**Supplement**

**S1. Early Growth and Development Study Recruitment**

Families were eligible to participate if: (a) adoption placement was domestic, (b) the adoption placement occurred within 3 months postpartum (Median age at placement = 2 days; *M =* 5.58, *SD* = 11.32; range = 0–91 days), (c) infants were placed with a genetically unrelated adoptive family, (d) there were no known major birth complications (e.g., extreme prematurity or extensive medical surgeries), and (e) birth and adoptive families read and understood English at the eighth-grade level. Families were recruited between 2003 and 2010 from 45 adoption agencies in the United States.

Most of the adoptive families included mother-father dyads (*n* = 340), while a subset included same-sex dyads (*n* = 21). Approximately 37% of possible birth fathers participated. As noted in Leve et al. (2019), birth father participation was limited by the inability of the agency or the project staff to identify, locate or contact the birth father.

**S2. Observed Behavioral Inhibition at 18 and 27 Months: Stranger Task**

The Lab TAB stranger task occurred first during the home visit, to maximize the unfamiliarity between the child and the interviewer. The task was divided into four 30 s intervals. During the first interval, the interviewer sat quietly and neutrally on the floor across from the child. In the second interval, the interviewer remained neutral while building a tower of stacked cups. In the following 30 s, the interviewer invited the child to play. The interviewer waited 5 s for the child to engage and proceeded to unstack the cups when children did not engage. In the final 30 s interval, the interviewer fully engaged the child by smiling and encouraging the child in play (i.e., building and knocking down the cups). At the end of the two minutes, the interviewer thanked the child and proceeded to the next activity. Parents were present and nearby throughout the task but remained uninvolved.

Trained coders coded children’s inhibited behaviors at 18 and 27 months using the following codes (inter-rater reliabilities provided in parenthesis). Child’s *proximity to caregiver* (ICC = 0.94) was coded when the child initiated physical touch, glanced at, or physically approached the caregiver. Child’s *proximity to the stranger* (i.e., the interviewer; ICC = 0.92) was coded when the child initiated physical touch, glanced at, or physically approached the interviewer. Child’s *inhibition to the stranger* (i.e., the interviewer; ICC = 0.91) was defined as the extent to which the child immediately attempted to increase the distance between themselves and the interviewer and maintained high-intensity avoidant behaviors, such as avoiding looking at the interviewer, turning away from the interviewer, leaning away from the interviewer, and trying to push the interviewer away, throughout the entire task. Child’s *inhibition of exploration* (ICC = 0.79) coded the extent to which the child stared at toys but was reluctant and generally anxious to touch/play with toys, move around and explore, and resisted attempts by caregivers or examiners to engage with toys. Child’s *active exploration* (ICC = 0.93) rated how much the child looked at, played with, and moved around exploring the toys. Interviewers provided their global impression on two behaviors: *inhibition/approach toward the toys* (ICC = 0.93), defined as the latency of child’s attempt to engage or touch the toys and the extent to which the child maintained contact with the toys for the duration of the task, and *child’s fearlessness with toys* (ICC = 0.91)*,* defined as the extent to which the child was not at all afraid of toys.

**S3. Observed Behavioral Inhibition at 4.5 Years: Scary Mask**

The Scary Mask episode was divided into five intervals. In the first interval, the interviewer entered the room wearing a wolf mask and silently looked at the child for 15 seconds. In the second interval, the interviewer walked toward the child and kneeled facing the child. Using a calm voice, the interviewer then invited the child to touch the mask. In the third interval, the interviewer removed the mask, and told the child they were not really a wolf, and that it was just a mask. The interviewer then invited the child to see what the mask looked like up closely, and to touch the nose and teeth on the mask. In the fourth interval, the interviewer invited the child to put the mask on and to show the parent. In the last interval, the interviewer brought the parent into the room and motivated the child to show the mask to the parent. If the child responded negatively to the mask encounter, the interviewer tried to soothe the child for 15 seconds before proceeding. When soothing did not occur, the interviewer terminated the task by putting the mask away, continuing to soothe the child, and bringing the parent back into the room.

