



The General Psychopathology ‘p’ Factor in Adolescence: Multi-Informant Assessment and Computerized Adaptive Testing

Jason D. Jones^{1,2} · Rhonda C. Boyd^{1,2} · Akira Di Sandro² · Monica E. Calkins² · Andres De Los Reyes³ · Ran Barzilay^{1,2} · Jami F. Young^{1,2} · Tami D. Benton^{1,2} · Ruben C. Gur² · Tyler M. Moore² · Raquel E. Gur^{1,2}

Accepted: 6 June 2024

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

Accumulating evidence supports the presence of a general psychopathology dimension, the p factor (‘p’). Despite growing interest in the p factor, questions remain about how p is assessed. Although multi-informant assessment of psychopathology is commonplace in clinical research and practice with children and adolescents, almost no research has taken a multi-informant approach to studying youth p or has examined the degree of concordance between parent and youth reports. Further, estimating p requires assessment of a large number of symptoms, resulting in high reporter burden that may not be feasible in many clinical and research settings. In the present study, we used bifactor multidimensional item response theory models to estimate parent- and adolescent-reported p in a large community sample of youth (11–17 years) and parents ($N=5,060$ dyads). We examined agreement between parent and youth p scores and associations with assessor-rated youth global functioning. We also applied computerized adaptive testing (CAT) simulations to parent and youth reports to determine whether adaptive testing substantially alters agreement on p or associations with youth global functioning. Parent-youth agreement on p was moderate ($r=.44$) and both reports were negatively associated with youth global functioning. Notably, 7 out of 10 of the highest loading items were common across reporters. CAT reduced the average number of items administered by 57%. Agreement between CAT-derived p scores was similar to the full form ($r=.40$) and CAT scores were negatively correlated with youth functioning. These novel results highlight the promise and potential clinical utility of a multi-informant p factor approach.

Keywords Psychopathology · P Factor · Multi-informant · Discrepancies · Computerized Adaptive Testing

In research and clinical practice, psychopathology has traditionally been viewed in terms of distinct diagnostic categories. Recently, however, evidence has been accumulating for the presence of a general psychopathology dimension, referred to as the ‘p’ factor, that spans mental disorders (e.g., Caspi et al., 2014; Lahey et al., 2012, 2021; Ronald, 2019).

The p factor is conceptualized as a transdiagnostic construct reflecting severity of overall psychopathology and/or the degree of comorbidity among a range of symptoms (Fried et al., 2021). The concept of the p factor is supported by an abundance of research demonstrating high comorbidity, transdiagnostic risk factors, and shared sequelae of various forms of psychopathology (Gili et al., 2019; Kessler et al., 2005; Nolen-Hoeksema & Watkins, 2011). The p factor shows high heritability and stability over time (Allegrini et al., 2020; Class et al., 2019; McElroy et al., 2018). Further, the p factor is associated with a number of relevant psychological constructs, such as functional impairment, psychiatric diagnoses, psychiatric medication use, family history of psychopathology, and neural correlates of mental disorders (Caspi et al., 2014; Moore et al., 2019; Pettersson et al., 2018).

Despite growing support for the use of the p factor in clinical research, questions remain about how the p factor

✉ Jason D. Jones
jonesjd@chop.edu

¹ Department of Child and Adolescent Psychiatry and Behavioral Sciences, Children’s Hospital of Philadelphia, Roberts Center for Pediatric Research, 2716 South Street, Philadelphia PA 19146, USA

² Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, USA

³ Department of Psychology, University of Maryland, College Park, Maryland, USA

is assessed. First, whereas research with adults has typically relied on self-reports to assess the *p* factor, research with youth typically takes a multi-informant approach to assessment wherein symptom ratings are obtained from youth and a collateral informant (typically a caregiver). Yet, research on the multi-informant assessment of *p* is lacking, resulting in major gaps in the literature. Second, estimating the *p* factor requires assessing a large number of symptoms that span a range of disorders, resulting in high reporter burden. Some recent evidence suggests *p* can be estimated more efficiently using computerized adaptive testing (CAT) procedures with little or no cost to construct or predictive validity (Moore et al., 2019). However, this research applied CAT only to youth self-reported symptoms. It is unclear if applying CAT to parent-reported youth symptoms would yield similarly promising results. The present investigation aims to address these gaps in the literature.

Multi-Informant Assessment of Youth Psychopathology

It is a well-established practice for researchers and clinicians interested in youth mental health to obtain ratings from multiple informants (e.g., youth, caregiver, teacher). However, an abundance of evidence suggests that informants often disagree, sometimes quite substantially, in their ratings of various symptoms and behaviors (e.g., Achenbach et al., 1987; De Los Reyes et al., 2023). A meta-analysis of 25 years of multi-informant studies reported an average overall correlation of 0.29 between parent and youth reports of internalizing ($r=0.26$) and externalizing ($r=0.32$) symptoms (De Los Reyes et al., 2015). Discrepancies in reports of youth psychopathology symptoms have important clinical implications. For example, in inpatient psychiatric settings, greater discrepancies in youth and parent reports of youth symptoms have been prospectively associated with use of intensive and restrictive treatment regimens (e.g., standing antipsychotics, locked door seclusion; Makol et al., 2019). Further, in outpatient settings, parent-child discrepancies at the beginning of therapy predicted poorer overall treatment response (Goolsby et al., 2018), and increased parent-youth concordance over the course of treatment predicted better outcomes (Becker-Haimes et al., 2018).

To our knowledge, there has been very limited investigation of cross-informant agreement on youth *p*. Allegrini et al. (2020) used principal components analysis (PCA) to derive a *p* factor score from youth and parent reports of youth psychopathology at multiple time points and found that correlations among the first components extracted (i.e., the general components) across time points ranged from approximately 0.30–0.40, similar to the correlations

reported by De Los Reyes et al. (2015). However, there are limitations to a PCA-derived *p* factor. Namely, PCA does not capture the hierarchical structure of psychopathology with a general *p* factor and several sub-factors (see Lahey et al., 2021). Indeed, loadings on the first principal component are almost certainly biased by the underlying multidimensionality (Reise et al., 2011, 2015). Other statistical approaches, such as bifactor modeling, correct for this multidimensionality. Bifactor modeling (Reise, 2012; Reise et al., 2010) is unique in that it allows each item to load on two factors simultaneously; one specific factor capturing covariance among items in a sub-domain (e.g., externalizing) and one general factor (*p*) capturing covariance among all items. A critical advantage of bifactor modeling in this context is that it includes direct relationships between the general factor (*p*) and the items themselves. Contrast this with a second-order model (see Reise et al., 2010; Fig. 1c) in which items load on their specific factor only and these specific factors load on a second-order general factor. This configuration will not work in the present context because one cannot estimate item parameters on the general factor without a direct relationship between the item and the factor. Item parameter estimates—specifically item response theory discrimination and threshold parameters—are necessary to build a CAT. In the present work, therefore, we employed bifactor multidimensional item response theory (MIRT) models (Reckase, 1997) to derive adolescent- and parent-reported *p* factor scores and evaluate agreement between scores.

