**Supplementary Appendix**

**Title: Joint developmental trajectories of internalizing and externalizing problems from mid-childhood to late adolescence and childhood risk factors: Findings from a prospective pre-birth cohort**

**Appendix 1**

**Internalizing and externalizing T-scores:**

The Manual for the Child Behavior Checklist/4-18 and 1991 Profile (Achenbach, 1991) specifies that T-scores should be used for the “broadband” or composite scales (internalizing problems, externalizing problems, total problems, total competence), and raw scores only for the eight syndrome (“narrowband”) scales and three competence scales. The reason for recommending using raw scores for the individual syndrome scales is that T-scores for these scales are truncated at 50 for the lower raw scores, and therefore, raw scores for these individual syndromes can reflect greater differentiation among non-deviant subjects than T-scores can on these scales. T-scores are not truncated for composite scales and using T-scores provides similar results to using raw scores. Therefore, T-scores should be used for composite scales. Our trajectories were significantly different from each other in terms of both mean intercept and slope growth factors (Appendix 3), indicating sufficient variations in the T-scores of internalizing and externalizing scales in our data.

For the composite scales, T-scores allow modelling sex and age differences because T-scores and percentiles are based on separate normative samples of each sex within each age range. T-scores are helpful for indicating the degree to which the child deviates from normative samples of peers on internalizing and externalizing problems.

**Checking bivariate association between predictors:**

The results from the Phi Coefficients values (for 2x2 contingency table) and Cramer’s V (for larger contingency tables) for bivariate association among the characteristic variables (Table S1) show that most of the associations are weak (<0.1). However, some bivariate associations are medium (<0.5 for 1 degrees of freedom), but not strong. Therefore, these associations will not have multicollinearity problems when using them together in a predictive model.

Table S1: Bivariate association statistics (Phi Coefficient or Cramer’s V values) among predicting variables

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 1 | Sex |  |  |  |  |  |  |  |  |  |  |  |
| 2 | Mother's age | -0.0032 |  |  |  |  |  |  |  |  |  |  |
| 3 | Mother's race | 0.0313 | -0.0012 |  |  |  |  |  |  |  |  |  |
| 4 | Mother's highest qualification (Y8) | 0.0535 | 0.2318 | 0.0414 |  |  |  |  |  |  |  |  |
| 5 | Maternal mental health problem (Y8) | -0.0052 | -0.0925 | -0.0814 | 0.0491 |  |  |  |  |  |  |  |
| 6 | Family income (Y5) | -0.0057 | 0.2256 | -0.0609 | 0.2904 | -0.0927 |  |  |  |  |  |  |
| 7 | Marital problems (Y5)  | 0.0177 | -0.0789 | -0.0004 | 0.0519 | 0.1441 | -0.016 |  |  |  |  |  |
| 8 | Father living at home with family (Y5) | -0.0223 | 0.2314 | 0.0515 | 0.145 | -0.1808 | 0.3667 | -0.2036 |  |  |  |  |
| 9 | Parent's tobacco smoking (Y3) | 0.0501 | -0.1883 | -0.0813 | 0.2246 | 0.107 | -0.2018 | 0.0741 | -0.2464 |  |  |  |
| 10 | Parent's smoking other substances (Y3) | 0.0208 | -0.121 | -0.0538 | 0.0787 | 0.0924 | -0.0988 | 0.0887 | -0.1151 | 0.2286 |  |  |
| 11 | Child's temper-tantrum (Y2) | 0.0144 | -0.125 | -0.0051 | 0.0821 | 0.0495 | -0.0388 | 0.0693 | -0.0723 | 0.1116 | 0.0924 |  |
| 12 | Speech ability (Y3)  | -0.1243 | 0.0152 | 0.0059 | 0.0345 | 0.0136 | -0.033 | 0.0037 | -0.0116 | -0.0167 | -0.0052 | 0.0413 |

Note: Phi Coefficient values obtained for 2x2 contingency table and Cramer's V values for larger contingency tables (e.g., 2x3). The variables, which had greater test values (>0.4) and therefore excluded from the predictive analysis, are not presented in this table.