Trained, reliable coders used video recordings of the Scary Mask episode to code the following inhibited behaviors (individual ICCs or Kappas are in parenthesis). *Fear* (ICC = 0.807) was coded as the latency to first expression of fear. *Distress* (ICC = 0.803) was a vocal code, and was defined by intensity of fearful vocalizations, ranging from silence to quiet or timid voice, nervous laughter, and engaging the masked stranger with actual comments in a fearful tone. *Approach* (ICC = 0.968) was defined as the extent to which the child moved towards the masked stranger, ranging from no movement to taking one or two hesitant steps, taking only three to four steps, and walking directly to the masked stranger. *Avoidance* (ICC = 0.889) was defined as behaviors initiated by the child to maintain or increase the distance between the child and the masked stranger, ranging from no movement to turning or leaning away, taking one or two steps away, and going to the furthest corner of the room and/or the parent. *Gaze aversion* (ICC = 0.825) coded the extent to which the child avoided looking at the masked stranger and focused on other objects or made no eye contact at all. *Fidgeting* (*kappa* = 0.770) was coded as present or absent and was defined as movement without a purpose. Apart from fidgeting, these behaviors were rated using a 4-point scale ranging from 0 (*behavior is absent*) to 3 (*behavior is strongly or frequently present*). The coding system was adapted from the Lab TAB Scary Mask coding manual (Goldsmith et al., 1993).

**S4. Observed Behavioral Inhibition at 7 Years: Speech Task**

During the Speech task, the interviewer told the child to pretend that someone had just given them $10, which they could use to do whatever they wanted and do something fun. Then the interviewer asked the child to look to the camera for 2 minutes and say what they would do. Prior to starting the task, the interviewer provided some talking points, asking the child if they would do something with friends or family, buy something, or go somewhere. The interviewer also told the child to say as much as they could, and that the interviewer would not be able to speak until the two minutes were over.

If children finished their speech before the two minutes, the interviewer prompted them 30 s after the pause by asking if there was anything else the child wanted to say. If the child continued their speech and paused again, the interviewer waited 30 s again, and repeated the prompt. If the child said “no” to the initial prompt the interviewer waited 15 s then ended the episode. When children did not initially speak, the interviewer waited 90 s and then prompted the child by saying “Can you tell us what you would do?”. If the child responded positively and began speaking, the task proceeded as described above. If the child said “no,” the interviewer waited 30 s and asked “Are you done?”. If the child indicated they were not done, the interviewer prompted them to speak as described above. If the child indicated they were done, the interviewer waited 15 s and ended the task.

The interviewers provided global impressions of children’s nervousness, excitement, fidgeting, whispered speech, comfort, and whether their speech lasted the full 2 minutes. Interviewers provided their impressions using a 4-point scale ranging from 1 (*Very true*) to 4 (*Not true*). Interviewers were trained on the rating system prior to going into the field and were instructed that the items should represent their “impressions,” as intended by the original measure (Weinrott et al., 1981). Thus, these data represent an independent assessment by a single individual, similar to teacher or parent ratings in that regard. Interviewer ratings done in this manner have been shown to correlate with coded observations and child behavioral outcomes in previous research (Weinrott et al., 1981) and have been used in similar ways in published work (e.g., Capaldi et al., 2012).