Computerized Adaptive Testing (CAT)

In the extant literature, *p* factor scores have typically been derived from lengthy and time-consuming clinical assessments that include numerous items spanning a range of disorders and symptom clusters. Other areas of research and assessment (e.g., standardized testing, cognitive/IQ tests) have adopted adaptive procedures that substantially reduce testing burden while maintaining a high level of precision relative to the non-adaptive version. In CAT, after the first item is completed, an algorithm estimates the responder's trait level of the construct being assessed and then chooses the most appropriate next item, where "most appropriate" is determined by the amount of information¹ an item will provide at that examinee's estimated interim trait level. After the second item is completed, the algorithm uses both responses to re-estimate the trait level to select the next most appropriate item, and so on until some stopping criterion is reached. Recently, Moore et al. (2019) applied CAT to data from the Philadelphia Neurodevelopmental Cohort

¹ Information is defined in the [Analytic Approach](#) section.

(PNC) study to create a publicly available adaptive screener for assessing p , which they called the Overall Mental Illness (OMI) screener. They reported that the CAT version (substantially shorter than the full form) performed nearly as well as the full version in predicting psychiatric diagnoses and brain parameters. However, Moore et al. (2019) only used youth report to create the adaptive assessment of p . We extended this work in the current investigation by applying CAT to both parent and adolescent reports of youth psychopathology in the PNC and examining agreement between CAT p factor scores. If p factor scores derived from CAT perform similarly to scores derived from lengthy p factor assessments, CAT could reduce assessment burden, and thus provide researchers and clinicians with an efficient way to collect multi-informant ratings of youth psychopathology.

The Present Study

In the present investigation, we utilized data from a large sample of youth and collateral informants in the PNC study (Calkins et al., 2015) to evaluate the degree of agreement between youth- and parent-reported p . The PNC includes over 5,000 youth-parent dyads who were independently administered a structured clinical interview by an assessor and thus provides a rich resource for robustly evaluating issues related to informant agreement. Prior multi-informant analyses with the PNC sample revealed substantial discrepancies in youth and parent reports of adolescent substance use (Jones et al., 2017a), suicidal ideation (Jones et al., 2019), and psychosis spectrum symptoms (Xavier et al., 2022). The present analyses extend this prior work by utilizing all the symptom-level data available and employing bifactor models to create overall psychopathology p factor scores for youth and parent reports. In addition to evaluating agreement between p factor scores, we also examined which items loaded most highly on youth- and parent-reported p to see which, if any, of the highest loading items were common across reporters. We also tested associations between each reporter's p factor score and assessor-rated youth global functioning. We chose this criterion of clinical validity because, like the p factor, it is transdiagnostic. Equally important, the validity criterion was independently rated by an assessor, rather than reported by the adolescent and/or parent. This aspect of our study design addressed issues recently raised with use of mono-source paradigms for measurement validation (see De Los Reyes et al., 2023; Watts et al., 2022). Finally, we applied simulated CAT to parent and youth reports of psychopathology to determine whether adaptive testing substantially alters the degree of agreement between parent and youth reports of p or the associations with youth global functioning.

Method

Participants

The PNC includes 9,498 community youth between the ages of 8 and 21 years from the greater Philadelphia area. Participants were recruited from the Children's Hospital of Philadelphia (CHOP) pediatric healthcare network. Importantly, participants were not recruited from mental health treatment centers; thus, the PNC is not enriched for individuals seeking psychiatric care. To be eligible for the PNC, participants had to be: (a) aged 8–21 years, (b) proficient in English, and (c) not have any mental or physical conditions that could interfere with the completion of study procedures.

Notably, for adolescents between the ages of 11 and 17 years, both the youth and a collateral informant completed a structured clinical interview that screens for a wide range of youth psychopathology symptoms. The present analyses included 5,060 adolescents (mean age = 14.54 years, $SD = 1.98$; 52% female) and collateral informants (87% mother/mother figure; 10% father/father figure; 3% other caregiver/legal guardian) from the PNC. The sub-sample was racially diverse: 57% White, 32% Black, and 11% other races (e.g., multiracial, Pacific Islander).

Procedures

After providing a detailed description of study procedures, written parental consent and youth assent were obtained. Parents and youth were assessed independently and were informed

that all their responses would be kept confidential, with the exception of legal reporting requirements (i.e., self/other harm, evidence of abuse). See Calkins et al. (2015) for additional details about study recruitment, sample, and procedures. The Institutional Review Boards of CHOP and the University of Pennsylvania approved all study procedures.

Measures

Clinical assessment. Trained assessors administered the computerized GOASSESS structured clinical interview to youth and collateral informants (Calkins et al., 2015). GOASSESS is derived from the Schedule for Affective Disorders and Schizophrenia for School-Aged Children (KSADS; Kaufman et al., 1997) and screens for major domains of psychopathology (e.g., mood, anxiety, attention/behavior, psychosis spectrum). In the present investigation, youth and parent p factor scores were derived from 107 dichotomous items from GOASSESS that assess a range of psychopathology symptoms (see Table S1 in Supplemental Materials for a list of all 107 items). Participants were also

asked about distress/impairment associated with symptoms and about history of mental health treatment for mood or behavioral problems. Based on all information provided during the clinical interviews, the assessor assigned a global functioning score using the Children’s Global Assessment Scale (CGAS; Shaffer et al., 1983) GOASSESS has been validated as a psychopathology screener in numerous studies (e.g., Barzilay et al., 2019a; Calkins et al., 2015; Jones et al., 2017b, 2021; Moore et al., 2019; Satterthwaite et al., 2014). Previous bifactor analyses of the GOASSESS youth symptom data found an overall psychopathology factor (i.e., p) and four subfactors: (1) anxious-misery (e.g., depression, generalized anxiety disorder), (2) fear (e.g., panic disorder, phobias), (3) externalizing (e.g., ADHD, conduct disorder), and (4) psychosis spectrum symptoms (Calkins et al., 2015; Moore et al., 2019, 2023; Shanmugan et al., 2016).

Analytic Approach

Multidimensional Item Response Theory Models

To determine the optimal factor configuration for youth and collateral informants, we used exploratory multidimensional item response theory (MIRT; McDonald, 2000; Reckase, 1997) models implemented in the *mirt* package (Chalmers, 2012) in R. The number of factors to extract (4) was based on previous analyses of these data (Moore et al., 2019, 2023; Shanmugan et al., 2016), as well as the theoretical 4-factor structure of psychopathology. This 4-factor structure combines the original 3-factor structure from Krueger (1999), itself roughly comprising anxious-misery, fear, and externalizing, with a fourth factor, psychosis spectrum symptoms (Calkins et al., 2015; cf. Markon, 2010). Subjective evaluation of the scree plot, as well as the minimum average partial (MAP) method (Velicer, 1976) for determining the number of factors, were consistent with the choice of 4 factors. With the optimal item configuration determined by MIRT, a confirmatory bifactor MIRT model was fit to youth and collateral informant data. This allowed the general p factor to be estimated for each with optimal weights uncontaminated by each reporter’s unique multidimensionality (Reise et al., 2015). Note that, if we wished to make claims about the “structure of psychopathology,” the above analyses would need to be performed in a cross-validated framework whereby the model configuration is determined in a separate sub-sample from the one used for the confirmatory analysis. However, here we do not wish to make such claims, and confirmatory models were used only for the purpose of calibrating the items needed for all downstream analyses. Item calibration produced parameter estimates (e.g., each item’s severity) that were then used as

an item bank for performing CAT simulations and calculating p factor scores (interim and final).

