, 11 published models are using the bifactor structure (Constantinou & Fonagy, 2019; Hoffmann, Moore, Axelrud, Tottenham, Zuo, et al., 2022). Therefore, given the diversity of modeling choices for assessment instruments, one must have a clear rationale for choosing one measurement model over another.

At least three indicators might be relevant for making such decisions. The first is model fit and reliability. These measures indicate how well the model explains the structural relationships among the included variables and the extent to which the bifactor model's dimensions are internally consistent (Bornovalova et al., 2020; Rodriguez et al., 2016). The second indicator is instrument measurement invariance. Instrument invariance test if differences between scores assessed by two questionnaires that aim to measure the same underlying construct are due only to differences in the underlying construct and not due to instrument differences. Finally, a third indicator that is less frequently

considered is authenticity: the degree of concordance between factor scores from the harmonized models (i.e., with the limited item pool) and factor scores from original models (i.e., with the full item pool). While model fit and instrument invariance are indicators of internal validity, authenticity is a measure of external validity—a way to assess the harmonization costs in terms of the deviation from published measurement models.

Most measurement models of psychopathology that used harmonized items from different instruments (i.e., item-harmonized models) have been tested using unidimensional structures, which presented good model fit and invariance testing (Gondek et al., 2021; McElroy et al., 2021; Ploubidis et al., 2019). However, the impact of item harmonization on bifactor models has been sparsely explored. A bifactor measurement model assumes that the variation among items can be explained by, aside from error, a general factor (shared variance among all items) and orthogonal specific factors (explained variance of a subset of items above and beyond the general factor). Previous evidence has demonstrated the predictive and biological validity of the general and specific factors of psychopathology using this model (Allegrini et al., 2020; Caspi & Moffitt, 2018; Kaczkurkin et al., 2018, 2019; Laceulle et al., 2020; Sallis et al., 2019; Shanmugan et al., 2016; Thompson et al., 2021; Waszczuk et al., 2021).

However, despite its clinical utility has been proposed (Caspi et al., 2020), there is a current debate on the substantive meaning of the derived factors (general/p- and specific factors), as they could represent impairment, cognitive dysfunction, negative affectivity, etc. (Caspi & Moffitt, 2018; Heinrich et al., 2020; Smith et al., 2020) which could depend on instrument and sample that are being used for modeling (Fernández de la Cruz et al., 2018; Levin-Aspenson et al., 2021; Watts et al., 2020). Thus, data aggregation (i.e., harmonizing different instruments and samples) might shed further light on the validity and utility of the bifactor model by overcoming its sensitivity on instruments and samples. In this study, we evaluated a method for testing bifactor models while performing item-wise harmonization of two mental health questionnaires, the CBCL and GOASSESS, a questionnaire derived from the Kiddie Schedule for Affective Disorders and Schizophrenia. The method consists of (1) harmonizing items using bifactor models previously used in studies of the CBCL ( $n = 11$  models) or GOASSESS ( $n = 1$  model), (2) testing model fit and factor reliability, (3) measured questionnaire's

invariance for the bifactor models using harmonized items, and (4) measured the correlation and factor score difference between bifactor models using harmonized and un-harmonized items with the full item set (i.e., authenticity). We hypothesized that the bifactor models using harmonized items would present reliable p-factors and be questionnaire-invariant. Moreover, we predicted that the p-factors from the models using harmonized items would be authentic with models containing the full item set.

## 2 | METHODS

### 2.1 | Sample

The samples consist of those subjects within the Reproducible Brain Charts (RBC) initiative. A complete description of the RBC samples can be found elsewhere (Hoffmann, Moore, Axelrud, Tottenham, Zuo, et al., 2022). Briefly, it contains phenotypic data from six large-scale developmental imaging cohorts, which are the Philadelphia Neurodevelopmental Cohort (PNC) (Satterthwaite et al., 2016), the Brazilian High-Risk Cohort Study for Mental Conditions (BHRCS) (Salum et al., 2015), the Healthy Brain Network (HBN) (Alexander et al., 2017), the Nathan Kline Institute-Rockland Sample (NKI-RS) (Nooner et al., 2012), the developmental component of the Chinese Color Nest Project (devCCNP) (Liu et al., 2020), and the Parents and Children Coming Together project (PACCT; Pls: Tottenham & Milham) (Nikolaidis et al., 2022). Healthy Brain Network is treatment-seeking samples. Philadelphia Neurodevelopmental Cohort, NKI-RS, and devCCNP are community-based samples. BHRCS is a community sample enriched for high family risk for psychopathology. PACCT is a community sample enriched for caregiving-related adversities.

For this study, we have included baseline data from PNC ( $n = 1,601$ , aged 8–22, 52.3% females), BHRCS ( $n = 2,511$ , aged 6–14, 45% females), HBN ( $n = 3,629$ , aged 5–22, 36% females), NKI-RS ( $n = 374$ , aged 6–17, 45% females), devCCNP ( $n = 181$ , aged 6–18, 52% females) and PACCT ( $n = 312$ , aged 6–12, 52% females). The final sample comprised 8606 subjects aged 5–22, 41.0% females. In PNC, psychopathology was assessed using the GOASSESS (described below). All the other samples used CBCL.

### 2.2 | Instruments

#### 2.2.1 | Child Behavior Checklist

The CBCL is a 120-item parent-report assessment of current emotional and behavioral symptoms in participants aged 6–18 over the past 6 months, answered on a 3-point scale (0 = not true, 1 = somewhat/sometimes true, and 2 = very true/often). It encompasses eight syndromes: anxious-depressed, withdrawn-depressed, somatic complaints, rule-breaking behavior, aggressive behavior, social problems, thought problems, and attention problems (Achenbach & Rescorla, 2001). To harmonize with GOASSESS, CBCL scores

1 and 2 were collapsed to generate a binary-scaled variable compatible with GOASSESS (i.e., 0 or 1). In the present study, we extended the CBCL preonized age (6–17) to between 5 and 22 years old based on a previous finding showing age-invariance when CBCL is used to this extended age range (Hoffmann, Moore, Axelrud, Tottenham, Zuo, et al., 2022). Items can be found in Table S1.

#### 2.2.2 | GOASSESS

GOASSESS is a 120-item instrument based on DSM-IV constructs, including symptoms of mood disorders (Major Depressive Episode, Manic Episode), anxiety disorders (Generalized Anxiety Disorder, Separation Anxiety Disorder, Specific Phobia, Social Phobia, Panic Disorder, Agoraphobia, Obsessive-Compulsive Disorder, Post-traumatic Stress Disorder), Attention Deficit/Hyperactivity Disorder (ADHD), behavioral (Oppositional Defiant Disorder, Conduct Disorders) and eating disorders (Anorexia, Bulimia), and suicidal thinking and behavior. Items are scored as 0 (absent) or 1 (currently present) and were parent-reported. The instrument is abbreviated and modified from the epidemiologic version of the NIMH Genetic Epidemiology Research Branch Kiddie-SADS, and its development is described and tested elsewhere (Calkins et al., 2015). Items can be found in Table S1.

### 2.3 | Study design

We aimed to evaluate how to best harmonize different questionnaires in different samples using the bifactor model framework for psychopathology. To reach this general aim, we performed the following five steps. First, we selected bifactor models using the CBCL and GOASSESS. Second, we selected items from these models to harmonize CBCL (HBN, BHRCS, NKI-RS, devCCNP, and PACCT samples) and GOASSESS (PNC sample) using item-wise expert-based semantic harmonization strategy (Hoffmann, Moore, Axelrud, Tottenham, Pan, et al., 2022, submitted). Third, we applied the selected harmonized items for each model and tested model fit and reliability. Fourth, we tested each questionnaire's invariance in the harmonized models. Finally, we tested the authenticity of harmonized scores by estimating the correlation and differences between factor scores from harmonized versus original models (i.e., the same factor structure for the full item set described in the original publications).

