



Differences in Emotional and Behavioral Problems of Students over Time: A 22-Year Cross-Sectional Cohort Study

W. Alex Mason¹ · Matthew C. Lambert² · Michael H. Epstein²

Accepted: 20 December 2022 / Published online: 29 December 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

Levels of emotional and behavioral problems in children and adolescents demonstrate secular changes over time, warranting ongoing investigation. Prior studies examining secular trends in a range of such problems have been conducted in the U.S. and internationally. Research in this area generally has not fully considered the school setting. This study compared emotional and behavioral problems across two cohorts of students in the U.S. assessed over a 22-year time period as part of measurement development efforts for the *Scales for Assessing Emotional Disturbance Rating Scale* (SAED-RSRS; Epstein et al., 2020). Specifically, analyses drew from data collected via teacher report on matched cohorts of students for the 1998 (data collected from 1996 to 1997; $n=1,148$) and 2020 (data collected from 2016 to 2018; $n=1,148$) editions of the SAED-RS. After establishing measurement invariance across cohorts and testing for gender differences, structural equation modeling revealed statistically significant cohort mean differences on two of the five factors of the SAED-RS, suggesting increases over time in Inability to Learn ($\beta=0.09$, $p=.024$) and Physical Symptoms and Fears ($\beta=0.14$, $p=.005$) that were comparable for girls and boys. There were no statistically significant differences on the remaining factors: Relationship Problems, Inappropriate Behavior, and Unhappiness/Depression. Supplemental item-level tests revealed differences on 8 of the 39 SAED-RS items. Findings suggest increases in specific problem areas that could benefit from ongoing monitoring and targeted interventions to support contemporary students.

Keywords Secular trends · Emotional problems · Behavioral problems · Students

Interest in changing secular trends in emotional and behavioral problems among children and adolescents proliferated at the end of the 20th century over concerns that such problems were increasing (Rutter & Smith, 1995). Indeed, there were substantial increases in psychosocial problems, such as depression and delinquency, among youth from the mid-20th century through the 1970s (Rutter & Smith, 1995). In response, investigators began using increasingly sophisticated research designs involving large cohorts that completed identical instruments over time. Several studies

collected data across two or more cohorts using the *Achenbach System of Empirically Based Assessment* (Achenbach & Rescorla, 2001), obtaining ratings from either parents using the Child Behavioral Checklist (CBCL), teachers using the Teacher Report Form (TRF), or young people themselves using the Youth Self Report (YSR). For example, Achenbach and colleagues (2002, 2003) reported increases in problem scores assessed via the CBCL from 1976 to 1989 and decreases from 1989 to 1999 in U.S. general population children and adolescents. Verhulst et al., (1997) found little evidence of increases in maladaptive functioning among Dutch general population children and adolescents from 1983 to 1993 based on CBCL and TRF assessments. By contrast, Collishaw et al., (2004) reported increases in conduct problems from 1974 to 1986 to 1999, combined with a spike in emotional problems in the 1990s, among United Kingdom general population adolescents based on parent reports using the Rutter scale (Elander & Rutter, 1996) and the Strengths and Difficulties Questionnaire (SDQ; Goodman 1997). In summary, although there

✉ W. Alex Mason
wmason2@unl.edu

¹ Department of Child, Youth, & Family Studies; Nebraska Center for Research on Children, Youth, Families, & Schools, University of Nebraska-Lincoln, Room 36, Carolyn Pope Edwards Hall, Lincoln, United States

² Department of Special Education & Communication Disorders, University of Nebraska-Lincoln, Lincoln, United States

were conflicting findings from these studies based on variations in the informants, measures, and target populations, there was generally evidence for increases in behavioral problems (e.g., conduct problems) up to the 1990s.

Numerous studies have extended the examination of secular trends in child and adolescent emotional and behavioral functioning into the 21st century within various country contexts, including the Netherlands (Duijnhof et al., 2015), Denmark (Henriksen et al., 2012), Sweden (Hagquist, 2010), Finland (Sourander et al., 2012), Scotland (Sweeting et al., 2009), England (Collishaw et al., 2010), Australia (Smart & Sanson, 2008), and the U.S. (Mojtabai et al., 2016; Twenge et al., 2019). As summarized in systematic review articles (Bor et al., 2014; Maughan et al., 2005), these and other studies have revealed a general pattern of stable or slightly decreasing behavioral problems and increasing emotional problems, the latter trend being particularly heightened for girls compared to boys in adolescence. Culture-specific exceptions do exist. For example, Sweeting et al., (2009) reported increases from 1986 to 2006 in “caseness” of social dysfunction as well as anxiety/depression according to the General Health Questionnaire in Scottish adolescent girls and boys. Again, variations in the observed patterns of secular trends are evident based on different study characteristics. In general, since the 2000s, behavioral problems stabilized or, in some cases, decreased, while emotional problems (e.g., depression, anxiety) increased steadily, particularly among adolescent girls.