Computerized Adaptive Testing

The measurement theory facilitating CAT is item response theory (IRT; Embretson & Reise, 2000), which describes how latent traits relate to the probability of item responses. The most demonstrative model of IRT is the unidimensional 2-parameter model described by Eq. 1:

$$p_i(X_i = 1|\theta) = \frac{e^{a_i(\theta-b_i)}}{1 + e^{a_i(\theta-b_i)}} \tag{1}$$

Where $p_i(X_i = 1|\theta)$ is the probability of endorsing item i (given θ), a_i is the item discrimination for item i , b_i is the item severity (or “difficulty”), and θ is the trait level (e.g., psychopathology severity) of the person. For example, if an item had discrimination (a)=2.0 and severity (b)=0.5, and an examinee had trait level (θ)=1.0, the probability that that examinee would endorse the item would be $e^{2.0(1.0-0.5)}/(1 + e^{2.0(1.0-0.5)}) = e^1/(1 + e^1) = 2.718^1/(1 + 2.718^1) = 0.731$. By providing an estimate of endorsement probability, the item parameter estimates (a_i and b_i) provide an estimate of item “quality” at any specific point along the trait dimension, because knowing a_i and the probability of endorsement allows one to calculate the information provided by that item:

$$I(\theta) = a_i^2 p_i(\theta) q_i(\theta) \tag{2}$$

Where $I(\theta)$ is the information produced by the item, a_i is the item discrimination, $p_i(\theta)$ is the probability of endorsement, and $q_i(\theta)$ is the probability of non-endorsement. Using the example above, the information provided by that item ($a=2.0$) for that examinee ($\theta=1.0$) would be $2^2(0.731)(0.269) = 4(0.197) = 0.788$. Item information, in turn, relates to the standard error of measurement as:

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \tag{3}$$

Where $SE(\theta)$ is the measurement error, meaning that as information increases, error decreases. For the example above, the $SE(\theta)$ would be $1/\sqrt{0.788} = 1.127$ standard deviations. A trait level (θ) estimate from that one item would therefore have quite wide 95% confidence intervals around it, spanning roughly 2.2 standard deviations above and below. As additional items are administered, more information is accumulated, reducing the standard error.

Equation 1 above can be expanded into a multidimensional model (Reckase, 2009) as:

$$p_i(X_i = 1 | \theta_1, \theta_2) = \frac{e^{a_{i1}(\theta_1 - b_i) + a_{i2}(\theta_2 - b_i)}}{1 + e^{a_{i1}(\theta_1 - b_i) + a_{i2}(\theta_2 - b_i)}} \quad (4)$$

Where $p_i(X_i = 1 | \theta_1, \theta_2)$ is the probability of endorsement of item i given two different trait dimensions (θ_1 and θ_2), a_{i1} is the item discrimination for dimension 1, a_{i2} is the item discrimination for dimension 2, and b_i is the item difficulty. However, while the present study does use a multidimensional model (bifactor; Reise, 2012), the focus is on only one dimension (i.e., the general factor).

With the ability to characterize item quality (information provided) as a function of examinee trait level, the ability to administer tests adaptively follows neatly. Imagine an examinee of unknown trait level, assumed to be average ($\theta=0$). Each item in a collection (item bank) of calibrated items has an associated amount of information provided for an examinee at any trait level, and a test can start by administering the item with maximum information at $\theta=0$. The response to that item (endorsed or not) would result in an updated estimated trait level—e.g., suppose the examinee endorsed it and θ is re-estimated to be 0.80. The next item to administer would be the one that provides maximum information at $\theta=0.80$. Suppose the examinee does not endorse this second item, resulting in a re-estimated trait level of 0.10. The item providing maximum information at $\theta=0.10$ would be administered, and so on. The test would stop when some pre-established stopping criterion is met. A common approach is to stop when the standard error [$SE(\theta)$ from Eq. 3 above] reaches a lower bound (e.g., 0.30).

Using the above analytic framework, CAT sessions were simulated for the purpose of calculating what score each person would have received had they taken an adaptive version of the psychopathology screener (and received far fewer items). This was possible because all participants had already answered all items (in the full form), meaning any item administered in the CAT simulation would have a corresponding response from that person (given when they took the full version). A standard error cutoff of 0.30 was used in the simulations, meaning each simulated CAT stopped when the standard error reached that level, and the person received whatever score the algorithm had estimated at that point in the CAT sequence. With the CAT simulation scores estimated (in a Z metric, as is typical of CAT), they could be compared to the full form scores (estimated using the full item bank). CAT scores were estimated using the default Bayesian modal method (Birnbaum, 1969) common in CAT, and full-form scores were estimated using expected a posteriori (EAP; Bock & Mislevy, 1982).

p Factor Agreement and Associations with Youth Global Functioning

We used Pearson correlations to evaluate youth-parent agreement on p factor scores and the associations between each reporter's p factor scores and youth CGAS scores. To compare agreement on p in the full version versus agreement on p in the CAT version we used Steiger tests (Steiger, 1980). To test whether the magnitude of the correlation between p factor and CGAS differed by reporter (parent vs. youth) or version (full vs. CAT), we used Williams' tests (Williams, 1959). Given the large sample size and the number of comparisons, we set a stringent significance threshold of $p < .001$.

Results

Bifactor Model Results

Fit of the bifactor models with four subfactors was acceptable for youth (CFI=0.97, SRMR=0.04, RMSEA=0.03) and collateral informants (CFI=0.97, SRMR=0.05, RMSEA=0.03). In addition, while interpretation is beyond the present scope, it is common to report several indices specific to bifactor models (Rodriguez et al., 2016a); these are presented in Supplementary Tables S2 and S3 for youth and collateral informants, respectively. The indices (such as omega, omega-hierarchical, and factor determinacy) are used to judge qualities such as: (1) factor reliability, especially that unique to a specific factor (e.g., externalizing relative to p), (2) the extent to which we can expect factor scores to represent the factor used to calculate them, (3) the proportion of inter-item correlations that are “uncontaminated” by multidimensionality, and other similar properties. All general factor indices were within the acceptable range (Rodriguez et al., 2016b).

Bifactor Item Loadings

The 10 highest loading items for youth- and parent-reported p are presented in Table 1 (see Table S1 in Supplemental Materials for all item loadings). It is noteworthy that 7 out of 10 items were common across reporters and that most of the items capture symptoms of obsessive-compulsive disorder (OCD). For example, the GOASSESS item assessing intrusive and repetitive bad or forbidden thoughts was the highest loading item on p factor scores for both parents and adolescents.