**S5. Longitudinal Measurement Invariance of Behavioral Inhibition Latent Factors**

In line with previous studies of behavioral inhibition (Fox et al., 2021), we intentionally used different tasks and coded different behaviors across assessments to capture the developmental progression of the behaviorally inhibited phenotype. The number and the definition of behavioral codes across some of these tasks were different, sometimes based on the developmental stage (e.g., toddler interactions with toys vs. preschooler interactions with a masked adult) and other times due to the different nature of the tasks (e.g., interaction with stranger vs. speaking in front of a camera). Based on the lack of one-to-one mapping across indicators for each task, we knew that *full measurement invariance* was not a possibility. However, we identified a few indicators with similar underlying definitions across tasks, accounting for the nature of each task. Specifically, we identified three underlying dimensions of behavioral inhibition across all the codes: 1) inhibition indicated by **physical behaviors/location**, which included *proximity to stranger* (18 and 27 months), *approach* (4.5 years), and *comfort level* (7 years); 2) inhibition indicated by **avoidance/anxiety**, which included *inhibition to stranger* (18 and 27 months)*, mask avoidance* (4.5 years)*,* and *nervous* (7 years); and 3) indicated by **fearlessness/excitement about the task**, which included *fearlessness with objects* (18 and 27 months)*, touch mask* (4.5 years)*,* and *excited* (7 years). We then established *partial measurement invariance* across these indicators using previous guidelines that suggest either *strong* or *strict* invariance are sufficient to establish measurement invariance (Ferrer et al., 2008), and that *partial* measurement invariance may be expected when modeling developmental constructs (Widaman et al., 2010).

First, we set up the original model reported in our initial submission, including the different indicators across age assessments reported in Supplement Table S2, with mean and covariance structures freely estimated (CFI: 0.865; TLI: 0.850; RMSEA: .09; SRMR: 0.05). Second, we tested *weak* partial invariance by constraining the loadings of the indicators identified above to be equal across time, while the rest of the indicators were free to vary (CFI: 0.917; TLI: 0.907; RMSEA: .08; SRMR: 0.10). Modification indices for this model also indicated that the residual variance for the BI factor at 4.5 years should be freely estimated, so this parameter was added to the model. These constraints improved the model fit (Δ*X2* (16) = 257.89, *p* = .001). Third, we tested a *strong* partial invariance model, in which across-time invariance constraints were added on the intercepts of the indicators identified above and to the intercepts of the latent BI factors. While these constraints did significantly worsen the model fit (Δ*X2* (9) = 251.04, *p* = .001), RMSE and SRMR were within the generally acceptable range (CFI: 0.866; TLI: 0.858; RMSEA: .09; SRMR: 0.10) for correctly identifying or rejecting model fit based on our sample size (*N* < 500). While CFI and TLI were below the desired 0.95 cut-off, these indices have been shown to perform poorly in model specifications with sample size < 1000 even when exact model fit is known, and therefore may be too rigorous for applied analyses that seek an approximation of the true process (Sivo et al., 2006). Finally, we tested a *strict* partial invariance model, in which across-time invariance constraints were further added to the unique variances of the indicators identified above and the variances of the latent BI factors. These strict constraints further worsened the model fit (Δ*X2* (8) = 474.39, *p* = .001) such that fit indices were no longer within acceptable range (CFI: 0.770; TLI: 0.764; RMSEA: .12; SRMR: 0.20). Together, these results suggest that our BI tasks show *strong partial measurement invariance*, which is sufficient to model the growth structure of BI latent factors over time.

**S6. Index of BP Psychopathology**

Composite scores were created to reflect BP psychopathology using information on: a) BP psychopathology symptoms (*Symptom Count*), b) BP diagnoses (*Diagnosis Count*), c) BP *Age of Onset* of disorders, and d) The proportion of BP’s *First Degree Relatives* experiencing problems with psychopathology. It was assumed that more symptoms indicate increased severity, and that having symptoms that do not reach the threshold for disorder also constitutes a genetic influence, albeit weaker (e.g., Andrews et al., 1990; Levy et al., 1997). It was also assumed that crossing from symptoms to disorder increases the genetic loading (e.g., Eley, 1997), earlier ages of onset are associated with higher genetic risk (supported particularly for depression, Cadoret et al., 1977; Levinson, 2006), and the number of first-degree relatives confers additional genetic risk, because it indicates higher genetic influence (Sullivan et al., 2000). Genetic risk for psychopathology was separated into three interrelated categories: *Internalizing, Externalizing,* and *Substance Use* problems. Finally, general psychopathology factorwas conceptualized and calculated as a higher-order factor based on the factor scores from these subfactors.