**Multicollinearity check:**

We further conducted tests for multicollinearity issues (Table S2). The results show no indication of multicollinearity issues among the predicting variables used in the logistic regression models in this study. In the results, no tolerance value fall below 0.1 (<0.1 indicates multicollinearity), the smallest tolerance value is 0.7616; all the VIF values are well below 10 (>10 indicates multicollinearity); none of the eigenvalues is close to zero and corresponding condition index is very large, which indicates no issues of multicollinearity among the variables used for the logistic regression models.

Table S2: Tests\* for multicollinearity among variables used for the predictive models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Tolerance | Variance Inflation Factor (VIF) | Eigenvalue | Condition Index |
| Sex | 0.9731 | 1.0277 | 1.1892 | 2.4026 |
| Mother's age | 0.8378 | 1.1936 | 0.9502 | 2.6879 |
| Mother's race | 0.9717 | 1.0291 | 0.8442 | 2.8515 |
| Mother's highest qualification (Y8) | 0.8649 | 1.1562 | 0.7668 | 2.9920 |
| Maternal mental health problem (Y8) | 0.9503 | 1.0523 | 0.6835 | 3.1692 |
| Family income (Y5) | 0.7689 | 1.3005 | 0.6355 | 3.2866 |
| Marital problems (Y5)  | 0.9104 | 1.0984 | 0.4638 | 3.8472 |
| Father living at home with family (Y5) | 0.7616 | 1.3131 | 0.2511 | 5.2282 |
| Parent's tobacco smoking (Y3) | 0.8437 | 1.1853 | 0.1856 | 6.0811 |
| Parent's smoking other substances (Y3) | 0.9192 | 1.0879 | 0.0974 | 8.3942 |
| Child's temper-tantrum (Y2) | 0.9543 | 1.0479 | 0.0544 | 11.2388 |
| Speech ability (Y3)  | 0.9699 | 1.0310 | 0.0138 | 22.3136 |

\*These tests were not available for logistic regression, therefore, these tests were performed using linear regression, because the multicollinearity among the predicting variables concerns about the relationship between predicting variables which are not affected by the form of the dependent variable. Therefore, using linear model will not affect these tests.

**Appendix 2**

**Examining shape of the longitudinal curves of internalizing and externalizing problems:**

Prior to specifying latent classes, we specified unconditional single-class linear, quadratic, and cubic latent growth models as the initial examination of the sample means as well as observed data for internalizing and externalizing variables over time using five time points (ages 5, 8, 10, 14, and 17). The results and spaghetti plots pointed to the quadratic model. Cubic models for externalising problems had a convergence problem in Mplus. It has also been suggested for researchers to use at least five measurement time points for the quadratic growth model and six time points for the cubic growth model to determine which growth model best represents the trend in the data (Whittaker and Khojasteh\_2017). This suggested using a quadratic model for our data with five measurement points.

Theoretically as well as empirically, it has been shown that externalizing problems in general, and internalizing problems, mostly among boys, increase during childhood and start decreasing during adolescence, although girls experience higher rates of internalizing problems over time than boys (Achenbach et al., 1991; Angold et al., 2002; Bongers et al., 2003, 2004; Cicchetti & Rogosch, 2002; Cicchetti & Toth, 1998; Gilliom & Shaw, 2004; Keiley et al., 2003; Leve et al., 2005; Shi & Ettekal, 2020; Shi et al., 2020). Each of these trends indicates that the corresponding psychopathology has a quadratic shape over time. Consistent with these findings, the average T-scores of internalizing and internalizing problems in our data were higher during childhood until age 8/10 and slightly declined during mid-adolescence, reflecting the quadratic trend. Thus, based on the shape of the curve for our observed data as well as the theoretical base, we used quadratic LCGA models for this study.

**Model Selection:**

In addition to model fit indices (smaller values of AIC, BIC, and SABIC scores indicate better model fit), we evaluated classification quality measured by entropy and the average posterior probability of assignment (APPA) for a model selection. Although entropy and APPA are not measures of model fit, they are important statistics to assess the quality and accuracy of the class separation. An entropy value close to 1 indicates accurate classification of a specific model, with generally accepted cutoff values >0.8 (Celeux, G., & Soromenho, G., 1996; Weller et al., 2020). However, there is no agreed cutoff entropy value (Muthen 2008; Weller et al\_2020). Diallo et. al (2017) consider entropy values >0.80 to reflect a high level of class separation and values ≤0.50 to reflect a poor level of class separation.