Table 1 Top 10 Highest Loading Items in Bifactor Models

| Youth-Reported <i>p</i> | | Parent-Reported <i>p</i> | |
|--|---------|---|---------|
| Item | Loading | Item | Loading |
| OCD006: Forbidden/bad thoughts | 0.83 | OCD006: Forbidden/bad thoughts | 0.87 |
| OCD004: Fear doing or saying something bad | 0.81 | OCD004: Fear doing or saying something bad | 0.86 |
| OCD001: Concern with harming self/others | 0.81 | OCD002: Images of violent things | 0.86 |
| OCD005: Bad things your fault | 0.80 | OCD001: Concern with harming self/others | 0.86 |
| MAN007: Grouchy, cranky, irritable | 0.78 | OCD005: Bad things your fault | 0.86 |
| OCD002: Images of violent things | 0.78 | OCD008: Religious thoughts | 0.84 |
| OCD014: Getting dressed over and over | 0.76 | PAN004: Felt losing control, something bad happen | 0.80 |
| OCD015: Open door over and over | 0.76 | OCD015: Open door over and over | 0.80 |
| OCD008: Religious thoughts | 0.76 | MAN005: Unusually elevated mood | 0.78 |
| MAN004: Fast talking/thinking | 0.72 | OCD012: Repetitive counting | 0.78 |

Note. Full item text available in supplemental Table S1

Table 2 Top 10 Most Frequently Administered Items in CAT Simulation

| Youth-Reported <i>p</i> | | Parent-Reported <i>p</i> | |
|--|-----|---|-----|
| Item | % | Item | % |
| ODD006: Irritable, grouchy, angry | 100 | ADD011: Trouble paying attention | 100 |
| MAN001: Active, excited, energetic | 100 | ADD012: Problems following instructions | 100 |
| MAN007: Grouchy, cranky, irritable | 90 | ODD002: Refusing to do as told; breaking rules | 99 |
| DEP004: Grouchy, irritable, bad mood | 88 | ADD016: Not listening; daydreaming | 94 |
| SIP003: Odd or unusual things going on | 80 | ODD001: Lose temper, argue, or be grouchy/irritable with adults | 89 |
| SIP007: Confused about whether something is real or imagined | 80 | ODD006: Irritable, grouchy, angry | 84 |
| DEP006: Nothing is fun; not interested in anything | 77 | DEP004: Grouchy, irritable, bad mood | 81 |
| SIP012: Hear talking when no one near me | 76 | ADD015: Trouble completing tasks with lots of different steps | 77 |
| SIP011: Mind is “playing tricks” on me | 75 | GAD001: Been a worrier | 67 |
| MAN003: Hardly needed sleep | 74 | ADD013: Dislike/avoid schoolwork | 63 |

Note. Full item text available in supplemental Table S1

Table 3 Associations between *p* and Youth Global Functioning

| | Comparison | | <i>p</i> -value |
|-----------------|---------------------------------------|--------------------------------------|-----------------|
| CGAS Score | Youth <i>p</i> (full) $r = -0.48$ | Youth <i>p</i> (CAT) $r = -0.45$ | <0.001 |
| CGAS Score | Parent <i>p</i> (full) $r = -0.46$ | Parent <i>p</i> (CAT) $r = -0.44$ | <0.001 |
| <i>p</i> -value | 0.25 | 0.39 | |

*Note. Threshold $p < .001$

CAT Simulations

In CAT simulations, youth were administered 46 items on average (range 7-107 items) and parents were administered 47 items on average (range = 6-107 items). This corresponds to approximately 43% of the original (full form, 107 items) test length across reporters. Notably, OCD symptoms were not among the 10 most frequently used items in the CAT simulations and there was minimal item overlap between youth and parents (see Table 2). The items most frequently used in CAT were related to mood, externalizing, and psychosis spectrum symptoms. Nevertheless, the within-reporter correlation between the full version *p* factor score and CAT-derived *p* factor score was high for both youth-report ($r = 0.95$) and parent-report ($r = 0.95$). Exploratory analyses including all 214 parent and youth items combined are reported in Supplemental Materials.

Agreement between Youth- and Parent-Reported *p* Factor Scores

The correlation between youth- and parent-reported *p* factor scores on the full form was moderate, $r = 0.44$, $p < .001$. Parent-youth agreement on CAT-derived *p* factors scores was similar to the full form version, $r = 0.40$, $p < .001$. The difference in the magnitude of these correlations, although small (0.04), was statistically significant, $p < .001$.

Associations of *p* Factor Scores with Overall Functioning

Both youth- and parent-reported *p* (full and CAT versions) were negatively associated with assessor-rated youth global functioning (r ranged from -0.44 to -0.48 ; $p < .001$, see Table 3). Within each reporter, using CAT only slightly reduced the magnitude of the association between *p* and CGAS scores (0.03 for youth and 0.02 for parents, $p < .001$; rows in Table 3). Across reporters, the correlation between *p* factor scores (full version or CAT version) and CGAS scores did not significantly differ for youth report versus collateral informant report (columns in Table 3).

Discussion

Interest in a general psychopathology *p* factor has grown rapidly in recent years among researchers, clinicians, and psychometricians. Yet several important questions about the *p* factor have been insufficiently addressed. Although multi-informant assessment of youth psychopathology is considered part of “best practices” in clinical research and practice, almost no research has taken a multi-informant approach to studying the *p* factor (see Allegrini et al., 2020 and Watts et al., 2022, for exceptions). Our results fill several gaps in the literature related to cross-informant agreement on *p* and the clinical validity of the *p* factor when reported by youth and a collateral informant. In addition, our results suggest that lengthy clinical assessments used to generate *p* factor scores could be substantially abbreviated through adaptive testing procedures with minimal consequences to parent-youth agreement or construct validity.

It is noteworthy that among the 10 items that loaded most highly on youth- and parent-reported *p* in the bifactor models, seven of the items were common across reporters. The majority of the highest loading items for each reporter assessed symptoms of OCD and were related to repetitive bad thoughts and repetitive behaviors. Other high loading items across reporters were related to mood dysregulation. In some of the earliest research on the *p* factor, Caspi et al. (2014) found that thought disorder symptoms (which included OCD symptoms) loaded most highly on the general psychopathology factor. This finding is consistent with recent evidence suggesting that OCD symptoms among youth in the PNC study, particularly repetitive bad thoughts, are associated with increased risk for depression, psychosis, and suicidal ideation, indicating thought disturbance as a transdiagnostic risk factor for psychopathology (Barzilay et al., 2019b). The study by Barzilay and colleagues relied only on youth self-reported OCD symptoms. The present results extend this prior work by suggesting that parent-reported youth OCD symptoms may meaningfully capture general psychopathology risk. Further, researchers have recently proposed a general cognitive vulnerability factor (dubbed the ‘*c*’ factor) that is significantly associated with the *p* factor and is a transdiagnostic risk factor for multiple psychopathologies in youth (Schweizer et al., 2020). Our results further support thought disturbance (particularly intrusive repetitive thoughts) as a major underlying component of the *p* factor, and it is noteworthy that this was the case across reporters.