Semantic matching was rated by two researchers (MSH and LKA). Any disagreement was decided by a third rater (GAS). We recently tested the performance of 11 CBCL bifactor models, which revealed differences in specific factors' characteristics depending on the items used (Hoffmann, Moore, Axelrud, Tottenham, Zuo, et al., 2022). We selected models that presented reliable specific factors from this previously published work that were also associated with functional impairment. Three models with internalizing and externalizing specific factors (named Achenbach 2S, Deutz GP, and

Clark 2S) and three with internalizing, externalizing, and attention specific factors (named McElroy, Moore 3S, and Clark 3S) were used as CBCL-based harmonized models. CBCL-derived original factor scores (i.e., full item set) were obtained from the same published work (Hoffmann, Moore, Axelrud, Tottenham, Zuo, et al., 2022). One GOASSESS-based bifactor model was published using mixed informants (Kaczurkin et al., 2018, 2019; Shanmugan et al., 2016). This model contains an item-wise GOASSESS-derived p-factor (112 items) and four specific factors, namely anxious-misery (38 items from depression, generalized anxiety, obsessive-compulsive, and panic symptoms), fear (31 items from separation and social anxiety, specific phobia, and agoraphobia), behavioral (23 items from attention/hyperactivity, conduct, and oppositional-defiant symptoms) and psychosis (20 items from manic and psychotic symptoms). Among these items, 36 were harmonized with CBCL in the model Shanmugan 4S (see Table S2). Because the previous GOASSESS bifactor model was published using mixed informants (parents and child reports), we estimated a new bifactor model using the full item set for parent-report only (model named Shanmugan et al., 2016). Therefore, all factor scores used in this study were based on parent-reported symptoms.

## 2.4 | Statistical analysis

### 2.4.1 | Global fit and model-based reliability testing

We applied confirmatory factor analysis (CFA) using delta parameterization and weighted least squares with diagonal weight matrix with standard errors and mean- and variance-adjusted chi-square test statistics (WLSMV) estimators. To evaluate global model fit, we used root mean square error of approximation (RMSEA), comparative fit index (CFI), Tucker–Lewis index (TLI), and standardized root mean square residual (SRMR). Root mean square error of approximation lower than 0.060 and CFI or TLI values higher than 0.950 indicate a good-to-excellent model. SRMR lower than or equal to 0.080 indicate acceptable fit, and lower than 0.060 in combination with previous indices indicates good fit (Hu & Bentler, 1999). The samples were used as clusters to adjust for the non-independence of the standard errors.

We used 10 model-based reliability indices to evaluate the bifactor models described in full in supporting information (page 2). Briefly, they were omega ( $\omega$ ), hierarchical omega ( $\omega_H$ ), factor determinacy (FD), H index, explained common variance (ECV), ECV of a specific factor due to itself (ECV-SS), ECV of a specific factor relative to the general factor (ECV-SG), ECV of the general factor relative to a specific factor (ECV-GS), and percent uncontaminated correlations (PUC) (Dueber, 2017; Rodriguez et al., 2016). The construct can be interpreted as unidimensional when  $\omega_H$  is  $> 0.8$  and ECV and PUC are  $> 0.7$  (Rodriguez et al., 2016). We also examined the percentage of items on specific factors with significant negative ( $\leq -0.3$ ) factor loading (PINFL) and the percentage of items on specific factors with high ( $\geq 0.3$ ) factor loading (PIHFL).

### 2.4.2 | Invariance testing

First, we tested whether the models are structurally similar (configural invariance). After that, we tested whether items were informing symptoms at an equivalent level and equally correlated with each questionnaire's latent factors (scalar invariance). Invariance was tested with multigroup CFA (MG-CFA) for ordinal data, establishing group equality in model configuration, thresholds, and loadings using the option "model = configural scalar" in Mplus. Invariance is established by comparing global model fit indices between constrained models (B. Muthén & Asparouhov, 2002; Ploubidis et al., 2019).  $\Delta CFI < 0.01$  supplemented by  $\Delta RMSEA < 0.015$  or  $\Delta SRMR < 0.010$  between models with increasing levels of constraints indicate invariance (configural vs. scalar) (Chen, 2007).

### 2.4.3 | Authenticity

Model authenticity was examined in two ways. First, we estimated the Pearson correlation of factor scores among harmonized versus original full item set from CBCL and GOASSESS models, applied to the study samples. This estimates how close a harmonized model coheres to the original models, providing a concise summary of the reproducibility and validity of the harmonized models. Second, we calculated the differences between harmonized and full-item set bifactor scores using Bland-Altman plots. We also calculated the proportion of subjects with factor score differences between harmonized and original models higher than 0.5 to indicate prediction error. Finally, as a sensitivity analysis, we estimated the same harmonized models for samples separated by instrument (i.e., using GOASSESS or CBCL only). We correlated these models with the full item set. This estimates how close a harmonized model is to the original models without the interference of other samples.

All CFAs were carried out using Mplus version 8.6 (Muthén & Muthén, 2017) and implemented in R version 4.0.3 using the *MplusAutomation* package (Hallquist & Wiley, 2018), which was also used to extract factor scores generated in Mplus. All bifactor reliability indices were calculated using the *BifactorIndicesCalculator* package in R (Dueber, 2017). Invariance was tested with multigroup CFA (MG-CFA) using the option "model = configural scalar" in Mplus. Authenticity was assessed using Pearson correlation, estimated, and plotted using the *rcorr* function in the *Hmisc* package (Harrell, 2021). Bland-Altman plots were generated to demonstrate the average factor score of the harmonized model relative to the measurement difference between harmonized and full-item set models. We conducted a supplementary analysis to understand if harmonized items were consistently associated with demographic variables across studies. For that, we estimate the tetrachoric correlation between harmonized items with age (below age 11 and higher or equal than 11) and sex (male as reference) for studies with these demographic variables (all but PACCT). Tetrachoric correlations were calculated using the *tetcor* function in the *fungible* R package (Waller et al., 2022). Code and supporting tables can be found at <https://osf.io/bg7zq/>.

TABLE 1 CBCL and GOASSESS harmonized items

CBCL		GOASSESS		Harmonized code
Item	Content	Item	Content	
CBCL_4	Fails to finish things he/she starts	ADD_012	Have problems following instructions or finishing things you meant to get done	att_1
CBCL_8	Can't concentrate, can't pay attention for long	ADD_011	Have trouble paying attention on activities that you were doing	att_2
CBCL_10	Can't sit still, restless or hyperactive	ADD_020	Have difficulty sitting still for more than a few minutes at a time	att_3
CBCL_13	Confused or seems to be in a fog	ADD_016	Been told that you did not seem to be listening when they spoke to you	att_4
CBCL_61	Poor school work	ADD_014	Make careless mistakes in school work or other activities	att_5
CBCL_5	There is very little he/she enjoys	DEP_006	Nothing was fun for you and you just weren't interested in anything	int_1
CBCL_9	Can't get his/her mind off certain thoughts; obsessions	OCD_006	Bothered by thoughts such as forbidden/bad thoughts	int_2
CBCL_11	Clings to adults or too dependent	SEP_508	Wanted to stay home or not leave without your attachment figures	int_3
CBCL_14	Cries a lot	DEP_002	Cried a lot or felt like crying	int_4
CBCL_29	Fears certain animals, situations, or places, other than school	PHB_008	Afraid of any other things or situations	int_5
CBCL_31	Fears he/she might think or do something bad	OCD_004	Bothered by fear that you would do/say something bad without intending to	int_6
CBCL_50	Too fearful or anxious	GAD_002	Worry a lot more than most people your age	int_7
CBCL_66	Repeats certain acts over and over; compulsions	OCD_016	Repetitive ordering or arranging things	int_8
CBCL_71	Self-conscious or easily embarrassed	SOC_005	Afraid being the center of attention and were concerned something embarrassing might happen	int_9
CBCL_75	Too shy or timid	SOC_001	Really shy with people meeting new people going to parties or doing things in front of others	int_10
CBCL_83	Stores up too many things he/she doesn't need	OCD_018	You saved up so many things that they got in the way	int_11
CBCL_91	Talks about killing self	SUI_002	You have thought about killing yourself	int_12
CBCL_103	Unhappy, sad, or depressed	DEP_001	You felt sad or depressed most of the time	int_13
CBCL_112	Worries	GAD_001	Have been a worrier	int_14
CBCL_15	Cruel to animals	CDD_006	Been physically cruel to an animal or person	ext_1
CBCL_16	Cruelty, bullying, or meanness to others	CDD_005	Often bully others	ext_2
CBCL_28	Breaks rules at home, school, or elsewhere	ODD_002	Breaking rules at home/school	ext_3
CBCL_43	Lying or cheating	CDD_001	Got into trouble with adults like lying or stealing	ext_4
CBCL_57	Physically attacks people	CDD_007	Try to hurt someone with a weapon	ext_5
CBCL_72	Sets fires	CDD_003	Set fires break into cars or destroy someone else's property on purpose	ext_6
CBCL_86	Stubborn, sullen, or irritable	ODD_006	Irritable or grouchy or get angry because you thought that things were unfair	ext_7
CBCL_87	Sudden changes in mood or feelings	MAN_007	You felt unusually grouchy cranky or irritable	ext_8
CBCL_88	Sulks a lot	DEP_004	Felt grouchy irritable or in a bad mood most of the time	ext_9
CBCL_95	Temper tantrums or hot temper	ODD_001	Losing temper arguing with adults or being grouchy or irritable with them	ext_10