APPA values are important to assess the quality of assigning individuals to each development trajectory group that best matches their behavior, i.e., individuals are assigned to the group to which their posterior membership probability is largest (Nagin D S\_1999). A high level of classification accuracy is present when individuals have highly differentiated posterior probabilities of membership into the various classes, with APPA (i.e., diagonal values in the posterior probability matrix) close to 1 for all classes (Weller 2020). However, there is no consensus regarding the cutoff value of APPA, some suggesting 0.80 as the acceptable cutoff probability, while others suggest a cutoff value of > 0.90 (Muthén & Muthén, 2000), and probabilities between 0.80 and 0.90 acceptable if the other criteria are met and the model is theoretically supported, but APPA < 0.80 unacceptable even if the other criteria are met (Weller 2020). However, some others suggest the minimum threshold of APPA should be >0.7, which is indicative of a good fit (Blaze 2014; Nagin 2005; Van der Nest, 2020).

In our selected model (five-class solution), the entropy was close to 0.8 (0.76), and all the APPA values >0.80, ranging between 0.8 and 0.89, indicating a good model fit with greater precision of classification.

**BCH method and missing data:**

To deal with missing values for predictors of latent class variables (i.e., trajectory classes), we used a multiple imputation approach using BCH method implemented in Mplus (Asparouhov and Muthen, 2021). This method is preferable to alternative approaches because it avoids heavy numerical integration computations, and it can incorporate both categorical and continuous predictors. This multiple imputation method takes advantage of existing correlations in the data to impute the missing values more accurately.

For this multiple imputation process, we used the BCH weights saved in the first step, where we estimated the latent class model using parallel-process LCGA approach, along with other covariates. BCH weights reduce bias by providing information on correlation between the covariates and the latent class variable in the imputation process. Under the missing at random (MAR) assumption, variables included in the imputation process (those with missing values) should be correlated with covariates which do not have any missing values. To test this assumption, we tested whether the missingness in each predictor depends on the covariates with complete data (i.e., child’s sex at birth, mother’s age at child’s birth, mother’s ethnicity, and father’s ethnicity) using logistic regression. The results showed that mother’s age at child’s birth was significantly associated (p<0.05) with the missingness in all predictors, and mother’s ethnicity, in some predictors. This suggested that the missingness was MAR and justified using multiple imputation with BCH method. Therefore, we included mother’s age and mother’s ethnicity along with BCH weights and other covariates (sex and father’s ethnicity) in the imputation model to impute the missing values in the latent class predictors. As the authors (Asparouhov and Muthen, 2021) suggest that there is important benefit of using a larger number of imputations, we used 100 imputations. All 100 imputed data sets were saved for final estimation of latent class predictors using multinomial logistic regression.

**Appendix 3**

**Optimal class solution:**

Parallel-process LCGA models with varying numbers of classes (i.e., two- to six- class models) were compared, which showed that the AIC, BIC, and SABIC scores decreased as the number of classes increased. The six-class model had the lowest values of these model fit indices (slightly lower than that for five-class model) and therefore the six-class model was the contender for the optimal model. However, the six-class model (55 parameters) was less parsimonious than the five-class model (48 parameters). Also, the Vuong-Lo-Mendell-Rubin Likelihood Ratio Test (VLMR-LRT) and the Bootstrapped Likelihood Ratio Test (BLRT) p-values were not statistically significant (p-values>0.05) for the six-class model (VLMR-LRT p-value= 0.2382; BLRT p-value= 0.6667), suggesting in favour of the model with one less class (i.e., five-class model) (Muthen et al\_2004; Nylund et al\_2007; Van der Nest et al\_2020). Additionally, the six-class model was not qualitatively better than the five-class model in terms of the quality of trajectories. While most trajectories in these two models were similar, the six-class model had two “co-occurring” classes, one similar to but smaller than the high internalizing/ very high externalizing class in the five-class model (6% vs. 10%), and one withwith T-scores within the normal range (50-55) for both internalizing and externalizing problems, which is not meaningfully consonant with the definition of “co-occurring problem scores” as these scores are not problematic.. Although the class-specific sample should not be <5% of the whole sample (Shanahan et al., 2013), borderline class sizes, for example 6%, can raise issues of interpretability – does the class make conceptual sense (Weller et al 2020)? There may also be an issue of power for detecting effects for small classes, particularly when conducting logistic regression of latent class variables on the predictors in the next step of modelling in this analysis. Thus, based on the combination of model selection criteria including the statistical support from the VLMR-LRT and BLRT tests, model parsimony, larger class-specific sample sizes, and substantive meaningfulness and interpretation of each trajectory class, the five-class LCGA model was preferred to the six-class model.