However, it is important to note that there is ongoing debate about what the *p* factor captures and possible underlying mechanisms that might explain a single general dimension of psychopathology (see Watts et al., 2024, for a review of key issues). One important criticism is that there

is currently no satisfactory explanation for why *p* exists; there are theories, but they are often unfalsifiable or otherwise unsatisfactory. Researchers have proposed several mechanisms that are transdiagnostic and cut across a range of disorders, including dispositional negative affectivity, emotion dysregulation, impairment, and thought dysfunction (Caspi & Moffit, 2018; Duetz et al., 2020; Lahey et al., 2021; Smith et al., 2020; Tackett et al., 2013). Further compounding the problem, the strongest indicators of *p* vary from study to study, where one study might define a *p* factor strongly determined by psychosis symptoms and another study’s *p* might be most strongly determined by mood symptoms (Watts et al., 2024). These differences are partially attributable to differences in measurement approach (as there is no standardized assessment of *p*) and sample characteristics across studies. Notably, even within the same sample, the strongest indicators of the *p* factor vary across reporters. Allegrini et al. (2020) found that autism traits and externalizing problems loaded highest on parent- and teacher-reported *p* factor, whereas internalizing symptoms loaded most highly on youth-reported *p*.

This lack of replication across *p* factor studies is a problem for the field, further exemplified by our finding that OCD symptoms most strongly defined the *p* factor in the present study. However, we argue this problem does not imply that *p* is an illusory or false construct. That *p* exists is demonstrated by the consistency of finding that any combination of psychopathology items (or collection of psychopathology measures) will produce a correlation matrix that is unmistakably unidirectional—i.e., all items correlate in the same direction. For researchers who wish to parse specific types of psychopathology, this “positive manifold” of symptoms is a nuisance. It results in collinearity, often severe, and causes uncertainty about the optimal way to model the distinct disorders or dimensions (such as internalizing versus externalizing). However, in some cases (as here), the distinctions among disorders or dimensions may be less important; the construct of interest may simply be psychopathology or mental health risk. In these cases, the “positive manifold” works in the researchers’ favor, and the *multidimensionality* is nuisance (accounted for by the bifactor configuration). To summarize, while we might not have a satisfactory explanation of what *p* is, it likely exists and serves the purpose of some research questions—e.g., as a general indicator of mental health risk—that can be used to compare parent and youth reports and is correlated with relevant psychological constructs (e.g., functioning). Therefore, it is worthwhile to pursue ways to make the assessment of *p*, as operationalized here, more time efficient than having to query about the entire range of symptoms. Such time savings could be particularly important in a low-yield population such as community samples or healthy controls

in clinical studies. An efficient measure of *p* could also be used as a screening tool, enabling identification of individuals where more detailed evaluation is indicated.

In contrast to the bifactor loadings, OCD symptoms were not among the most frequently used items in CAT simulations for either reporter. Critically, a high factor loading (analogous to a high item discrimination parameter, where higher is preferable) does not necessarily indicate a “high-quality”² item. This is because item quality depends not only on its discrimination but also on its difficulty (severity in this case), where items of extreme severity (high or low) tend to be less informative on average. The OCD symptoms here are a good example: they have very high loadings (and discrimination parameters) but are not selected very often in the CAT simulations because they tend to be severe/rare. Further, unlike the highest loading items which were largely similar across reporters, there was little overlap between reporters in the items most frequently selected in the CAT simulations. The youth-reported items that were most frequently selected in CAT were related to internal experiences and, interestingly, 40% of the items assess psychosis spectrum symptoms. By contrast, the parent-reported items that were most frequently selected in CAT largely capture externalizing symptoms and no psychosis spectrum symptoms were selected. These results are consistent with prior work suggesting that parents may be more attuned to observable, problematic behaviors but may miss less salient internal states that youth experience (De Los Reyes et al., 2015; Xavier et al., 2022). Despite these differences, correlations between full form *p* factor scores and CAT-derived *p* factor scores were very high regardless of reporter ($r=.95$).

Youth-parent agreement on the full form *p* factor was moderate ($r=.44$). This correlation compares favorably to the average level of parent-youth agreement on youth psychopathology symptoms reported in studies over a 25-year period ($r=.29$; De Los Reyes et al., 2015). Unlike prior cross-informant studies, which have typically compared average or sum scores across reporters on a particular category of symptoms (e.g., depression symptoms), we examined agreement on a latent general psychopathology factor derived from ratings on a range of symptoms that span multiple disorders. In bifactor models of psychopathology, the general factor loadings are optimized by accounting for the correlations among the symptoms specific to each sub-factor

(Lahey et al., 2021), meaning in some cases (as here) the specific sub-factors are modeled merely as “nuisance” to arrive at the true *p* (general factor) loadings. Consistent with the growing endorsement of a dimensional rather than categorical approach to psychopathology, symptoms are correlated across disorders and dimensions (e.g., internalizing/externalizing) and youth may present with a range of symptoms that do not fit neatly within a single disorder category. Thus, as our results suggest, a latent dimensional approach to capturing a general psychopathology factor may result in better youth-parent agreement on psychopathology than comparing mean scores on disorder-specific assessments.

Higher concordance between adolescent and parent reports of youth psychopathology is not a trivial result, as parent-youth divergence in symptom reports has important clinical implications. For example, in a large community sample, parent-adolescent disagreements in youth symptoms were prospectively associated with youth self-harm, substance use, and referral to mental health services (Ferdinand et al., 2004). Further, in clinical samples, pre-treatment divergence in symptom reports has been associated with less parental involvement in therapy and poorer treatment outcomes among anxious youth (Becker-Haimes et al., 2018; Israel et al., 2007). Importantly, reducing reporter discrepancies over the course of treatment was associated with better treatment outcomes (Becker-Haimes et al., 2018). Based on these prior results, and the degree of concordance of *p* factor scores reported in the present study, additional research on parent-youth agreement on *p* and implications for clinical outcomes is warranted.

A critical barrier to implementing the *p* factor approach to multi-informant assessment of youth psychopathology is informant response burden. To assess the broad range of symptoms from which a *p* factor score is derived takes substantial time and effort (e.g., clinical interview or comprehensive self-report battery). In a multi-informant context, which is typical in youth clinical assessment, this measurement burden is doubled. This type of assessment is likely not feasible in many clinical and research settings. Thus, investigations into possible ways to abbreviate *p* factor assessments are warranted. In the current work, our *p* factor score was derived from over 100 items rated by each reporter. Importantly, our results suggest that using CAT could substantially reduce burden on informants, cutting the average number of items administered by more than half with minimal impact on parent-youth concordance or the clinical validity of *p* factor scores. All differences in correlations between full form and CAT-derived *p* factor scores across reporters were ≤ 0.04 . In previous research with the PNC sample, Moore et al. (2019) demonstrated that a youth *p* factor score could be derived from as few as 10 items and still show clinical validity that is comparable to the full

² The term “quality” when referring to items should be interpreted with caution, because “quality” depends on the purpose of the item. For example, here we assume the GOASSESS was designed to get optimal measurement in a community sample, making extremely difficult (severe) items less informative than items with average severity. However, if the goal of the test were to distinguish among people at very high levels of the trait (e.g., to make a decision about whether an individual is a suicide risk), then items of average severity are “worth” far less.

form version derived from over 100 items. Thus, adaptive assessments that include a broad range of psychopathology symptoms may be a promising and efficient approach to multi-informant assessment of youth psychopathology.