(Continues)

TABLE 1 (Continued)

CBCL		GOASSESS		Harmonized code
Item	Content	Item	Content	
CBCL_97	Threatens people	CDD_008	Threaten someone	ext_11
CBCL_101	Truancy, skips school	CDD_002	Skip school stay out later than you were supposed to or run away from home overnight	ext_12
CBCL_34	Feels others are out to get him/her	PSY_071	Believed people being out to get you or controlling what you do or think	psy_1
CBCL_40	Hears sounds or voices that aren't there	SIP_012	Heard sounds of people talking when there is no one near me.	psy_2
CBCL_70	Sees things that aren't there	PSY_029	Have seen visions or seen things which other people could not see	psy_3
CBCL_85	Strange ideas	PSY_070	Believed in things that most other people don't believe in	psy_4
CBCL_93	Talks too much	MAN_004	Racing thoughts or pressured speech	psy_5

Note: 1G4S model is the model in which most items from CBCL and GOASSESS are compatible and have psychosis specific factor added to the previous specific factors. Achenbach 2S, Deutz GP and Clark 2S included internalizing and externalizing specific factors. Moore 3S and Clark 3S have attention/hyperactivity in addition to the previous specific factors. Shanmugan 4S contained 36 items which were possible to harmonized with CBCL, and included anxious-misery, fear, behavior and psychosis specific factors.

Abbreviation: CBCL, Child Behavior Checklist.

### 3 | RESULTS

The semantic harmonization between CBCL and GOASSESS items resulted in 91.6% agreement (33/36 items, Table 1) between the initial raters. Table 2 describes the configuration of the harmonized bifactor models. Supplementary tetrachoric correlation between harmonized items with age and sex for each study (Figure S1) demonstrates that there are no major differences in item-demographics correlations between studies that used CBCL and GOASSESS (PNC study).

#### 3.1 | Model fit and reliability for item-wise harmonized bifactor models

Model fits indices for the seven harmonized bifactor models are presented in Table 3. Factor loadings and factor reliability of each model can be found in supporting tables (Table S2–S8). All CBCL-GOASSESS harmonized models fitted the data well, except Clark 2S, which presented CFI and TLI indices below 0.950.

For each factor, the harmonized models used 29.0%–50.0% of the original item set in the CBCL-based models and 16.0%–62.5% for the GOASSESS-based model (Table 4). Reliability indices demonstrated that the harmonized models were multidimensional ( $p$ -factor  $\omega_H < 0.8$  and ECV-GS and PUC  $< 0.7$ ), except for the Clark 2 and 3S models (the only S-1 models), which presented poor reliability for the specific factors (Table S2–S8). The behavior and psychosis specific factors from the Shanmugan 4S model presented good FD and H-index (Table 4), indicating that the factor scores could be used in the analysis and are well-defined latent variables. Factor determinacy and H index were borderline acceptable or poor for all remaining specific factors in all models. Moreover, the average percentage of items with factor

loading  $\geq 0.3$  on specific factors was 57.1% (Table 4), corroborating with the information that specific factors from CBCL-GOASSESS harmonized bifactors models are poorly defined.

The original PNC's bifactor model (Shanmugan et al., 2016) presented a good fit (RMSEA = 0.019; 90%, CI = 0.18–0.20; CFI = 0.935; TLI = 0.933; SRMR = 0.101) and results can be found in Table S9 (note that in the present study, we used parent-reports only instead of self-reports for adolescents).

#### 3.2 | Instrument measurement invariance

Sample invariance testing demonstrated that Achenbach 2S and Deutz GP models did not reach scalar invariance; therefore, instruments in these models may be a source of variance and can explain mean differences between scores while using CBCL or GOASSESS (Figure 1). Accordingly, these models were not further used for the factor correlation analysis. Values for invariance testing can be found in Table S10.

#### 3.3 | Authenticity

We found that the correlation of harmonized scores with full item set scores was similar for the Clark 2S, McElroy, Moore 3S, and Clark 3S models (Figure 2a to 2d). These correlations varied from 0.95 to 0.96 for the  $p$ -factors of these models and from 0.65 to 0.81 for the specific factors. However, the Shanmugan 4S model presented a wider range of correlations between factors derived from the full item set model (Shanmugan et al., 2016) (Figure 2e and Table 4).

The proportion of subjects with a factor score difference between harmonized and original models higher than 0.5 varied from

TABLE 2 Configuration of the harmonized CBCL and GOASSESS bifactor models

Harmonized code	Achenbach 2S	Deutz GP	Clark 2S	McElroy	Moore 3S	Clark 3S	Shanmugan 4S
att_1			P-factor only		Attention	Attention	Behavior
att_2			P-factor only	Attention	Attention	Attention	Behavior
att_3			P-factor only	Attention	Attention	Attention	Behavior
att_4			P-factor only	Attention	Attention	Attention	Behavior
att_5			P-factor only	Attention	Attention	Attention	Behavior
int_1			P-factor only		Internalizing	Internalizing	Anxious-misery
int_2			P-factor only		Attention	Attention	Anxious-misery
int_3			P-factor only			Internalizing	Fear
int_4	Internalizing	Internalizing	Internalizing	Internalizing		Internalizing	Anxious-misery
int_5	Internalizing	Internalizing	P-factor only	Internalizing		Internalizing	Fear
int_6	Internalizing	Internalizing	Internalizing	Internalizing	Internalizing	Internalizing	Anxious-misery
int_7	Internalizing	Internalizing	Internalizing	Internalizing	Internalizing	Internalizing	Anxious-misery
int_8			P-factor only		Attention	Attention	Anxious-misery
int_9	Internalizing	Internalizing	Internalizing	Internalizing	Internalizing	Internalizing	Fear
int_10	Internalizing	Internalizing	Internalizing	Internalizing	Internalizing	Internalizing	Fear
int_11			P-factor only			P-factor only	Anxious-misery
int_12	Internalizing	Internalizing	P-factor only	Internalizing		Internalizing	Anxious-misery
int_13	Internalizing	Internalizing	Internalizing	Internalizing	Internalizing	Internalizing	Anxious-misery
int_14	Internalizing	Internalizing	Internalizing	Internalizing	Internalizing	Internalizing	Anxious-misery
ext_1			P-factor only		Externalizing	Externalizing	Behavior
ext_2	Externalizing	Externalizing	Externalizing	Externalizing	Externalizing	Externalizing	Behavior
ext_3	Externalizing	Externalizing	P-factor only		Externalizing	Externalizing	Behavior
ext_4	Externalizing	Externalizing	Externalizing	Externalizing	Externalizing	Externalizing	Behavior
ext_5	Externalizing	Externalizing	Externalizing	Externalizing	Externalizing	Externalizing	Behavior
ext_6	Externalizing	Externalizing	Externalizing	Externalizing	Externalizing	Externalizing	Behavior
ext_7	Externalizing	Externalizing	Externalizing	Externalizing	Externalizing	Externalizing	Behavior
ext_8	Externalizing	Externalizing	Externalizing	Externalizing	Externalizing	Internalizing	Anxious-misery
ext_9	Externalizing	Externalizing	Internalizing	Externalizing	Externalizing	Externalizing	Anxious-misery
ext_10	Externalizing	Externalizing	Externalizing	Externalizing	Externalizing	Externalizing	Behavior
ext_11	Externalizing	Externalizing	Externalizing	Externalizing	Externalizing	Externalizing	Behavior
ext_12	Externalizing	Externalizing	Externalizing			Externalizing	Behavior
psy_1			Internalizing		Externalizing	Internalizing	Psychosis
psy_2		P-factor only	P-factor only			P-factor only	Psychosis
psy_3		P-factor only	P-factor only			Internalizing	Psychosis
psy_4		P-factor only	P-factor only		Attention	Attention	Psychosis
psy_5			Internalizing		Attention	Attention	Psychosis

Note: 1G4S model is the model in which most items from CBCL and GOASSESS are compatible and have psychosis specific factor added to the previous specific factors. Achenbach 2S, Deutz GP and Clark 2S included internalizing and externalizing specific factors. Moore 3S and Clark 3S have attention/hyperactivity in addition to the previous specific factors. Shanmugan 4S contained 36 items which were possible to harmonized with CBCL, and included anxious-misery, fear, behavior and psychosis specific factors.