All other two- to five-class models had significant P-values (<0.05) for VLMR-LRT and BLRT, indicating the k-class model had improved model fit compared to the model with one less class, i.e., k-1 class model (Nylund et al., 2007). Although the four-class model was also a candidate for the model selection based on the smaller number of parameters, and significant VLMR-LRT and BLRT p-values, the five-class model identified a clinically meaningful additional co-occurring trajectory class with very high externalizing mean T-scores (>63, i.e, clinically significant) and high internalizing mean T-scores (>59, i.e., sub-clinical), which was not identified by the four-class model. Thus, based on a combination of model selection criteria, including the lower values of model fit indices, significant VMLR LRT and BLRT p-values, quality of classification (APPA>0.8), and the substantive meaning of the trajectories, the five-class model had the best model fit (AIC=131682.59, BIC=131960.05, SaBIC=131807.54; VLMR-LRT p-value=0.0365; BLRT p-value <0.0001; APPA>0.8 for each class), compared to four-class and six-class models. Therefore, five-class model was selected as an optimal class solution for this study.

**Table S3: Means of intercept, slope and quadratic slope growth factors**

**of each class of the five-class joint trajectory model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | **Estimate** | **S.E.** | **Z-value** | **P-Value** |
| **Low-Problems (Low-INT/Low-EXT) Class** |
| IINT | 43.45 | 0.53 | 82.75 | <.001 |
| SINT | -0.32 | 0.11 | -2.93 | 0.003 |
| QINT | -0.02 | 0.01 | -2.41 | 0.016 |
| IEXT | 42.71 | 0.43 | 98.50 | <.001 |
| SEXT | -1.17 | 0.10 | -11.80 | <.001 |
| QEXT | 0.07 | 0.01 | 7.95 | <.001 |
| **Moderate Externalizing (Moderate-EXT/Low-INT) Class** |
| IINT | 47.03 | 0.64 | 73.88 | <.001 |
| SINT | -0.04 | 0.14 | -0.28 | 0.781 |
| QINT | -0.04 | 0.01 | -3.27 | <.001 |
| IEXT | 53.80 | 0.50 | 107.29 | <.001 |
| SEXT | -0.71 | 0.13 | -5.38 | <.001 |
| QEXT | 0.02 | 0.01 | 1.57 | 0.117 |
| **Primary Internalizing (Moderate high-INT/Low-EXT) Class** |
| IINT | 54.63 | 1.09 | 50.02 | <.001 |
| SINT | 0.17 | 0.17 | 0.99 | 0.321 |
| QINT | -0.07 | 0.02 | -4.80 | <.001 |
| IEXT | 49.22 | 1.03 | 47.94 | <.001 |
| SEXT | -1.42 | 0.14 | -9.92 | <.001 |
| QEXT | 0.07 | 0.01 | 5.12 | <.001 |
| **Co-occurring (High-INT/High-EXT) Class** |
| IINT | 59.02 | 0.91 | 65.11 | <.001 |
| SINT | 0.66 | 0.18 | 3.68 | <.001 |
| QINT | -0.11 | 0.01 | -7.51 | <.001 |
| IEXT | 58.66 | 1.19 | 49.29 | <.001 |
| SEXT | -0.40 | 0.19 | -2.16 | 0.031 |
| QEXT | -0.01 | 0.02 | -0.83 | 0.406 |
| **High Co-occurring (Very High-EXT/High-INT) Class** |
| IINT | 59.55 | 1.17 | 50.92 | <.001 |
| SINT | 0.41 | 0.19 | 2.11 | 0.035 |
| QINT | -0.11 | 0.01 | -7.51 | <.001 |
| IEXT | 68.17 | 0.85 | 79.85 | <.001 |
| SEXT | 0.30 | 0.28 | 1.10 | 0.271 |
| QEXT | -0.09 | 0.03 | -3.61 | <.001 |