Compared to studies of time trends in various child and adolescent emotional and behavioral problems conducted elsewhere, those conducted in the U.S., with a few noteworthy exceptions (e.g., Centers for Disease Control and Prevention, 2013), have had a narrower focus on specific types of problems. Although narrowly focused studies are informative, they do not permit the surveillance of a broad range of emotional and behavioral problems in one snapshot. Most such studies have addressed changes in depression and related concerns (Bitsko et al., 2018; Keyes et al., 2019; Mojtabai et al., 2016; Twenge et al., 2019), including suicide (Miron et al., 2019), reflecting the general trend of increases in recent years, particularly among adolescent girls. To illustrate, Twenge et al., (2019) reported a 52% increase from 2005 to 2017 in the prevalence of past-year major depressive episode among girls, aged 12–17 years, in the U.S. Other studies have restricted their focus to externalizing problems (Grucza et al., 2018). Borodovsky et al., (2019) found evidence for a decline from 1997 to 2017 among U.S. general population adolescents in their propensity to engage in substance use, delinquency, and risky sexual behaviors. Wang et al., (2020) examined changes since the turn of the century in crime and safety within school settings across the U.S., reporting decreases in certain problem

indicators (e.g., bullying and victimization) and relative stability in others (e.g., disorder in the classroom).

Significant gaps in knowledge remain. Most studies obtained parent- or self-reports of child and adolescent emotional and behavioral problems. Fewer studies obtained teacher reports, and those that did relied almost exclusively on the TRF (e.g., Henriksen et al., 2012). Teacher reports are informative because they capture student behaviors in the school environment, where most children and adolescents spend large portions of their time. Recent studies using teacher reports in the U.S. are particularly rare, yet such reports could provide the basis for large-scale, national surveillance efforts that monitor the emotional and behavioral functioning of students within an accessible and highly salient developmental setting. Such efforts likely cannot implement longer assessments, such as the TRF, with a heavy administrative burden, but will need to rely on brief, psychometrically-sound instruments that are optimized for efficient repeated administration on a large scale in the school environment. Collecting data from multiple informants (e.g., parents, students, and teachers) is ideal in research studies, since each rater brings their own unique perspective, but doing so is usually not feasible for routine surveillance, which typically relies on querying only particular raters (e.g., teachers) who have access to large numbers of individuals (e.g., students) in a salient context (e.g., schools).

Repeated administration of the same instrument is an important methodological strength of the studies of time trends reviewed above. Still, a significant weakness in prior research is that very few investigators tested and established measurement invariance of their instruments across cohorts. In the current context, measurement invariance ensures that the instrument measures the same fundamental construct in the same way at each administration, which is an essential precondition for testing differences across cohorts. It is possible that the cohort differences observed in previous studies are due, in part, to underlying changes in the measurement characteristics of the instruments. In an exemplary study, Duijnhof et al., (2015) conducted confirmatory factor analyses to establish measurement invariance followed by tests of latent mean differences over a 10-year period, from 2003 to 2013, in Dutch adolescents’ reports of emotional and behavioral problems using the SDQ. They found that problem levels were fairly stable over time. Similar studies are needed, particularly in the U.S. context, since the findings from prior research conducted in other cultural contexts might not generalize to the U.S. Large-scale studies addressing long-term secular trends in multiple youth problems via teacher reports using validated instruments hold promise for advancing knowledge with practical implications.

The present study helps fill gaps in the literature by conducting secondary analyses of data collected from two cohorts over two time periods for the original purpose of establishing the measurement properties of the *Scales for Assessing Emotional Disturbance Rating Scale* (SAED-RS). The two cohorts are labeled according to the publication years of the SAED-RS editions, one published in 1998 (drawing on data collected from 1996 to 1997) and the other published in 2020 (drawing on data collected from 2016 to 2018). Using these data, our prior project analyses compared the two cohorts of students who had school-identified emotional disturbance (Epstein et al., 2022), finding that the 1998 cohort was elevated on relationship problems (e.g., avoidance of peers and/or rejection by them) and inappropriate behavior (e.g., being disrespectful toward teachers or aggressive toward peers) compared to the 2020 cohort, reflecting decreases in these behavior problems over the 22-year time period. Due to the high-risk status of these students, the findings likely do not generalize to the majority of students in the general population. Here, we extend our prior work by replicating these comparisons among students who were not identified with any disability. Findings, therefore, can have greater relevance for most elementary, middle, and high school students across the country. Based on reviewed research, we expected to find support for the general trend of lower behavioral problems and higher emotional problems in the 2020 cohort compared to the 1998 cohort of general population students, with the increase in emotional problems being particularly elevated for girls compared to boys. Our analyses address measurement invariance of the SAED-RS across the two cohorts to ensure that any observed differences reflect true scores and not fluctuations in measurement characteristics of the instrument.