Abbreviation: CBCL, Child Behavior Checklist.

**TABLE 3** CBCL and GOASSESS harmonized bifactor model fit indices

Model	RMSEA	RMSEA 90% CI	CFI	TLI	SRMR
Achenbach 2S	0.009	0.007 0.011	0.985	0.980	0.059
Deutz GP	0.009	0.007 0.011	0.978	0.974	0.068
Clark 2S	0.009	0.008 0.010	0.947	0.942	0.084
McElroy	0.011	0.009 0.012	0.972	0.965	0.063
Moore 3S	0.009	0.008 0.011	0.968	0.962	0.065
Clark 3S	0.008	0.007 0.009	0.961	0.956	0.072
Shanmugan 4S	0.008	0.007 0.009	0.961	0.956	0.073

Note: Achenbach 2S, Deutz GP and Clark 2S included internalizing and externalizing specific factors. McElroy, Moore 3S and Clark 3S have attention/hyperactivity in addition to the previous specific factors. Shanmugan 4S contained 36 items which were harmonized with CBCL, and included anxious-misery, fear, behavior and psychosis specific factors.

Abbreviations: CBCL, Child Behavior Checklist; CFI, Comparative Fit Index; RMSEA, Root Mean Square Error of Approximation; SRMR, Standardized Root Mean-square Residual; TLI, Tucker-Lewis Index.

6.3% to 13.7% in the CBCL-based bifactor models, and it was 50.9% for the Shanmugan 4S model. For the specific factors, this proportion varied from 23.4% to 47.1% (Table 4).

Table 4 summarizes the key model features and results, including multiple indicators of model fit and reliability and results from the authenticity assessment. Based on these results, Achenbach 2S and Deutz GP were non-invariant, and the Clark 2 and 3S demonstrated poor reliability for the specific factors described above. In addition, the Shanmugan 4S presented two specific factors with poor authenticity (anxious-misery and psychosis). Therefore, the McElroy and Moore 3S model resulted in the best models to harmonize CBCL and GOASSESS, with differences concerning the reliability and authenticity of attention and externalizing specific factors.

### 3.4 | Sensitivity analysis—Estimating models in separated samples

Previous models were estimated by combining all six samples. The low correlation between harmonized (Shanmugan 4S) and original Shanmugan et al. (2016) models might be due to the instrument imbalance among samples in the harmonized model. Given sample size for CBCL is larger than GOASSESS, we also have conducted a sensitivity analysis using only the sample that derived the model. When bifactor models were estimated in separated samples depending on the questionnaire used, they fitted the data well (Table S11). The proportion of subjects with p-factor score differences higher than 0.5 between harmonized versus original Shanmugan model decreased from 50.9% of the sample to 22.2% using the PNC sample. The psychosis specific factor from the Shanmugan 4S (harmonized) model still presented a low correlation ( $r = 0.31$ ) with the original psychosis factor from the Shanmugan model (2016)

(Table S12). However, the correlation between anxious-misery raised scores o from 0.12 to 0.87 when the bifactor models were restricted to the PNC sample only (Table S12). Restricting the analysis to the CBCL samples only did not present different results, as mentioned above (Table S12). Bifactor models for sensitivity analysis are described in full for the Clark 2S (Tables S13 and S14), McElroy (Tables S15 and S16), Moore 3S (Tables S17 and S18), Clark 3S (Tables S19 and S20), and Shanmugan S4 (Tables S21 and S22). Correlation and factor score differences between harmonized versus full item set restricting the samples for CBCL or GOASSESS only are described in Table S12.

## 4 | DISCUSSION

Our study provides a framework for testing bifactor models of psychopathology while harmonizing two instruments using three main aspects: model fit and reliability, instrument measurement invariance, and authenticity. We demonstrated that fewer than 50% of theoretical and empirically driven bifactor models from the literature using CBCL and GOASSESS (Constantinou & Fonagy, 2019; Hoffmann, Moore, Axelrud, Tottenham, Zuo, et al., 2022; Shanmugan et al., 2016) presented reliable specific factors and are instrument-invariant. Moreover, item-wise harmonized models containing around a third of the original item set generated a highly correlated p-factor between harmonized and original full-item models.

Enormous effort has been made to move psychiatry and psychopathology to the next level of nosology (Insel et al., 2010; Kotov et al., 2021; Lahey et al., 2022). Consortia gather different samples to understand psychiatric conditions' genetics and neuroimaging underpinnings. In that effort, mental health is operationalized in different ways, from symptom checklists to clinical categorical diagnoses. However, harmonizing these phenotypic assessments is frequently overlooked, and it is unknown to what extent symptoms and diagnoses assessed with different questionnaires using different samples can be combined. For example, the PGC assumes that all diagnoses are equivalent, and the ENIGMA consortium uses diagnosis and symptom-level analysis (Sullivan et al., 2018; Thompson et al., 2014). In the present study, we demonstrate that not all bifactor models can be applied to harmonize different questionnaires in different samples, as there is a risk of non-authentic factors. However, some factors in these models might be useful as they are authentic dimensions of psychopathology, such as the fear factor in the Shanmugan 4S model.

The Shanmugan 2016 model (Shanmugan et al., 2016) was generated within the PNC study (Satterthwaite et al., 2016) and presented a good bifactor solution for this study. However, the harmonized version of this model (Shanmugan 4S) generated factors with lower correlation when compared with the full item set model. This might be due to the patterns of factor loadings. For example, the psychosis specific factor loaded highly onto hallucination-related symptoms in the harmonized Shanmaugan 4S model. In contrast, the original published model (Shanmugan et al., 2016) loaded more highly onto delusion-related symptoms that were not harmonized.



TABLE 4 Result summary

Models	Model fit			RBC Model-based reliability							Harmonized and original models				
	RMSEA	CFI	SRMR	Factors	N° of original indicators	N° of harmonized indicators	Harmonized/original indicators	wH	H index	FD	PINFL	PIHFL	Questionnaire invariance	Correlation	Proportion of subjects with factor score difference module higher than 0.5
Achenbach 2S	0.009	0.985	0.059	P-factor	68	20	29.4%	0.717	0.936	0.956	0.0%	90.0%	Non-invariant		
				Internalizing	33	9	27.3%	0.251	0.625	0.810	0.0%	66.7%			
				Externalizing	35	11	31.4%	0.233	0.756	0.888	0.0%	63.6%			
Deutz GP	0.009	0.978	0.068	P-factor	71	23	32.4%	0.730	0.936	0.955	0.0%	95.7%	Non-invariant		
				Internalizing	31	9	29.0%	0.203	0.575	0.782	0.0%	55.6%			
				Externalizing	26	11	42.3%	0.262	0.772	0.890	0.0%	72.7%			
Clark 2S	0.009	0.947	0.084	P-factor	116	36	31.0%	0.870	0.959	0.975	0.0%	100.0%	Invariant	0.96	9.1%
				Internalizing	31	10	32.3%	0.188	0.616	0.821	0.0%	50.0%		0.69	35.9%
				Externalizing	29	9	31.0%	0.139	0.610	0.858	0.0%	44.4%		0.67	34.0%
McElroy	0.011	0.972	0.063	P-factor	66	22	33.3%	0.776	0.945	0.968	0.0%	100.0%	Invariant	0.96	6.7%
				Attention	8	4	50.0%	0.234	0.541	0.833	0.0%	75.0%		0.81	23.4%
				Internalizing	31	9	29.0%	0.309	0.712	0.869	0.0%	77.8%		0.69	36.0%
				Externalizing	27	9	33.3%	0.103	0.619	0.871	0.0%	44.4%		0.73	29.1%
Moore 3S	0.009	0.968	0.065	P-factor	75	28	37.3%	0.837	0.954	0.969	0.0%	100.0%	Invariant	0.96	6.3%
				Attention	18	9	50.0%	0.063	0.558	0.886	0.0%	44.4%		0.72	31.8%
				Internalizing	23	7	30.4%	0.226	0.602	0.809	0.0%	71.4%		0.67	34.7%
				Externalizing	34	12	35.3%	0.224	0.744	0.892	0.0%	66.7%		0.77	26.9%
Clark 3S	0.008	0.961	0.072	P-factor	116	36	31.0%	0.835	0.958	0.969	0.0%	100.0%	Invariant	0.95	13.7%
				Attention	25	9	36.0%	0.112	0.586	0.871	0.0%	44.4%		0.74	32.5%
				Internalizing	43	14	32.6%	0.166	0.654	0.822	0.0%	57.1%		0.65	39.0%
				Externalizing	32	11	34.4%	0.244	0.714	0.881	0.0%	63.6%		0.77	26.2%