**Tests for mean intercept and slope differences between trajectory classes:**

To examine whether the internalizing and externalizing classes of the five-class parallel-process model were significantly different to each other to be meaningful, we performed two different tests (Wald chi-squared tests and z-tests) to check the significant differences in mean intercepts (i.e., the initial mean value of trajectories at age 5) as well as mean slopes (i.e., the mean rate of change of trajectories over time) of the trajectory classes. We initially performed Wald chi-squared tests to test the mean intercept differences across the classes of internalizing trajectories (and then externalizing trajectories separately) of the joint trajectory classes. The test result showed that p-values were significant (<0.05) for internalizing as well as externalizing trajectories, indicating there were significant differences among intercepts between internalizing trajectory classes (and between externalizing classes), but this test did not pin-point which classes were different each other in terms of intercepts. Therefore, in the next step, we conducted pair-wise z-tests for all classes to test the intercept differences. These tests showed almost all intercepts were different across internalizing as well as externalizing classes of the five-class parallel-process model, i.e., classes were significantly different (p<.01) in terms of mean intercepts of internalizing trajectories and externalizing trajectories, except the difference between mean intercepts of internalizing trajectories of *Co-occurring* and *High Co-occurring* classes was not significant; however, the mean intercepts of externalizing trajectories were significantly different between these two co-occurring classes (p<.01).

In the same manner, we repeated these tests for the mean slope differences across the internalizing and the externalizing classes of the five-class parallel-process model. The pair-wise test results showed that most of the classes were significantly different in their mean slopes (p<0.01) as well, such as, *Primary Internalizing* with *Low-Problems*, with *Co-occurring* classes; *Low-Problems* with *Co-occurring*, with *High Co-occurring* classes; *High Co-occurring* with *Moderate Externalizing* class. Mean slopes of externalizing trajectories were significantly different between all classes, except between *Primary Internalizing* and *Low-Problems* classes.

**Appendix 4**

**Table S4: Multinomial logistic regression: predictors of joint trajectory classes of five-class model from childhood to mid-adolescent internalizing and externalizing symptoms, comparing each class with another class (i.e., other than *Low-Problems* class), (N=2393)**