*Internalizing problems* included symptoms, diagnoses, and age of onset from major depression, brief recurrent depression, dysthymia, separation anxiety, adult separation anxiety, social phobia, agoraphobia (with and without panic), panic disorder, specific phobia, and generalized anxiety, assessed via the Composite International Diagnostic Interview (CIDI, Kessler and Ustun, 2004). The proportion of first-degree relatives with internalizing problems was calculated on a single item, “ever been diagnosed with depression or anxiety problems that have been treated or recommended for treatment with medication or counseling”.

*Externalizing problems* included symptoms, diagnoses, and age of onset from conduct disorder and antisocial personality assessed via the Diagnostic Interview Schedule (DIS, Robins et al., 1981). The proportion of first-degree relatives was calculated on the maximum score the BP rated each relative on two items, “ever had a hot temper, been in fights frequently, or been involved in stealing regularly”, and “ever come into contact with the legal system because of things s/he has done (e.g., been arrested, spent time in jail, had a driver’s license revoked)”.

*Substance use* included symptoms, diagnoses, and age of onset of alcohol and drug dependence and abuse and tobacco dependence assessed using the CIDI at the 18-month assessment. The proportion of first-degree relatives (of mother, father, and up to three siblings) with substance use problems was calculated on a single item, “ever had problems with drugs or alcohol (e.g., drank too much or used drugs on a regular basis, got mean while drinking”.

Diagnosis and symptom count from the CIDI and DIS were computed on the following categories:

1. Mood Disorders: Including major depression episode (MJE), dysthymia (DY), recurrent brief depression (RBD), and major depression disorder (MDD)
2. Anxiety Disorders: Including separation anxiety (SA), adult separation anxiety (ASA), agoraphobia (AGO), agoraphobia without panic (AGP), social phobia (SO), panic disorder (PD) specific phobia (SP), and generalized anxiety (G).
3. Substance Use: including Alcohol abuse (AA) and dependence (AD), other drug abuse (DA) and dependence (DD), and nicotine dependence (ND).
4. Externalizing: including Conduct and Antisocial Personality Disorders.

Some inclusion/ exclusion decisions were made for the internalizing scales based on the symptom structure of DSM-IV disorders, which is often quite complex. MJE is a necessary component of major depressive disorder, and the sole symptom of MDD. Therefore, MJE was not included in the diagnosis list. It was however, included as a symptom in the symptom count, as it is the sole symptom for MDD.  MJE and Recurrent Brief Depression have the exact same 9 symptoms. Therefore, we elected to include the nine (RBD) symptoms in the Symptom Count (e.g., only once). Note, however, that because of the overlap of RBD symptoms (which are included in the Symptom Count), functionally, both MJE and its symptoms are represented in the Symptom Count. MJE and Dysthymia have somewhat overlapping symptoms. Five out of the six DY symptoms are repeated in the MJE list verbatim. Three of these overlapping items were perfectly correlated (*r =* 1.0) within participant in our sample, while the others were strongly correlated (*r =* .66 - .88). Therefore, the three perfectly overlapping DY symptoms were not included in the Symptom Count, as to not “double count” them, but the others were retained (including those that were strongly correlated), to ensure that DY symptoms were represented. Finally, Panic Attack is a necessary component of the DSM-IV Criterion A of Panic Disorder, and nearly 50% of the sample reporting having at least a panic attack. However, Panic Disorder criteria A is computed as reporting a panic attack plus reporting at least one attack followed by one month or more of at least one of three symptoms. Thus, the Panic Attack item was excluded from both the Diagnosis Count and Symptom Count.

Principal component analyses (PCA) including each of the four indicators (Symptom Count, Diagnosis Count, Age of Onset, First-Degree Relatives) were conducted for each type of psychopathology (*Internalizing, Externalizing,* and *Substance Use*) to create a more robust index of genetic risk in the adoption sample. Composite scores were created by saving factor scores from these PCAs. Then, a separate PCA was conducted where a second-order composite score was created by saving factor scores of PCAs including each of the three indicators from the first set of PCAs (*Internalizing, Externalizing,* and *Substance Use*). This second-order composite score reflects a “P-factor” indicating more general genetic risk for psychopathology.