(Continues)

TABLE 4 (Continued)

Models	Model fit				RBC Model-based reliability						Harmonized and original models				
	RMSEA	CFI	SRMR	Factors	N° of original indicators	N° of harmonized indicators	Harmonized/original indicators	$\omega$ H	H index	FD	PINFL	PIHFL	Questionnaire invariance	Correlation	Proportion of subjects with factor score difference module higher than 0.5
Shanmugan 4S	0.008	0.961	0.073	P-factor	104	36	34.6%	0.837	0.957	0.973	0.0%	97.2%	Invariant	0.89	50.9%
				Anxious-Misery	33	12	36.4%	0.000	0.452	0.794	25.0%	8.3%		0.12	47.1%
				Fear	25	4	16.0%	0.205	0.639	0.857	0.0%	50.0%		0.74	27.1%
				Behavior	24	15	62.5%	0.288	0.795	0.907	0.0%	80.0%		0.74	32.2%
				Psychosis	24	5	20.8%	0.194	0.711	0.911	0.0%	40.0%		0.39	25.6%

Note: Non-harmonized models were factor models containing the full set of items as recommended by the original models. Correlation with non-harmonized models was measured using Pearson correlations among factor scores.  $\omega$ H, omega-hierarchical; H index, index of construct replicability (>0.8 suggests a well-defined latent variable); FD, factor determinacy (>0.9 indicates that the factor score can be used). PINFL, percentage of items with negative factor loading (<=-0.3); PIHFL, percentage of items with high ( $\geq 0.3$ ) factor loading.

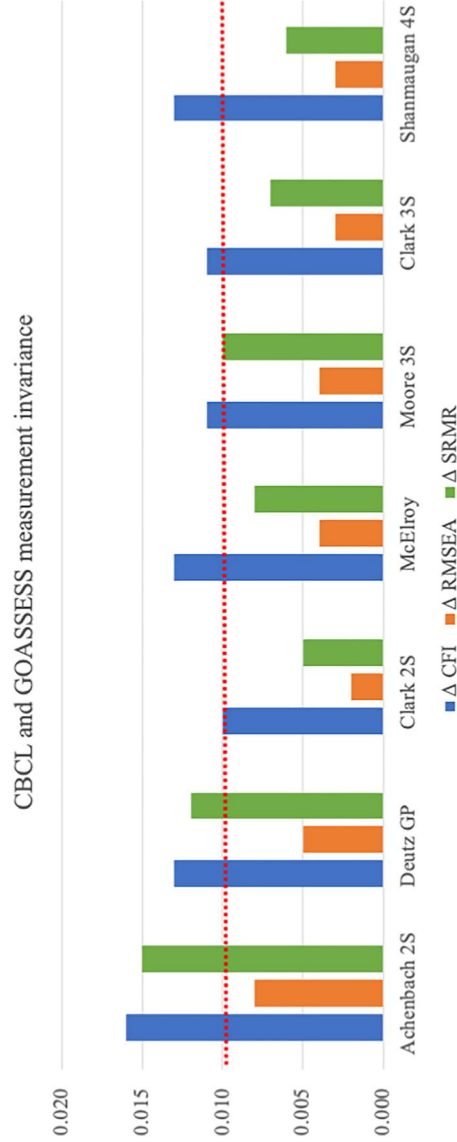
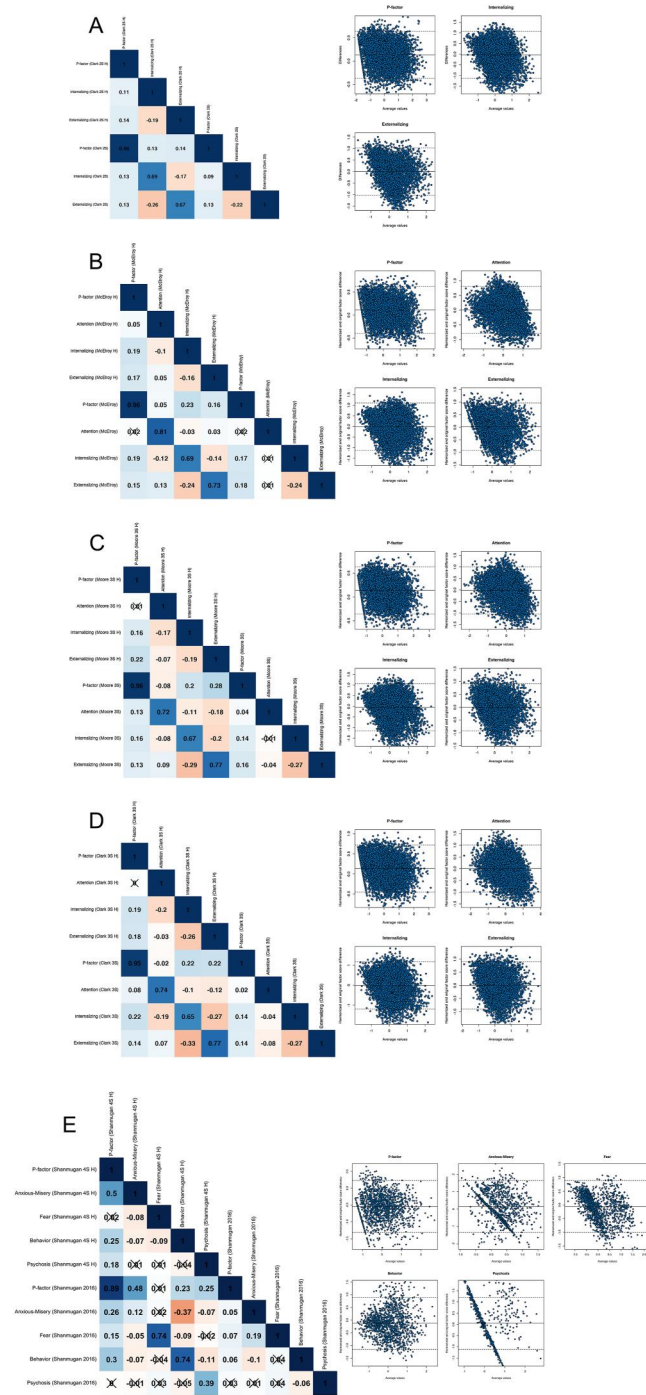


FIGURE 1 Y-axis represents the global model fit indices differences between configural and scalar models. X-axis depicts the bifactor models, namely Achenbach 2S, Deutz GP, Clark 2S, McElroy, Moore 3S, Clark 3S, and Shanmugan 4S. CBCL, Child Behavior Checklist; RMSEA, Root Mean Square Error of Approximation; CFI, Comparative Fit Index; TLI, Tucker-Lewis Index; SRMR, Standardized Root Mean-square Residual;  $\Delta$ , differences between fit indices.



**FIGURE 2** Correlation and Bland-Altman plots for Clark 2s (a), McElroy (b), Moore 3S (c), Clark 3S (d) and Shanmugan 4S (e) bifactor models. In the Bland-Altman plots (plots in the right), the x-axis demonstrates the average factor score of the harmonized and full item set models and the y-axis displays the difference in measurements between them.

Moreover, the p-factor from the harmonized Shanmugan 4S model loaded highly on irritability-related items, a pattern commonly observed in CBCL bifactor models (Hoffmann, Moore, Axelrud, Tottenham, Zuo, et al., 2022). The p-factor loading pattern of the PNC's

published model includes high loadings for mania and obsessive-compulsive symptoms rather than irritability, which might explain the correlation among p-factors of 0.89; while still high, this correlation is below what was observed using the CBCL models.