|  |
| --- |
|  |
| **Childhood factors (assessed at follow-up years 3-8)** | **Co-occurring vs Primary Internalizing** | **High Co-occurring vs Primary Internalizing** | **Co-occurring vs High Co-occurring** | **Primary Internalizing vs Moderate Externalizing** | **Co-occurring vs Moderate Externalizing** | **High Co-occurring vs Moderate Externalizing** |
|   | OR (95% CI)\* | OR (95% CI) | OR (95% CI) | OR (95% CI) | OR (95% CI) | OR (95% CI) |
| Sex |  |  |  |  |  |  |
| Female (vs Male) | 0.94 (0.63-1.38) | **0.56 (0.37-0.86)** | **1.66 (1.07-2.58)** | 1.15 (0.82-1.61) | 1.08 (0.76-1.53) | **0.65 (0.44-0.96)** |
| Mother's age at child's birtha (unit/SD) | 0.96 (0.78-1.17) | **0.77 (0.61-0.97)** | 1.25 (0.98-1.59) | 1.06 (0.89-1.28) | 1.02 (0.84-1.23) | 0.81 (0.65-1.02) |
| Mother's race  |  |  |  |  |  |  |
| Non-White (vs White)  | **0.56 (0.32-0.99)** | **0.47 (0.25-0.91)** | 1.19 (0.57-2.47) | **3.31 (1.93-5.67)** | **1.86 (1.01-3.43)** | 1.57 (0.78-3.15) |
| Mother's highest qualification (Y8b) |  |  |  |  |  |  |
|  No qualification (vs Highschool/Trade/TAFE)  | 1.20 (0.73-1.97) | 1.29 (0.77-2.17) | 0.93 (0.55-1.57) | 0.99 (0.63-1.55) | 1.18 (0.77-1.81) | 1.27 (0.78-2.05) |
| University degree (vs Highschool/Trade/TAFE)  | 0.72 (0.42-1.26) | **0.36 (0.15-0.86)** | 2.00 (0.78-5.13) | 1.23 (0.78-1.93) | 0.89 (0.52-1.51) | 0.44 (0.19-1.05) |
| Mother's mental health problem (Y 8)  |  |  |  |  |  |  |
| Yes (vs No) | 1.05 (0.67-1.64) | 1.51 (0.94-2.41) | 0.70 (0.42-1.14) | **2.89 (1.82-4.57)** | **3.02 (1.88-4.86)** | **4.34 (2.66-7.08)** |
| Family income per year (Y 5) |  |  |  |  |  |  |
| > $40,000 (vs ≤$40,000) | **0.54 (0.34-0.87)** | **0.35 (0.20-0.61)** | 1.56 (0.85-2.89) | 1.38 (0.92-2.06) | 0.75 (0.48-1.16)  | **0.48 (0.28-0.82)** |
| Marital problem last year (Y5) |  |  |  |  |  |  |
| Yes (vs No) | 1.63 (0.97-2.75) | **1.92 (1.13-3.27)** | 0.85 (0.51-1.40) | 1.44 (0.85-2.45) | **2.35 (1.42-3.87)** | **2.77 (1.68-4.57)** |
| Father living at home with family (Y5) |  |  |  |  |  |  |
| Yes (vs No) | 0.63 (0.38-1.03) | 0.64 (0.38-1.06) | 0.98 (0.59-1.63) | 1.37 (0.86-2.17) | 0.86 (0.55-1.33) | 0.87 (0.55-1.39) |
| Parent's smoking (Y3) |  |  |  |  |  |  |
| Yes (vs No) | 0.74 (0.45-1.21) | 1.17 (0.70-1.95) | 0.63 (0.37-1.06) | 0.93 (0.60-1.43) | 0.68 (0.44-1.06) | 1.08 (0.68-1.73) |
| Parent's smoking other substances (Y3) |  |  |  |  |  |  |
| Yes (vs No) | 1.47 (0.79-2.71) | 0.75 (0.37-1.51) | **1.95 (1.01-3.81)** | 1.07 (0.49-1.14) | 1.56 (0.92-2.66) | 0.80 (0.42-1.53) |
| Child's temper tantrum (Y2) |  |  |  |  |  |  |
| Very/ often (vs none/sometimes) | **2.15 (1.22-3.77)** | **3.16 (1.81-5.54)** | 0.68 (0.40-1.16) | 0.92 (0.54-1.58) | **1.98 (1.20-3.26)** | **2.92 (1.73-4.92)** |
| Speech ability (Y3)  |  |  |  |  |  |  |
| Not clear/clear<75% time (vs Clear>75% time/all time) | 0.60 (0.32-1.13) | 0.78 (0.41-1.47) | 0.77 (0.38-1.57) | **2.58 (1.42-4.67)** | 1.54 (0.77-3.08) | **2.00 (1.01-3.98)** |
|  |  |  |  |  |  |  |

\*Note: Significant (p-values<.05) OR and 95% CI values are bolded.

aMean standardized. bY = follow-up year, which approximately equates to mean age of child

Further analyses were performed to compare trajectory classes with each other (Table S4). Compared with *High Co-occurring* group, children in the *Co-occurring* group were more likely to be girls and have parents smoking other substances. Compared with the *Primary Internalizing*, children in the *Co-occurring* group were more likely to have temper tantrums, White mothers, and low family income. Compared with the *Primary Internalizing* group, children in the *High Co-occurring* group were more likely to be boys, and have temper tantrums, younger mothers, White mothers, parents with marital problems, low family income, and less educated mothers. Compared with the *Primary Externalizing* group, children in the *Co-occurring* group were likely to have temper tantrums, have non-White mothers, have parents with marital problems and have mothers with mental health problems. Compared with *Moderate Externalizing,* the *High Co-occurring* group had children who were more likely to be boys, have temper tantrums, have less/not clear speech, have parents with marital problems, have mothers with mental health problems, and have low family income.