Method

Study Design and Participants

This secondary analysis study capitalizes on a unique opportunity to compare two large cohorts of students (aged 5–18 years) in schools across the U.S. at two time points based on repeated administrations of the *Scales for Assessing Emotional Disturbance Rating Scale* (SAED-RS; Epstein et al., 2020). Data were originally collected to establish the instrument's group norms to which individual students' behaviors subsequently can be compared to facilitate identifying students with emotional disturbance and writing individualized educational and treatment plans. Data were collected from 1996 to 1997 for the SAED-RS 1998 cohort, which included 1,189 students attended schools in 35 states and the District of Columbia, representing all four census regions of

the U.S. (Northeast, Midwest, South, and West). Data were collected from 2016 to 2018 for the SAED-RS 2020 cohort, which included 1,908 students attended schools in 28 states and the District of Columbia, representing all four census regions. Note that in both cohorts, students ranged in age from 5 to 18 years.

For the current analyses, study participants were 2,296 students drawn from the two data sources who did not have any school-identified disabilities (e.g., emotional disturbance, learning disabilities, sensory impairments) as reported by teachers. There were differences in overall sample size and demographic characteristics between the 1998 and 2020 samples; therefore, students from the two samples were matched on school level (e.g., elementary schools), gender, and race/ethnicity using the case-control matching approach in STATA v17. This approach, which selects a random subset of exact matching cases from the two samples, resulted in the inclusion of a cohort of 1,148 students from the 1998 sample and a cohort of 1,148 students from the 2020 sample. The matching variables selected were school level, gender, and race/ethnicity, because research has linked each of these variables to emotional and behavioral disorders of young people (e.g., Anderson & Mayes 2010; Hartung & Lefler, 2019). The resulting 1998 and 2020 cohorts were identical in terms of school level: elementary (63.0%), middle school (22.3%), high school (14.7%); gender: male (51.4%); and race/ethnicity: white/non-Hispanic (69.5%), black/non-Hispanic (15.5%), Hispanic (8.5%), and other race/ethnicity (6.5%). See supplemental figure for a depiction of the sampling strategy and matching process.

Procedures

Data collection efforts contributing to both the 1998 and 2020 editions of the SAED-RS had similar sample recruitment procedures. Educators from across the U.S. were recruited by the SAED-RS authors who contacted potential participants by mail, email, or telephone and asked them to participate in the instrument's norming process. In addition to direct recruitment, some teachers were recruited indirectly through other educators who had been contacted by the SAED-RS authors. Participating teachers were asked to either complete the SAED-RS on all students on their class roster(s) or to select an unbiased sample of their students according to the following procedure: "First, decide how many students you wish to rate. Then, start at either the top or bottom of your class roster and rate every student. Do not skip any student unless you have known this student for less than two months. Stop selecting and rating students when you have reached the number of students you decided to rate." Raters also provided demographic information for rated students (e.g., age in years, grade, race, ethnicity) as

well as for themselves (e.g., gender, race, ethnicity). Periodically, the student demographic information was cross tabulated to determine whether there were insufficient proportions of rated students in any geographical region, school level, gender, or race/ethnicity subcategory, with reference to the U.S. Department of Education statistics on age, gender, and race/ethnicity (<https://www2.ed.gov/programs/osepidea/618-data/static-tables/index.html>). When insufficient proportions were identified, the authors recruited additional teachers who could rate students in a particular subcategory until all demographic subcategories showed proportions that approximated those of the U.S. Department of Education. Data collection for the 1998 edition of the SAED represented “normal educational practice” at the time and was deemed exempt from IRB review. Although the procedures were the same in 2020, additional steps were taken at that time and all procedures, including teacher informed consent prior to completing the surveys, were approved by the Institutional Review Boards of the University of Nebraska-Lincoln, the University of Northern Colorado, and Elon University.