**S7. Missing Data and Imputation Methods for BP Psychopathology**

Missing data analyses indicated that across datasets for the Internalizing, Externalizing, and Substance Use indicators, the pattern of missingness was connected and non-monotone (or general). Imputation analyses were conducted in R, using the missMDA and FactoMineR packages, specifically with the imputePCA command. The imputePCA command uses a regularized iterative PCA algorithm to impute missing data (Josse & Husson, 2013), an approach originally designed to address missing data in principal component methods, such as PCA, particularly when dealing with a large number of missing values that may lead to overfitting issues. This approach creates imputations by drawing from iterated conditional models, which is the recommended strategy for general missing patterns (van Buuren, 2018).

Summarized, the regularized iterative PCA algorithm achieves imputation of missing values during the estimation process, where 1) missing values are first replaced by the mean of each variable, 2) PCA is then performed on the completed dataset to estimate the PCA parameters and retain the best fitting number of components (first estimated to be between 1 and 5), and 3) missing values are then imputed with the fitted values, in a “shrunk” imputation manner, where noise variance is estimated by the mean of the last eigenvalues under the assumption that the last dimensions of the PCA are restricted to noise, with the purpose of avoiding instabilities in predictions. Steps 2 and 3 (estimation of PCA parameters and imputation) are repeated (iterated) until convergence. This regularized iterative PCA approach for handling a large number of missing values in PCA has been shown to outperform other methods of imputation and was therefore chosen as the ‘best available method’ to impute missing data and use all possible/available data (i.e., Birth Mothers and Birth Fathers) when creating our genetic risk scores. Indeed, while there are limitations to using these scores based on patterns of missing data and data reduction techniques, we believe we adopted the best available option to create such genetic risk scores that comprehensively reflect genetic overlaps between parents and children. Full information for this imputation method and guidelines for implementation are provided by Josse and Husson (2013).

Furthermore, in the process of computing these genetic risk scores, six different scores were created for Listwise Deletion vs. Imputation, separate for Birth Mothers, Birth Fathers, and both. This strategy was intentionally taken with the purpose of comparing factor loadings between these approaches. As indicated in the table below, we see that the factor loadings were largely not significantly different between these two approaches. For the Birth Father data specifically, which has the highest percent of missingness, the Imputed factor loadings were not significantly different from the Listwise Deletion loadings for the Externalizing or Substance Use indicators. In contrast, the Imputed factor loading was smaller than the Listwise Deletion loading for the Internalizing indicator, suggesting that the effect of Birth Father variables on the principal component was *stronger* when using listwise deletion, potentially biasing the true effect via overestimation. When comparing factor loadings for the Birth Mother and Birth Father data combined, which are the risk scores used in the present analyses, we see no significant differences for the Internalizing and Substance Use indicators, and only a smaller effect of the Imputed factor loading compared to the Listwise Deletion loading, again suggesting potential bias of the true effect via overestimation when using listwise deletion. Based on this assessment, we concluded that it was more appropriate to use the Imputed scores, rather than the Listwise Deletion scores, which could potentially overestimate genetic associations.

**Table S1.** *Factor loadings for genetic risk scores across listwise deletion and imputation methods*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Listwise Deletion** | | | **Imputation** | | |
| *Indicator Source* | **Internalizing Symptoms** | **Externalizing Symptoms** | **Substance Use** | **Internalizing Symptoms** | **Externalizing Symptoms** | **Substance Use** |
| *BM only* | .66 | .80 | .84 | .68 | .81 | .85 |
| *BF only* | .67 | .85 | .84 | .50 | .86 | .86 |
| *BM + BF* | .64 | .87 | .85 | .67 | .78 | .83 |

We also examined how birth father indicators specifically were related to the BI factors. Missing data visualizations indicated that missingness in the following variables could be significantly related to BI levels across the different ages: 1st degree relative with internalizing, externalizing, or substance use disorder, and the number of internalizing, externalizing, or substance use symptoms. We then tested these relations we only found nonsignificant trends between the BI factor at 18 months with missingness in 1st degree relatives with a substance use disorder (*p* = .099), and missingness in the birth father externalizing symptom count (*p* = .087).