**References**

Achenbach, T. M. (1991). Manual for child behavior checklist 4 18, 1991 profile. *Department of Psychiatry, University of Vermont, Burlington*. <https://lccn.loc.gov/90072106>

Achenbach, T. M., Howell, C. T., Quay, H. C., Conners, C. K., & Bates, J. E. (1991). National survey of problems and competencies among four-to sixteen-year-olds: Parents' reports for normative and clinical samples. *Monographs of the society for research in child development*, i-130. https://doi.org/10.2307/1166156

Angold, A., Erkanli, A., Silberg, J., Eaves, L., & Costello, E. J. (2002). Depression scale scores in 8–17‐year‐olds: effects of age and gender. *Journal of Child Psychology and Psychiatry, 43*(8), 1052-1063. DOI: [10.1111/1469-7610.00232](https://doi.org/10.1111/1469-7610.00232)

Bongers, I. L., Koot, H. M., Van der Ende, J., & Verhulst, F. C. (2003). The normative development of child and adolescent problem behavior. *Journal of abnormal psychology, 112*(2), 179. <https://doi.org/10.1037/0021-843X.112.2.179>

Bongers, I. L., Koot, H. M., Van Der Ende, J., & Verhulst, F. C. (2004). Developmental Trajectories of Externalizing Behaviors in Childhood and Adolescence. *Child development, 75*(5), 1523-1537. <https://doi.org/10.1111/j.1467-8624.2004.00755.x>

Cicchetti, D., & Rogosch, F. A. (2002). A developmental psychopathology perspective on adolescence. *Journal Of Consulting And Clinical Psychology, 70*(1), 6. [https://doi.org/10.1037/0022-006X.70.1.6](https://psycnet.apa.org/doi/10.1037/0022-006X.70.1.6)

Cicchetti, D., & Toth, S. L. (1998). The development of depression in children and adolescents. *American Psychologist, 53*(2), 221. <https://doi.org/10.1037/0003-066X.53.2.221>

Gilliom, M., & Shaw, D. S. (2004). Codevelopment of externalizing and internalizing problems in early childhood. *Development and psychopathology, 16*(2), 313-333. <https://doi.org/10.1017/S0954579404044530>

Keiley, M. K., Lofthouse, N., Bates, J. E., Dodge, K. A., & Pettit, G. S. (2003). Differential risks of covarying and pure components in mother and teacher reports of externalizing and internalizing behavior across ages 5 to 14. *Journal of Abnormal Child Psychology, 31*(3), 267-283. [https://doi.org/10.1023/A:1023277413027](https://doi.org/10.1023/A%3A1023277413027)

Leve, L. D., Kim, H. K., & Pears, K. C. (2005). Childhood temperament and family environment as predictors of internalizing and externalizing trajectories from ages 5 to 17. *Journal of Abnormal Child Psychology, 33*(5), 505-520. <https://doi.org/10.1007/s10802-005-6734-7>

Muthén, B., & Muthén, L. K. (2000, Jun). Integrating person-centered and variable-centered analyses: growth mixture modeling with latent trajectory classes. *Alcohol Clin Exp Res, 24*(6), 882-891. <https://doi.org/10.1111/j.1530-0277.2000.tb02070.x>

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(4), 535-569. DOI:[10.1080/10705510701575396](http://dx.doi.org/10.1080/10705510701575396)

Shanahan, L., Copeland, W. E., Worthman, C. M., Erkanli, A., Angold, A., & Costello, E. J. (2013). Sex-differentiated changes in C-reactive protein from ages 9 to 21: The contributions of BMI and physical/sexual maturation. *Psychoneuroendocrinology, 38*(10), 2209-2217. DOI: [10.1016/j.psyneuen.2013.04.010](https://doi.org/10.1016/j.psyneuen.2013.04.010)

Shi, Q., & Ettekal, I. (2020). Co-occurring trajectories of internalizing and externalizing problems from grades 1 to 12: Longitudinal associations with teacher-child relationship quality and academic performance. *Journal of Educational Psychology*, No Pagination Specified-No Pagination Specified. <https://doi.org/10.1037/edu0000525>

Shi, Q., Ettekal, I., Deutz, M. H. F., & Woltering, S. (2020). Trajectories of pure and co-occurring internalizing and externalizing problems from early childhood to adolescence: Associations with early childhood individual and contextual antecedents. *Developmental Psychology, 56*(10), 1906. <https://doi.org/10.1037/dev0001095>