Measure

The SAED-RS is a 45-item teacher rating scale for students aged 5–18 years. The SAED-RS is standardized and norm-referenced, and has five subscales across 39 items that align with the five characteristics in the Individuals with Disabilities Education Act (IDEA) definition of emotional disturbance. *Inability to Learn* items involve problems with academic and other school-related tasks (e.g., “Listening and note-taking skills are weak”). *Relationship Problems* items concern a student’s difficulties in beginning and continuing relationships with peers and teachers (e.g., “Has few or no friends”). *Inappropriate Behavior* items address aggressive and/or disruptive acts toward peers and teachers (e.g., “Disrespectful; defiant of authority”). *Unhappiness or Depression* items measure negative thoughts and related behaviors (e.g., “Experiences little pleasure or joy”). *Physical Symptoms and Fears* items entail indicators of anxiety and physical distress (e.g., “Anxious, worried, tense”). A sixth SAED-RS subscale, *Socially Maladjusted* (6 items), was excluded from this study because it is not a characteristic of IDEA emotional disturbance. Also, several of the social maladjustment items are behaviors that typically occur outside of school and are not necessarily observed by teachers, which further reinforced the decision to delete the items.

Teachers rate each item on a 4-point scale (0 = “not a problem,” 1 = “mild problem,” 2 = “considerable problem,” and 3 = “severe problem”). Subscales are formed by summing the respective items. Cronbach’s alphas for

the analytic sample were high, ranging from 0.81 (Physical Symptoms and Fears) to 0.94 (Inappropriate Behaviors). Indeed, prior research has demonstrated that the scores from both the 1998 and 2020 SAED-RS editions meet acceptable standards for reliability (i.e., test-retest, interrater) and validity (content, construct, convergent; see, e.g., Epstein & Cullinan, 1998; Epstein et al, 2020; Lambert, Cullinan, et al., 2021a). Furthermore, the SAED-RS scores have demonstrated validity evidence based on internal structure including a well-defined confirmatory factor model (Lambert, Cullinan, et al., 2021b; Lambert, Martin, Epstein, Cullinan, & Katsiyannis, 2021) and measurement invariance for students with and without disabilities; that is, the scores reflect the same underlying dimensions of emotional and behavioral problems across race (Lambert, Martin, Epstein, Cullinan, & Katsiyannis, 2021), ethnicity (Lambert, Martin, Epstein, & Cullinan, 2021), age, and gender (Epstein & Cullinan, 1998; Epstein et al., 2020). However, until now, no research has evaluated the measurement invariance of scores between the 1998 and 2020 cohorts for students *without disabilities*.

Data Analysis

The data analysis proceeded in five main steps. First, we implemented an iterative confirmatory factor analysis (CFA) modeling approach that used the items from the SAED-RS to operationalize the five emotional and behavioral problems factors. Second, we examined factor measurement invariance across the 1998 and 2020 cohorts. Third, we evaluated latent mean differences between the two cohorts on each of the five emotional and behavioral problems factors. Fourth, as a series of supplemental analyses to follow-up on the factor tests, we evaluated differences at the item level between the 1998 and 2020 cohorts. Finally, we tested for differences in secular trends by student gender at both the factor and item levels in a series of moderation analyses. Note that as an added sensitivity analysis we replicated the factor-level tests of cohort differences in the 1998 and 2020 total samples without matching.

Confirmatory Factor Analysis

Primary analyses were conducted using Mplus Version 8.6 (Muthén & Muthén, 1998–2021). Using items from the SAED-RS, CFA models operationalized the five emotional and behavioral problems as correlated latent factors (without correlated residual variances between items). The ratings were modeled as ordinal rather than continuous indicators of the latent factors due to the 4-point response scale of the SAED-RS, and the model parameters were estimated using weighted least squares with mean and variance

adjustments (WLSMV). The small amount of missing data (<0.01% of item responses) was handled using a pairwise-present method as is default in Mplus when using the WLSMV estimator.

The five correlated factors CFA model, which included all 39 items, was fit to the data and goodness of fit was interpreted. We evaluated the global fit of the model and also examined structure coefficients (Bentler & Yuan, 2000; Graham et al., 2003) to identify two kinds of potentially problematic items that did not fit the CFA model well: (a) items that were significantly correlated with more than one latent factor and (b) items for which structure coefficients were larger than pattern coefficients (i.e., factor loadings). That is, a problematic item would be more highly correlated with a different factor than the factor onto which it was originally specified to load (e.g., if an item specified to load on Physical Symptoms and Fears was more highly correlated with Unhappiness or Depression).