Five CBCL-based models presented authentic dimensions when compared with the full item set. Even when applied to the PNC sample (which used a different instrument), it fits the data well (Table S11) and produced factors with acceptable model-based reliability indices (see S13, S15, S17, S19, and S21). These models are generally based on internalizing and externalizing specific factors and, in some, an additional attention factor (Achenbach, 1966; Hoffmann, Moore, Axelrud, Tottenham, Zuo, et al., 2022). Attention specific factors are usually neglected by versions of the Hierarchical Taxonomy of Psychopathology (Kotov et al., 2021; Krueger et al., 2021). However, previous and recent evidence supports the role of this dimension (Barkley, 2015; Gomez et al., 2021), and previous studies using CBCL demonstrated its validity (Clark et al., 2021; McElroy et al., 2018; Moore et al., 2020). In a bifactor context, the attention factor might capture aspects of impulsivity overlapping with the p-factor, while the residual specific factor is possibly related to the attention-hyperactivity phenotype (Gomez et al., 2021). Our findings emphasize that models that include internalizing, externalizing, and attention specific dimensions are reproducible and likely to generate harmonizable dimensions across studies and questionnaires.

Several limitations of this study should be acknowledged. First, this study considered two parent-report questionnaires, and the estimated differences among constructs may not apply if other informants and questionnaires are used. Second, the previously published GOASSESS model used self-reports for subjects beyond age 10. Here we estimated it using parent reports to keep it consistent with CBCL. Nonetheless, the analytical pipeline of this study can be applied to analyze differences in other instruments and informants. Third, the GOASSESS scale measures symptoms using a binary response, while the CBCL is a three-level Likert-type scale. This has the potential to introduce some bias. We minimized that by using a scale-level harmonization that has been used in other studies in the field, which involves collapsing the CBCL to produce a binary variable.

This study provides a framework to evaluate bifactor models using harmonized data from different questionnaires and different samples. Harmonized CBCL-GOASSESS bifactor models contained a third of the original item set in general. Most but not all models with harmonized data resulted in highly authentic p-factors, while the authenticity for the specific factors varied. CBCL-based bifactor models (i.e., internalizing, externalizing, and attention) performed better relative to the GOASSESS-based bifactor model (i.e., anxious-misery, fear, behavior, and psychosis factors) when their items were harmonized. This approach could be expanded to dimensions based on diagnosis, other instruments, and diverse informants. Future studies are advised to estimate how psychopathology dimensions in aggregated datasets are harmonizable and represent the same dimensions of psychopathology.

## AUTHOR CONTRIBUTIONS

**Mauricio Scopel Hoffmann:** Writing the original draft; Conceptualization; Methodology; Analysis; Visualization. **Tyler Maxwell Moore:** Conceptualisation; Writing—review & editing. **Luiza Kvitko Axelrud:** Conceptualisation; Writing—review & editing. **Nim Tottenham:** Funding acquisition; Writing—review & editing; Luis Augusto Rohde: Funding acquisition; Writing—review & editing. **Michael Peter Milham:** Funding acquisition; Conceptualisation; Writing—review & editing. **Theodore Daniel Satterthwaite:** Funding acquisition; Conceptualisation; Writing—review & editing. **Giovanni Abrahão Salum:** Funding acquisition; Conceptualisation; Methodology; Writing—review & editing; Supervision.

## ACKNOWLEDGMENTS

This work was, and the RBC initiative is supported by the United States National Institutes of Health grant R01MH120482. BHRCs was supported by grants from the National Institute of Development Psychiatric for Children and Adolescent (INPD) (Grants: CNPq 465550/2014-2, FAPESP 2014/50917-0 and European Research Council (FP7/2007–2013)/grant agreement n° 337673, The Medical Research Council-UK). Chinese Color Nest Project received funding support from the Child Brain-Mind Development Cohort Study in China Brain Initiative (SQ2021AAA010024), the National Basic Science Data Center “Chinese Data-sharing Warehouse for In-vivo Imaging Brain” (NBSDC-DB-15), the Major Project of National Social Science Foundation of China (20&ZD296), the Beijing Municipal Science and Technology Commission (Z161100002616023, Z181100001518003), the Key-Area Research and Development Program of Guangdong Province (2019B030335001) and the Startup Funds for the Leading Talents at Beijing Normal University. Additional support was provided by the Penn-CHOP Lifespan Brain Institute. Analysis code and supporting information can be accessed at <https://osf.io/uwy5n/>.

## CONFLICT OF INTEREST

Luis Augusto Rohde has received grant or research support from, served as a consultant to, and served on the speakers' bureau of Aché, Bial, Medice, Novartis/Sandoz, Pfizer/Upjohn, and Shire/Takeda in the last 3 years. The ADHD and Juvenile Bipolar Disorder Outpatient Programs chaired by Dr Rohde have received unrestricted educational and research support from the following pharmaceutical companies in the last 3 years: Novartis/Sandoz and Shire/Takeda. Dr Rohde has received authorship royalties from Oxford Press and ArtMed. Other authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author at the moment. As one of the Reproducible Imaging-based Brain Growth Charts for Psychiatry's main goal (NIMH number R01MH120482), the full data set will be freely available at the end of the project. Thus, the data are not publicly available yet due to ethical restrictions (i.e., deidentification

procedures are being carried out). Analysis code and project details are available at <https://osf.io/bg7zq/>

## ORCID

Mauricio Scopel Hoffmann  <https://orcid.org/0000-0003-4232-3169>

## REFERENCES

- Achenbach, T. M. (1966). The classification of children's psychiatric symptoms: A factor-analytic study. *Psychological Monographs*, 80(7), 1–37. <https://doi.org/10.1037/h0093906>
- Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles*. University of Vermont, Research Center for Children, Youth & Families: Library of Congress.
- Alexander, L. M., Escalera, J., Ai, L., Andreotti, C., Febre, K., Mangone, A., Vega-Potler, N., Langer, N., Alexander, A., Kovacs, M., Litke, S., O'Hagan, B., Andersen, J., Bronstein, B., Bui, A., Bushey, M., Butler, H., Castagna, V., Camacho, N., & Milham, M. P. (2017). An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific Data*, 4(1), 170181. <https://doi.org/10.1038/sdata.2017.181>
- Allegrini, A. G., Cheesman, R., Rimfeld, K., Selzam, S., Pingault, J.-B., Eley, T. C., & Plomin, R. (2020). The p factor: Genetic analyses support a general dimension of psychopathology in childhood and adolescence. *Journal of Child Psychology and Psychiatry*, 61(1), 30–39. <https://doi.org/10.1111/jcpp.13113>
- Barkley, R. A. (2015). History of ADHD. In *Attention-deficit hyperactivity disorder: A handbook for diagnosis and treatment* (4th ed., pp. 3–50). The Guilford Press.
- Bornoalova, M. A., Choate, A. M., Fatimah, H., Petersen, K. J., & Wiernik, B. M. (2020). Appropriate use of bifactor analysis in psychopathology research: Appreciating benefits and limitations. *Biological Psychiatry*, 88(1), 18–27. <https://doi.org/10.1016/j.biopsych.2020.01.013>
- Calkins, M. E., Merikangas, K. R., Moore, T. M., Burstein, M., Behr, M. A., Satterthwaite, T. D., Ruparel, K., Wolf, D. H., Roalf, D. R., Mentch, F. D., Qiu, H., Chiavacci, R., Connolly, J. J., Sleiman, P. M., Hakonarson, H., & Gur, R. E. (2015). The Philadelphia neurodevelopmental cohort: Constructing a deep phenotyping collaborative. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 56(12), 1356–1369. <https://doi.org/10.1111/jcpp.12416>
- Caspi, A., Houts, R. M., Ambler, A., Danese, A., Elliott, M. L., Hariri, A., Harrington, H., Hogan, S., Poulton, R., Ramrakha, S., Rasmussen, L. J. H., Reuben, A., Richmond-Rakerd, L., Sugden, K., Wertz, J., Williams, B. S., & Moffitt, T. E. (2020). Longitudinal assessment of mental health disorders and comorbidities across 4 decades among participants in the dunedin birth cohort study. *JAMA Network Open*, 3(4), e203221. <https://doi.org/10.1001/jamanetworkopen.2020.3221>
- Caspi, A., & Moffitt, T. E. (2018). All for one and one for all: Mental disorders in one dimension. *American Journal of Psychiatry*, 175(9), 17121383–17121844. <https://doi.org/10.1176/appi.ajp.2018.17121383>
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464–504. <https://doi.org/10.1080/10705510701301834>
- Clark, D. A., Hicks, B. M., Angstadt, M., Rutherford, S., Taxali, A., Hyde, L., Weigard, A. S., Heitzeg, M. M., & Sripada, C. (2021). The general factor of psychopathology in the adolescent brain cognitive development (ABCD) study: A comparison of alternative modeling approaches. *Clinical Psychological Science*, 2167702620959317(2), 169–182. <https://doi.org/10.1177/2167702620959317>
- Constantinou, M., & Fonagy, P. (2019). Evaluating bifactor models of psychopathology using model-based reliability indices. *PsyArXiv*. <https://doi.org/10.31234/osf.io/6tf7j>