**Tables**

**Table S1.** *Description of Study Measures*

|  |  |  |
| --- | --- | --- |
| **Scale Name** | **Scale Description** | **Method** |
| Antisocial Action | Some items were adapted to be more applicable for stay-at-home parents and included statements such as “I tell lies”, and “I don’t pay parking tickets” | Self-Report |
| Beck Anxiety Inventory (BAI) | The BAI is a 21-item measure of anxiety validated in clinical and non-clinical samples. Parents indicated the degree to which they were bothered by symptoms of anxiety (i.e., difficulty breathing, nervous) in the past week using a 4-point scale ranging from 1= *Not at all* to 4= *Severely.* | Self-Report |
| Beck Depression Inventory (BDI) | The BDI is a 21-item measure of depressive symptoms with acceptable reliability and validity. Parents rated symptoms of depression in the past week on a 0-3 scale (the item that measures suicidal ideation was dropped to minimize situations requiring clinical follow-up).Cronbach’s α for adoptive mother and father BDI reports indicated good reliability (α = 0.79 and 0.81, respectively). | Self-Report |
| Composite International Diagnostic Interview | The CIDI is a fully structure interview for the assessment of mental disorders based on criteria and accepted definitions of DSM-IV. This assessment has been validated in clinical and non-clinical samples with excellent reliability. | Interview |
| Composite International Diagnostic Interview - Short Form (CIDI-SF) | The CIDI-SF includes a series of short screening scales to assess the same psychiatric disorders measured by the full CIDI. These scales comprise a fewer number of questions that were determined by stepwise regression analysis to maximize the reproduction of the full CIDI diagnosis. Classification accuracy with these short forms ranged between 77% and 100%. | Interview |
| Family History-Research Diagnostic Criteria | This instrument provides operational criteria for determining a diagnosis based on family history. Relevant diagnoses to the present study included alcoholism, drug abuse, and antisocial personality. This instrument has good reliability based on agreement of prespecified diagnostic standard and diagnostic categories. | Self-Report |

**Table S2.** *Cross-sectional Confirmatory Factor Analyses of Behavioral Inhibition (BI).*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Study Wave** | **Behavioral Codes** | **BI Factor Loadings** | **CFA Fit Indices** | | |
| **CFI** | **TLI** | **SRMR** |
| 18 Months  *Stranger Approach Task* | *Proximity to caregiver* | 0.940 | 0.952 | 0.916 | 0.030 |
| *Proximity to stranger* | 1.076 |  |  |  |
| *Inhibition to stranger* | 0.995 |  |  |  |
| *Inhibition of exploration* | 0.433 |  |  |  |
| *Active exploration* | 0.930 |  |  |  |
| *Fearlessness with objects* | 1.038 |  |  |  |
| *Global inhibition with objects* | 0.947 |  |  |  |
| 27 Months  *Stranger Approach Task* | *Proximity to caregiver* | 0.868 | 0.954 | 0.920 | 0.030 |
| *Proximity to stranger* | 0.952 |  |  |  |
| *Inhibition to stranger* | 0.945 |  |  |  |
| *Inhibition of exploration* | 0.387 |  |  |  |
| *Active exploration* | 0.796 |  |  |  |
| *Fearlessness with objects* | 0.923 |  |  |  |
| *Global inhibition with objects* | 0.808 |  |  |  |
| 4.5 Years  *Scary Mask Task* | *Gaze aversion* | 0.188 | 1.000 | 1.008 | 0.003 |
| *Mask avoidance* | -0.551 |  |  |  |
| *Touch mask* | -0.298 |  |  |  |
| *Approach* | 0.672 |  |  |  |
| 7 Years  *Speech Task* | *Speaking prompts* | 0.83 | 0.992 | 0.981 | 0.020 |
| *Nervous* | -0.68 |  |  |  |
| *Excited* | 0.36 |  |  |  |
| *Low speech* | -0.53 |  |  |  |
| *Comfort level* | 0.96 |  |  |  |
| *Note:* CFI: Comparative Fit Index; TLI: Tucker-Lewis Index; SRMS: Standardized Root Mean Square Residual. All factor loadings *p* > 0.05. | | | | | |

**Supplement References**

Andrews, G., Stewart, G., Allen, R., & Henderson, A. S. (1990). The genetics of six neurotic disorders: a twin study. *Journal of Affective Disorders*, *19*, 23-29.