Our results further support the construct validity of a general psychopathology p factor among adolescents. Both youth and parent reports of p (full form and CAT versions) were negatively associated with youth global functioning. It is noteworthy that, in this study, youth global functioning was independently rated by an assessor rather than self-reported by youth and/or parents. This methodological approach addressed important issues recently raised related to mono-informant reports for measurement validation (De Los Reyes et al., 2023; Watts et al., 2022). Some researchers have proposed that the p factor captures impairment, rather than severity of psychopathology (Smith et al., 2020; Watts et al., 2020). Psychopathology, by definition, is characterized by functional impairment and, therefore, a significant association between a general psychopathology factor and impairment is expected. However, the moderate correlations reported in this study ($r_s \approx 0.40$) between p factor scores and assessor-rated functioning suggest that p is capturing more than just impairment. An important direction for future research is to further elucidate the potential mechanisms underlying the general psychopathology factor by including independent assessments of putative processes and examining associations with p factor scores.

Limitations and Future Directions

The present findings should be interpreted in the context of several study limitations that point to future avenues for research. Our results highlight the potential benefits of employing the p factor approach in multi-informant assessments of youth psychopathology and the ability to reduce response burden through adaptive testing. However, a barrier to implementing this strategy is that there is no widely-used, standardized measure of p or its sub-factors. Moore and colleagues (2019) created a publicly available Overall Mental Illness (OMI) screener (full and CAT versions) to measure p using items from the PNC study. The initial psychometrics of this measure are promising, and the results of the present study further support this measure. Additional research aimed at validating and standardizing a comprehensive measure of p across diverse sociodemographic groups would benefit the field.

Use of the general psychopathology p factor in research has increased substantially in recent years and evidence continues to accumulate to support its construct validity. However, there are ongoing debates about p both from a statistical angle and a conceptual perspective (e.g., Lahey et al., 2021; Watts et al., 2020, 2024). Statistically, there are

debates about the use of bifactor models in p factor research and what they can and cannot tell researchers about model validity and causality (e.g., Heinrich et al., 2023; Watts et al., 2024). One concern is that bifactor models will always tend to fit better than alternative models simply because bifactor models are less parsimonious—i.e., there are more parameters (loadings) estimated in the bifactor model—and this superior fit is often erroneously used to justify or argue in favor of the p factor, bifactor model, or both. However, there are many (perhaps most) cases where the lower parsimony (higher complexity) of the bifactor model is necessary to obtain unbiased parameter (loading) estimates. This is because the bifactor model does not impose proportionality constraints on the loadings (Gignac, 2016). For example, Supplemental Figure S1 shows two theoretical bifactor configurations, where the loading pattern in panel “a” fulfills the proportionality assumption of the second-order model (if it were estimated instead of the bifactor), and the loading pattern in panel “b” does not. The model in panel “a” is an example of a case where the bifactor model is unnecessary; the general and specific factor loadings are proportional, so all information could be captured in a second-order model. Any superior fit of a bifactor model in the panel “a” example would be due only to overfitting caused by the added complexity of the bifactor. However, the information in the panel “b” model could not be captured by a second-order model; only a bifactor model could estimate those loadings accurately. This issue is clarified in Moore and Lahey (2022), in which Fig. 1 shows what happens to score estimates—i.e., how wrong scores can be—as the proportionality assumption of second-order models is increasingly violated. We have reason to believe that the true factor structure underlying item responses in the present study is more like panel “b” than panel “a”, because (1) the estimated loadings of the bifactor show clear disproportionality (e.g. general/specific loading ratios ranging from 0.59 to 2.87 for the psychosis spectrum items in youth) and (2) model fit of the bifactor model is far better than the second-order model, beyond what would be expected from the overfitting example given above. Conceptually, as noted above, there is no agreed upon theory or mechanism to explain a single general dimension of psychopathology (but see Lahey et al., 2017). Future research using the p factor will continue to inform these important issues.

The present study included a community sample of youth and parents from a major metropolitan area in the north-eastern United States. It is unclear if our results would generalize to clinical samples or to samples from other areas. Further, for our purposes in this study, the exploratory and confirmatory MIRT models and CAT simulations were all performed on the same sample. Future research could

employ a cross-validation approach in which item calibration and CAT simulations are performed on independent samples.

Conclusion

In sum, we observed moderate parent-youth agreement on p and significant associations between p and youth global functioning across reporters. Further, our results suggest that the degree of parent-youth agreement and associations between p and global functioning are only slightly diminished by reducing reporting burden through adaptive testing procedures. Although there are ongoing debates and unanswered questions about the p factor (Watts et al., 2024), these novel results highlight the promise and potential clinical utility of a multi-informant p factor approach and set the stage for additional investigations of youth p.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10802-024-01223-8>.

Acknowledgements We thank the participants of this study and all the members of the Recruitment, Assessment, and Data Teams whose individual contributions collectively made this work possible. This research was supported by the NIH (RC2 MH089983, MH117014, MH119219, MH117014 and MH089924; K08MH079364; K23MH120437; NIDA supplement to MH089983), the Dowshen Program for Neuroscience, and the Lifespan Brain Institute (LiBI) of Children's Hospital of Philadelphia and University of Pennsylvania Perelman School of Medicine.

Disclosures Ran Barzilay serves on the scientific Advisory Board and holds equity in 'Taliaz Health', with no relevance to the current work.