- Dueber, D. (2017). Bifactor indices calculator: A microsoft excel-based tool to calculate various indices relevant to bifactor CFA models. *Educational, School, and Counseling Psychology Research Tools*. <https://doi.org/10.13023/edp.tool.01>
- Fernández de la Cruz, L., Vidal-Ribas, P., Zahreddine, N., Mathiassen, B., Brøndbo, P. H., Simonoff, E., Goodman, R., & Stringaris, A. (2018). Should clinicians split or lump psychiatric symptoms? The structure of psychopathology in two large pediatric clinical samples from England and Norway. *Child Psychiatry and Human Development*, 49(4), 607–620. <https://doi.org/10.1007/s10578-017-0777-1>
- Gomez, R., Liu, L., Krueger, R., Stavropoulos, V., Downs, J., Preece, D., Houghton, S., & Chen, W. (2021). Unraveling the optimum latent structure of attention-deficit/hyperactivity disorder: Evidence supporting ICD and HiTOP frameworks. *Frontiers in Psychiatry*, 12, 666326. <https://doi.org/10.3389/fpsy.2021.666326>
- Gondek, D., Bann, D., Patalay, P., Goodman, A., McElroy, E., Richards, M., & Ploubidis, G. B. (2021). Psychological distress from early adulthood to early old age: Evidence from the 1946, 1958 and 1970 British birth cohorts. *Psychological Medicine*, 52(8), 1–10. <https://doi.org/10.1017/S003329172000327X>
- Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(4), 621–638. <https://doi.org/10.1080/10705511.2017.1402334>
- Harrell, F. E. (2021). Hmisc: Harrell miscellaneous package. Retrieved from <https://CRAN.R-project.org/package=Hmisc>
- Heinrich, M., Zagorscak, P., Eid, M., & Knaevelsrud, C. (2020). Giving G a meaning: An application of the bifactor-(S-1) approach to realize a more symptom-oriented modeling of the beck depression inventory-II. *Assessment*, 27(7), 1429–1447. <https://doi.org/10.1177/1073191118803738>
- Hoffmann, M. S., Moore, T. M., Axelrud, L. K., Tottenham, N., Pan, P. M., Miguel, E. C., Rohde, L. A., Milham, M. P., Satterthwaite, T. D., & Salum, G. A. (2022). An evaluation of item matching strategies to harmonize assessment tools for psychopathology in children and adolescents. (Submitted to the Assessment, in Second Round of Review).
- Hoffmann, M. S., Moore, T. M., Axelrud, L. K., Tottenham, N., Zuo, X.-N., Rohde, L. A., Milham, M. P., Satterthwaite, T. D., & Salum, G. A. (2022). Reliability and validity of bifactor models of dimensional psychopathology in youth. *Journal of Psychopathology and Clinical Science*, 131(4), 407–421. <https://doi.org/10.1037/abn0000749>
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., & Wang, P. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, 167(7), 748–751. <https://doi.org/10.1176/appi.ajp.2010.09091379>
- Kaczurkin, A. N., Moore, T. M., Calkins, M. E., Ciric, R., Detre, J. A., Elliott, M. A., Foa, E. B., Garcia de la Garza, A., Roalf, D. R., Rosen, A., Ruparel, K., Shinohara, R. T., Xia, C. H., Wolf, D. H., Gur, R. E., Gur, R. C., & Satterthwaite, T. D. (2018). Common and dissociable regional cerebral blood flow differences associate with dimensions of psychopathology across categorical diagnoses. *Molecular Psychiatry*, 23(10), 1981–1989. <https://doi.org/10.1038/mp.2017.174>
- Kaczurkin, A. N., Park, S. S., Sotiras, A., Moore, T. M., Calkins, M. E., Cieslak, M., Rosen, A. F., Ciric, R., Xia, C. H., Cui, Z., Sharma, A., Wolf, D. H., Ruparel, K., Pine, D. S., Shinohara, R. T., Roalf, D. R., Gur, R. C., Davatzikos, C., Gur, R. E., & Satterthwaite, T. D. (2019). Dimensions of psychopathology are dissociably linked to brain structure in youth. *American Journal of Psychiatry*, 176(12), 1000–1009. <https://doi.org/10.1176/appi.ajp.2019.18070835>
- Kotov, R., Krueger, R. F., Watson, D., Cicero, D. C., Conway, C. C., DeYoung, C. G., Eaton, N. R., Forbes, M. K., Hallquist, M. N., Litzman, R. D., Mullins-Sweatt, S. N., Ruggero, C. J., Simms, L. J., Waldman, I. D., Waszczuk, M. A., & Wright, A. G. C. (2021). The hierarchical Taxonomy of psychopathology (HiTOP): A quantitative nosology based on consensus of evidence. *Annual Review of Clinical Psychology*, 17(1), 83–108. <https://doi.org/10.1146/annurev-clinpsy-081219-093304>
- Krueger, R. F., Hobbs, K. A., Conway, C. C., Dick, D. M., Dretsch, M. N., Eaton, N. R., Forbes, M. K., Forbush, K. T., Keyes, K. M., Litzman, R. D., Michelini, G., Patrick, C. J., Sellbom, M., Slade, T., South, S., Sunderland, M., Tackett, J., Waldman, I., Waszczuk, M. A., & Kotov, R. (2021). Validity and utility of hierarchical Taxonomy of psychopathology (HiTOP): II. Externalizing superspectrum. *World Psychiatry*, 20(2), 171–193. <https://doi.org/10.1002/wps.20844>
- Laceulle, O. M., Chung, J. M., Vollebergh, W. A. M., & Ormel, J. (2020). The wide-ranging life outcome correlates of a general psychopathology factor in adolescent psychopathology. *Personality and Mental Health*, 14(1), 9–29. <https://doi.org/10.1002/pmh.1465>
- Lahey, B. B., Tiemeier, H., & Krueger, R. F. (2022). Seven reasons why binary diagnostic categories should be replaced with empirically sounder and less stigmatizing dimensions. *JCPP Advances*, 2(n/a), e12108. <https://doi.org/10.1002/jcv.2.12108>
- 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>
- Liu, S., Zhang, Z., Yang, N., Zhang, Q., Zhou, Q., & Zuo, X.-N. (2020). Cohort profile: Chinese color nest project. *PsyArXiv*. <https://doi.org/10.31234/osf.io/d8kpx>
- Mansolf, M., Vreeker, A., Reise, S. P., Freimer, N. B., Glahn, D. C., Gur, R. E., Moore, T. M., Pato, C. N., Pato, M. T., Palotie, A., Holm, M., Suvisaari, J., Partonen, T., Kieseppa, T., Paunio, T., Boks, M., Kahn, R., Ophoff, R. A., Bearden, C. E., & Bilder, R. M. (2020). Extensions of multiple-group item response theory alignment: Application to psychiatric phenotypes in an international genomics consortium. *Educational and Psychological Measurement*, 80(5), 870–909. <https://doi.org/10.1177/0013164419897307>
- McElroy, E., Belsky, J., Carragher, N., Fearon, P., & Patalay, P. (2018). Developmental stability of general and specific factors of psychopathology from early childhood to adolescence: Dynamic mutualism or p-differentiation? *Journal of Child Psychology and Psychiatry*, 59(6), 667–675. <https://doi.org/10.1111/jcpp.12849>
- McElroy, E., Villadsen, A., Patalay, P., Goodman, A., Richards, M., Northstone, K., Fearon, P., Tibber, M., Gondek, D., & Ploubidis, G. B. (2021). Harmonisation and measurement properties of mental health measures in six British cohorts. CLOSER. Retrieved from <https://www.closer.ac.uk/wp-content/uploads/210715-Harmonisation-measurement-properties-mental-health-measures-british-cohorts.pdf>
- Moore, T. M., Kaczurkin, A. N., Durham, E. L., Jeong, H. J., McDowell, M. G., Dupont, R. M., Applegate, B., Tackett, J. L., Cardenas-Iniguez, C., Kardan, O., Akcelik, G. N., Stier, A. J., Rosenberg, M. D., Hedeker, D., Berman, M. G., & Lahey, B. B. (2020). Criterion validity and relationships between alternative hierarchical dimensional models of general and specific psychopathology. *Journal of Abnormal Psychology*, 129(7), 677–688. <https://doi.org/10.1037/abn0000601>
- Muthén, B., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus.
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén. Retrieved from [www.StatModel.com](http://www.StatModel.com)
- Nikolaidis, A., Heleniak, C., Fields, A., Bloom, P. A., VanTieghem, M., Vannucci, A., Camacho, N. L., Choy, T., Gibson, L., Harmon, C., Hadis, S. S., Douglas, I. J., Milham, M. P., & Tottenham, N. (2022). Heterogeneity in caregiving-related early adversity: Creating stable dimensions and subtypes. *Development and Psychopathology*, 34(2), 621–634. <https://doi.org/10.1017/S0954579421001668>