Browne, M., MacCallum, R. C., Kim, C.-T., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods, 7,* 403–421.

Cadoret, R. J., Leve, L. D., & Devor, E. (1997). Genetics of aggressive and violent behavior. *Psychiatric Clinics of North America*, *20*, 301-322.

Capaldi, D. M., Pears, K. C., Kerr, D. C., Owen, L. D., & Kim, H. K. (2012). Growth in externalizing and internalizing problems in childhood: A prospective study of psychopathology across three generations. *Child Development*, *83*, 1945-1959.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., ... & Moffitt, T. E. (2014). The p factor: one general psychopathology factor in the structure of psychiatric disorders?. *Clinical Psychological Science*, *2*, 119-137.

Eley, T. C., & Plomin, R. (1997). Genetic analyses of emotionality. *Current Opinion in Neurobiology*, *7*, 279-284.

Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order latent growth models. *Methodology, 4*, 22.

Fox, N. A., Buzzell, G. A., Morales, S., Valadez, E. A., Wilson, M., & Henderson, H. A. (2021). Understanding the emergence of social anxiety in children with Behavioral Inhibition. *Biological Psychiatry*, *89*, 681–689.

Goldsmith, H. H., Reilly, J., Lemery, K. S., Longley, S., and Prescott, A. (1993). Preliminary manual for the Preschool Laboratory Temperament Assessment Battery (version 1.0). Technical Report, Department of Psychology, University of Wisconsin-Madison.

Josse, J., & Husson, F. (2012). Handling missing values in exploratory multivariate data analysis methods. *Journal for the Society of French Statistics*, *153*, 79-99.

Lahey, B. B., Rathouz, P. J., Keenan, K., Stepp, S. D., Loeber, R., & Hipwell, A. E. (2015). Criterion validity of the general factor of psychopathology in a prospective study of girls. *Journal of Child Psychology and Psychiatry*, *56*, 415-422.

Levinson, D. F. (2006). The genetics of depression: a review. *Biological psychiatry*, *60*, 84-92.

Levy, F., Hay, D. A., McSTEPHEN, M. I. C. H. A. E. L., Wood, C., & Waldman, I. (1997). Attention-deficit hyperactivity disorder: a category or a continuum? Genetic analysis of a large-scale twin study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *36*, 737-744.

Murray, A. L., Eisner, M., & Ribeaud, D. (2016). The development of the general factor of psychopathology ‘p factor’through childhood and adolescence. *Journal of Abnormal Child Psychology*, *44*, 1573-1586.

Neumann, A., Pappa, I., Lahey, B. B., Verhulst, F. C., Medina-Gomez, C., Jaddoe, V. W., ... & Tiemeier, H. (2016). Single nucleotide polymorphism heritability of a general psychopathology factor in children. *Journal of the American Academy of Child & Adolescent Psychiatry*, *55*, 1038-1045.

Sivo, S. A., Fan, X., Witta, E. L., & Willse, J. T. (2006). The search for" optimal" cutoff properties: Fit index criteria in structural equation modeling. *The Journal of Experimental Education, 74*, 267-288.

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement*, *79*, 310-334.

Sullivan, P. F., Neale, M. C., & Kendler, K. S. (2000). Genetic epidemiology of major depression: review and meta-analysis. *American journal of psychiatry*, *157*, 1552-1562.

Weinrott, M. R., Reid, J. B., Bauske, B. W., & Brummett, B. (1981). Supplementing naturalistic observations with observer impressions. *Behavioral Assessment*, *3*, 151–159.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*, 10-18.

Wright, A. G., Krueger, R. F., Hobbs, M. J., Markon, K. E., Eaton, N. R., & Slade, T. (2013). The structure of psychopathology: toward an expanded quantitative empirical model. *Journal of Abnormal Psychology*, *122*, 281.