References

- 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>.
- 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>.
- Barzilay, R., Calkins, M. E., Moore, T. M., Boyd, R. C., Jones, J. D., Benton, T. D., Oquendo, M., Gur, R. C., & Gur, R. E. (2019a). Neurocognitive functioning in community youth with suicidal ideation: Gender and pubertal effects. *The British Journal of Psychiatry*, *215*(3), 552–558.
- Barzilay, R., Patrick, A., Calkins, M. E., Moore, T. M., Wolf, D. H., Benton, T. D., Leckman, J. F., Gur, R. C., & Gur, R. E. (2019b). Obsessive-compulsive symptomatology in community youth: Typical development or a red flag for psychopathology? *Journal of the American Academy of Child & Adolescent Psychiatry*, *58*(2), 277–286. <https://doi.org/10.1016/j.jaac.2018.06.038>.
- Becker-Haimes, E. M., Jensen-Doss, A., Birmaher, B., Kendall, P. C., & Ginsburg, G. S. (2018). Parent–youth informant disagreement: Implications for youth anxiety treatment. *Clinical Child Psychology and Psychiatry*, *23*(1), 42–56. <https://doi.org/10.1177/1359104516689586>.
- Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology*, *6*, 258–276. [https://doi.org/10.1016/0022-2496\(69\)90005-4](https://doi.org/10.1016/0022-2496(69)90005-4).
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431–444. <https://doi.org/10.1177/0146622168200600405>.
- 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. A., Gur, R. C., Hakonarson, H., & Gur, R. E. (2015). The Philadelphia Neurodevelopmental Cohort: Constructing a deep phenotyping collaborative. *Journal of Child Psychology and Psychiatry*, *56*(12), 1356–1369. <https://doi.org/10.1111/jcpp.12416>.
- 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>.
- Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., & Moffitt, T. E. (2014). The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, *2*(2), 119–137. <https://doi.org/10.1177/2167702613497473>.
- Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1–29. <https://doi.org/10.18637/jss.v048.i06>.
- Class, Q. A., Van Hulle, C. A., Rathouz, P. J., Applegate, B., Zald, D. H., & Lahey, B. B. (2019). Socioemotional dispositions of children and adolescents predict general and specific second-order factors of psychopathology in early adulthood: A 12-year prospective study. *Journal of Abnormal Psychology*, *128*(6), 574–584. <https://doi.org/10.1037/abn0000433>.
- De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A., 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., Wang, M., Lerner, M. D., Makol, B. A., Fitzpatrick, O., & Weisz, J. R. (2023). The operations Triad Model and youth mental health assessments: Catalyzing a paradigm shift in measurement validation. *Journal of Clinical Child and Adolescent Psychology*, *52*(1), 19–54. <https://doi.org/10.1080/15374416.2022.2111684>.
- Deutz, M. H., Geeraerts, S. B., Belsky, J., Deković, M., van Baar, A. L., Prinzie, P., & Patalay, P. (2020). General psychopathology and dysregulation profile in a longitudinal community sample: Stability, antecedents and outcomes. *Child Psychiatry & Human Development*, *51*(1), 114–126. <https://doi.org/10.1007/s10578-019-00916-2>.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum.
- Ferdinand, R. F., van der Ende, J., & Verhulst, F. C. (2004). Parent-adolescent disagreement regarding psychopathology in adolescents from the general population as a risk factor for adverse outcome. *Journal of Abnormal Psychology*, *113*(2), 198–206. <https://doi.org/10.1037/0021-843X.113.2.198>.
- Fried, E. I., Greene, A. L., & Eaton, N. R. (2021). The p factor is the sum of its parts, for now. *World Psychiatry*, *20*(1), 69. <https://doi.org/10.1037/0021-843X.113.2.198>.
- Gignac, G. E. (2016). The higher-order model imposes a proportionality constraint: That is why the bifactor model tends to fit better. *Intelligence*, *55*, 57–68. <https://doi.org/10.1016/j.intell.2016.01.006>.

- Gili, M., Castellví, P., Vives, M., de la Torre-Luque, A., Almenara, J., Blasco, M. J., Cebrià, A. I., Gabilondo, A., Pérez-Ara, M. A., Lagares, M.-M.-A., Parés-Badell, C., Piqueras, O., Rodríguez-Jiménez, J. A., Rodríguez-Marín, T., Soto-Sanz, J., Alonso, V., J., & Roca, M. (2019). Mental disorders as risk factors for suicidal behavior in young people: A meta-analysis and systematic review of longitudinal studies. *Journal of Affective Disorders, 245*, 152–162. <https://doi.org/10.1016/j.jad.2018.10.115>.
- Goolsby, J., Rich, B. A., Hinnant, B., Habayeb, S., Berghorst, L., De Los Reyes, A., & Alvord, M. K. (2018). Parent–child informant discrepancy is associated with poorer treatment outcome. *Journal of Child and Family Studies, 27*(4), 1228–1241. <https://doi.org/10.1007/s10826-017-0946-7>.
- Heinrich, M., Geiser, C., Zagorscak, P., Burns, G. L., Bohn, J., Becker, S. P., Eid, M., Beauchaine, T. P., & Knaevelsrud, C. (2023). On the meaning of the P factor in symmetrical bifactor models of psychopathology: Recommendations for future research from the bifactor-(S–1) perspective. *Assessment, 30*(3), 487–507. <https://doi.org/10.1177/10731911211060298>.
- Israel, P., Thomsen, P. H., Langeveld, J. H., & Stormark, K. M. (2007). Parent–youth discrepancy in the assessment and treatment of youth in usual clinical care setting: Consequences to parent involvement. *European Child & Adolescent Psychiatry, 16*(2), 138–148. <https://doi.org/10.1007/s00787-006-0583-y>.
- Jones, J. D., Scott, J. C., Calkins, M. E., Ruparel, K., Moore, T. M., Gur, R. C., & Gur, R. E. (2017a). Correspondence between adolescent and informant reports of substance use: Findings from the Philadelphia Neurodevelopmental Cohort. *Addictive Behaviors, 65*, 13–18. <https://doi.org/10.1016/j.addbeh.2016.09.006>.
- Jones, J. D., Calkins, M. E., Scott, J. C., Bach, E. C., & Gur, R. E. (2017b). Cannabis use, polysubstance use, and psychosis spectrum symptoms in a community-based sample of US youth. *Journal of Adolescent Health, 60*(6), 653–659. <https://doi.org/10.1016/j.jadohealth.2017.01.006>.
- Jones, J. D., Boyd, R. C., Calkins, M. E., Ahmed, A., Moore, T. M., Barzilay, R., Benton, T. D., & Gur, R. E. (2019). Parent-adolescent agreement about adolescents' suicidal thoughts. *Pediatrics, 143*(2), e20181771. <https://doi.org/10.1542/peds.2018-1771>.
- Jones, J. D., Boyd, R. C., Calkins, M. E., Moore, T. M., Ahmed, A., Barzilay, R., Benton, T. D., Gur, R. E., & Gur, R. C. (2021). Association between family history of suicide attempt and neurocognitive functioning in community youth. *Journal of Child Psychology and Psychiatry, 62*(1), 58–65. <https://doi.org/10.1111/jcpp.13239>.
- Kaufman, J., Birmaher, B., Brent, D., Rao, U. M. A., Flynn, C., Moreci, P., Williamson, D., & Ryan, N. (1997). Schedule for affective disorders and Schizophrenia for School-Age Children-Present and Lifetime version (K-SADS-PL): Initial reliability and validity data. *Journal of the American Academy of Child & Adolescent Psychiatry, 36*, 980–988.
- Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry, 62*(6), 617–627. <https://doi.org/10.1001/archpsyc.62.6.617>.
- Krueger, R. F. (1999). The structure of common mental disorders. *Archives of General Psychiatry, 56*(10), 921–926. <https://doi.org/10.1001/archpsyc.56.10.921>.
- 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>.
- Lahey, B. B., Krueger, R. F., Rathouz, P. J., Waldman, I. D., & Zald, D. H. (2017). A hierarchical causal taxonomy of psychopathology across the life span. *Psychological Bulletin, 143*(2), 142–186. <https://doi.org/10.1037/bul0000069>.
- Lahey, B. B., Moore, T. M., Kaczurkin, A. N., & Zald, D. H. (2021). Hierarchical models of psychopathology: Empirical support, implications, and remaining issues. *World Psychiatry, 20*(1), 57–63. <https://doi.org/10.1002/wps.20824>.
- Makol, B. A., De Los Reyes, A., Ostrander, R., & Reynolds, E. K. (2019). Parent-youth divergence (and convergence) in reports of youth internalizing problems in psychiatric inpatient care. *Journal of Abnormal Child Psychology, 47*(10), 1677–1689. <https://doi.org/10.1007/s10802-019-00540-7>.
- Markon, K. E. (2010). Modeling psychopathology structure: A symptom-level analysis of Axis I and II disorders. *Psychological Medicine, 40*(2), 273–288. <https://doi.org/10.1017/S0033291709990183>.
- McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24*(2), 99–114. <https://doi.org/10.1177/01466210022031552>.
- 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>.
- Moore, T. M., & Lahey, B. B. (2022). Issues in estimating interpretable lower order factors in second-order hierarchical models: Commentary on Clark et al. (2021). *Clinical Psychological Science, 10*(3), 593–598. <https://doi.org/10.1177/21677026211035114>.
- Moore, T. M., Calkins, M. E., Satterthwaite, T. D., Roalf, D. R., Rosen, A. F., Gur, R. C., & Gur, R. E. (2019). Development of a computerized adaptive screening tool for overall psychopathology (p). *Journal of Psychiatric Research, 116*, 26–33. <https://doi.org/10.1016/j.jpsychires.2019.05.028>.
- Moore, T. M., Di Sandro, A., Scott, J. C., Lopez, K. C., Ruparel, K., Njokweni, L. J., Santra, S., Conway, D. S., Port, A. M., D'Errico, L., Rush, S., Wolf, D. H., Calkins, M. E., Gur, R. E., & Gur, R. C. (2023). Construction of a computerized adaptive test (CAT-CCNB) for efficient neurocognitive and clinical psychopathology assessment. *Journal of Neuroscience Methods, 109795*. <https://doi.org/10.1016/j.jneumeth.2023.109795>.
- Nolen-Hoeksema, S., & Watkins, E. R. (2011). A heuristic for developing transdiagnostic models of psychopathology: Explaining multifinality and divergent trajectories. *Perspectives on Psychological Science, 6*(6), 589–609. <https://doi.org/10.1177/1745691611419672>.
- Pettersson, E., Lahey, B. B., Larsson, H., & Lichtenstein, P. (2018). Criterion validity and utility of the general factor of psychopathology in childhood: Predictive associations with independently measured severe adverse mental health outcomes in adolescence. *Journal of the American Academy of Child & Adolescent Psychiatry, 57*(6), 372–383. <https://doi.org/10.1016/j.jaac.2017.12.016>.
- Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25–36. <https://doi.org/10.1177/0146621697211002>.
- Reckase, M. D. (2009). *Multidimensional item response theory*. Springer.
- 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., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*(6), 544–559. <https://doi.org/10.1080/00223891.2010.496477>.
- Reise, S., Moore, T., & Maydeu-Olivares, A. (2011). Target rotations and assessing the impact of model violations on the parameters of unidimensional item response theory models. *Educational and Psychological Measurement, 71*(4), 684–711. <https://doi.org/10.1177/0013164410378690>.