- Nooner, K. B., Colcombe, S. J., Tobe, R. H., Mennes, M., Benedict, M. M., Moreno, A. L., Panek, L. J., Brown, S., Zavitz, S. T., Li, Q., Sikka, S., Gutman, D., Bangaru, S., Schlachter, R. T., Kamiel, S. M., Anwar, A. R., Hinz, C. M., Kaplan, M. S., Rachlin, A. B., & Milham, M. P. (2012). The NKI-rockland sample: A model for accelerating the pace of discovery science in psychiatry. *Frontiers in Neuroscience*, *6*, 152. <https://doi.org/10.3389/fnins.2012.00152>
- Ploubidis, G. B., McElroy, E., & Moreira, H. C. (2019). A longitudinal examination of the measurement equivalence of mental health assessments in two British birth cohorts. *Longitudinal and Life Course Studies*, *10*(4), 471–489. <https://doi.org/10.1332/175795919X15683588979486>
- 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>
- Sallis, H., Szekely, E., Neumann, A., Jolicoeur-Martineau, A., Ijzendoorn, M. van, Hillegers, M., Greenwood, C. M., Meaney, M. J., Steiner, M., Tiemeier, H., Wazana, A., Pearson, R. M., & Evans, J. (2019). General psychopathology, internalising and externalising in children and functional outcomes in late adolescence. *Journal of Child Psychology and Psychiatry*, *60*(11), 1183–1190. <https://doi.org/10.1111/jcpp.13067>
- Salum, G. A., Gadelha, A., Pan, P. M., Moriyama, T. S., Graeff-Martins, A. S., Tamanaha, A. C., Alvarenga, P., Krieger, F. V., Fleitlich-Bilyk, B., Jackowski, A., Sato, J. R., Brietzke, E., Polanczyk, G. V., Brentani, H., de Jesus Mari, J., Do Rosario, M. C., Manfro, G. G., Bressan, R. A., Mercadante, M. T., & Rohde, L. A. (2015). High risk cohort study for psychiatric disorders in childhood: Rationale, design, methods and preliminary results. *International Journal of Methods in Psychiatric Research*, *24*(1), 58–73. <https://doi.org/10.1002/mpr.1459>
- Satterthwaite, T. D., Connolly, J. J., Ruparel, K., Calkins, M. E., Jackson, C., Elliott, M. A., Roalf, D. R., Hopson, R., Prabhakaran, K., Behr, M., Qiu, H., Mentch, F. D., Chiavacci, R., Sleiman, P. M., Hakonarson, H., & Gur, R. E. (2016). The Philadelphia neurodevelopmental cohort: A publicly available resource for the study of normal and abnormal brain development in youth. *NeuroImage*, *124*(Pt B), 1115–1119. <https://doi.org/10.1016/j.neuroimage.2015.03.056>
- Shanmugan, S., Wolf, D. H., Calkins, M. E., Moore, T. M., Ruparel, K., Hopson, R. D., Vandekar, S. N., Roalf, D. R., Elliott, M. A., Jackson, C., Gennatas, E. D., Leibenluft, E., Pine, D. S., Shinohara, R. T., Hakonarson, H., Gur, R. C., Gur, R. E., & Satterthwaite, T. D. (2016). Common and dissociable mechanisms of executive system dysfunction across psychiatric disorders in youth. *American Journal of Psychiatry*, *173*(5), 517–526. <https://doi.org/10.1176/appi.ajp.2015.15060725>
- Smith, G. T., Atkinson, E. A., Davis, H. A., Riley, E. N., & Oltmanns, J. R. (2020). The general factor of psychopathology. *Annual Review of Clinical Psychology*, *16*(1), 75–98. <https://doi.org/10.1146/annurev-clinpsy-071119-115848>
- Sullivan, P. F., Agrawal, A., Bulik, C. M., Andreassen, O. A., Børglum, A. D., Breen, G., Cichon, S., Edenberg, H. J., Faraone, S. V., Gelernter, J., Mathews, C. A., Nievergelt, C. M., Smoller, J. W., O'Donovan, M. C., & O'Donovan, M. C. (2018). Psychiatric genomics: An update and an agenda. *American Journal of Psychiatry*, *175*(1), 15–27. <https://doi.org/10.1176/appi.ajp.2017.17030283>
- Thompson, E. J., Richards, M., Ploubidis, G. B., Fonagy, P., & Patalay, P. (2021). Changes in the adult consequences of adolescent mental ill-health: Findings from the 1958 and 1970 British birth cohorts. *Psychological Medicine*, 1–10. <https://doi.org/10.1017/S0033291721002506>
- Thompson, P. M., Jahanshad, N., Schmaal, L., Turner, J. A., Winkler, A. M., Thomopoulos, S. I., Egan, G. F., & Kochunov, P. (2022). The enhancing Neuroimaging genetics through meta-analysis consortium: 10 Years of global collaborations in human brain mapping. *Human Brain Mapping*, *43*(1), 15–22. <https://doi.org/10.1002/hbm.25672>
- Thompson, P. M., Stein, J. L., Medland, S. E., Hibar, D. P., Vasquez, A. A., Renteria, M. E., Toro, R., Jahanshad, N., Schumann, G., Franke, B., Wright, M. J., Martin, N. G., Agartz, I., Alda, M., Alhusaini, S., Almasy, L., Almeida, J., Alpert, K., Andreassen, N. C., & Alzheimer's Disease Neuroimaging Initiative, EPIGEN Consortium, IMAGEN Consortium, Saguenay Youth Study (SYS) Group. (2014). The ENIGMA Consortium: Large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging and Behavior*, *8*(2), 153–182. <https://doi.org/10.1007/s11682-013-9269-5>
- Waller, N., Jones, J., & Giordano, C. (2022). fungible: Psychometric functions from the waller lab. Retrieved from <https://CRAN.R-project.org/package=fungible>
- Waszczuk, M. A., Miao, J., Docherty, A. R., Shabalin, A. A., Jonas, K. G., Michelini, G., & Kotov, R. (2021). General v. specific vulnerabilities: Polygenic risk scores and higher-order psychopathology dimensions in the Adolescent Brain Cognitive Development (ABCD) Study. *Psychological Medicine*, 1–10. <https://doi.org/10.1017/S0033291721003639>
- Watts, A. L., Lane, S. P., Bonifay, W., Steinley, D., & Meyer, F. A. C. (2020). Building theories on top of, and not independent of, statistical models: The case of the p-factor. *Psychological Inquiry*, *31*(4), 310–320. <https://doi.org/10.1080/1047840x.2020.1853476>

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Hoffmann, M. S., Moore, T. M., Axelrud, L. K., Tottenham, N., Rohde, L. A., Milham, M. P., Satterthwaite, T. D., & Salum, G. A. (2023). Harmonizing bifactor models of psychopathology between distinct assessment instruments: Reliability, measurement invariance, and authenticity. *International Journal of Methods in Psychiatric Research*, e1959. <https://doi.org/10.1002/mpr.1959>