- Reise, S. P., Cook, K. F., & Moore, T. M. (2015). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. Reise, & D. Revicki (Eds.), *Handbook of Item Response Theory modeling*. Taylor & Francis.
- Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137–150. <https://doi.org/10.1037/met0000045>.
- Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). 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>.
- Ronald, A. (2019). The psychopathology p factor: Will it revolutionise the science and practice of child and adolescent psychiatry? <https://doi.org/10.1111/jcpp.13063>.
- Satterthwaite, T. D., Elliott, M. A., Ruparel, K., Loughead, J., Prabhakaran, K., Calkins, M. E., Hopson, R., Jackson, C., Keefe, J., Riley, M., Mentch, F. D., Sleiman, P., Verma, R., Davatzikos, C., Hakonarson, H., Gur, R. C., & Gur, R. E. (2014). Neuroimaging of the Philadelphia Neurodevelopmental Cohort. *Neuroimage, 86*, 544–553. <https://doi.org/10.1016/j.neuroimage.2013.07.064>.
- Schweizer, T. H., Snyder, H. R., Young, J. F., & Hankin, B. L. (2020). The breadth and potency of transdiagnostic cognitive risks for psychopathology in youth. *Journal of Consulting and Clinical Psychology, 88*(3), 196–211. <https://doi.org/10.1037/ccp0000470>.
- Shaffer, D., Gould, M. S., Brasic, J., Ambrosini, P., Fisher, P., Bird, H., & Aluwahlia, S. (1983). A children's global assessment scale (CGAS). *Archives of General Psychiatry, 40*(11), 1228–1231. <https://doi.org/10.1001/archpsyc.1983.01790100074010>.
- 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*, 75–98. <https://doi.org/10.1146/annurev-clinpsy-071119-115848>.
- Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*(2), 245–251. <https://doi.org/10.1037/0033-2909.87.2.245>.
- Tackett, J. L., Lahey, B. B., Van Hulle, C., Waldman, I., Krueger, R. F., & Rathouz, P. J. (2013). Common genetic influences on negative emotionality and a general psychopathology factor in childhood and adolescence. *Journal of Abnormal Psychology, 122*(4), 1142–1153. <https://doi.org/10.1037/a0034151>.
- Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*(3), 321–327. <https://doi.org/10.1007/BF02293557>.
- Watts, A. L., Lane, S. P., Bonifay, W., Steinley, D., & Meyer, F. A. (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>.
- Watts, A. L., Makol, B. A., Palumbo, I. M., De Los Reyes, A., Olino, T. M., Latzman, R. D., DeYoung, C. G., Wood, P. K., & Sher, K. J. (2022). How robust is the p factor? Using multitrait-multimethod modeling to inform the meaning of general factors of youth psychopathology. *Clinical Psychological Science, 10*(4), 640–661. <https://doi.org/10.1177/21677026211055170>.
- Watts, A. L., Greene, A. L., Bonifay, W., & Fried, E. I. (2024). A critical evaluation of the p-factor literature. *Nature Reviews Psychology, 3*, 108–122. <https://doi.org/10.1038/s44159-023-00260-2>.
- Williams, E. J. (1959). *Regression analysis*. New York: Wiley.
- Xavier, R. M., Calkins, M. E., Bassett, D. S., Moore, T. M., George, W. T., Taylor, J. H., & Gur, R. E. (2022). Characterizing youth-caregiver concordance and discrepancies in psychopathology symptoms in a US community sample. *Issues in Mental Health Nursing, 43*(11), 1004–1013. <https://doi.org/10.1080/01612840.2022.2099494>.
- Xavier, R. M., Calkins, M. E., Bassett, D. S., Moore, T. M., George, W. T., Taylor, J. H., & Gur, R. E. (2022). Characterizing youth-caregiver concordance and discrepancies in psychopathology symptoms in a US community sample. *Issues in Mental Health Nursing, 43*(11), 1004–1013. <https://doi.org/10.1080/01612840.2022.2099494>.

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.