To assess goodness of fit of the CFA models we used the comparative fit index (CFI; Bentler 1990), the standardized root mean square residual (SRMR; Hu & Bentler 1999), and the root mean square error of approximation (RMSEA; Browne & Cudeck 1993) at its 90% confidence interval. CFI represents the degree of improvement over the worst-fitting model (Boomsma, 2000) and is scaled from 0 to 1, with values closer to 1 indicating better fit. A model with a CFI value greater than or equal to 0.95 demonstrates acceptable fit (Browne & Cudeck, 1993). SRMR and RMSEA are absolute measures of fit and represent the degree of model misfit. Both metrics are reported on a scale of 0 to 1; values closer to zero indicate better fit, with values less than 0.08 and 0.06 indicating acceptable fit for SRMR and RMSEA, respectively (Hu & Bentler, 1999). The 90% confidence interval surrounding the RMSEA point estimate was also used to examine misfit, with an upper limit lower than 0.06 suggesting acceptable fit.

Measurement Invariance

To assess measurement invariance between students in the 1998 cohort and the 2020 cohort, a CFA model was conducted to compare the fit of a multiple-group configural invariance model to the fit of a multiple-group scalar invariance model. A multiple-group configural invariance model allows factor loadings and thresholds to vary across groups, while fixing factor means and scaling parameters for the latent response variables (see Muthén & Muthén 1998–2021). A multiple-group scalar invariance model constrains factor loadings and thresholds across groups, while fixing the factor means and scaling parameters in one group and freely estimating them in the other group. Note that testing the intermediate step of metric invariance in the typical

sequence of measurement invariance tests is not recommended under WLSM estimation with categorical factor indicators because factor loadings and thresholds are most interpretable when either both are fixed or both are freely estimated (Millsap, 2011).

We used the Satorra–Bentler chi-square difference test as well as differences in CFI and RMSEA values (Chen et al., 2008; Cheung & Rensvold, 2002) to determine the tenability of measurement invariance. Although a non-significant chi-square difference test indicates that measurement parameters are statistically equivalent, this test is highly sensitive to sample size (Brannick, 1995) and does not take into account the substantive importance of differences in model parameters (Cheung & Rensvold, 2002). For this reason, we relied more heavily on the change in approximate fit indices, with small changes in CFI and RMSEA values ($\Delta\text{CFI} < 0.01$ and $\Delta\text{RMSEA} < 0.015$) indicating that measurement parameters were substantively equivalent.

Factor-Level Comparisons

After settling on a measurement model that established measurement invariance across the 1998 and 2020 cohorts, we specified a structural equation model (SEM) to test for mean differences between the groups on the five latent emotional and behavioral problems factors. Importantly, latent factor scores, not manifest scores (e.g., raw scores or scale scores), were evaluated because they represent “true scores” corrected for random measurement error and, therefore, are perfectly reliable measures of the constructs. Many scholars consider the use of latent variable measurement models as a gold standard for developmental research (e.g., Bandalos 2018). To test for group mean differences, each latent factor was regressed on a dummy-coded variable representing the contrast between cohorts, with the 1998 cohort specified as the reference group, as well as on dummy-coded gender, school level, and race/ethnicity variables to adjust for the effects of those student characteristics on the latent factor scores. We included the covariates in the analytic model because, after matching, (a) the covariates were orthogonal to the cohort variable (i.e., uncorrelated because of perfect balance across cohorts) and (b) accounted for a significant proportion of variance in the outcomes, so therefore increased the statistical power to detect differences between cohorts because of the resulting reduction in residual (error) variance. Coefficients reported in the results section were “standardized on Y” and are interpreted as the standardized mean difference, either positive (indicating a *higher* mean in the 2020 cohort than in the 1998 cohort) or negative (indicating a *lower* mean in the 2020 cohort than in the 1998 cohort), between students in the two groups. Because the distributions of the five latent factor scores were each

kurtotic and positively skewed, we used non-parametric bootstrapping to estimate empirical standard errors based on 1,000 bootstrapped replications. In addition to statistical significance, which was evaluated at the 0.05 per-test level, we also considered the magnitude of each standardized mean difference according to Cohen's guidelines (Cohen, 1988) for determining Trivial (0.00–0.19), Small (0.20–0.49), Medium (0.50–0.79), and Large (≥ 0.80) effect sizes and with reference to the systematic review of secular trends in behavioral and emotional problems conducted by Bor et al., (2014), who reported absolute values of statistically significant changes over time ranging from $d=0.13$ to 0.52.

Supplemental Item-Level Comparisons

Next, to follow-up on our primary factor-level analyses, we tested item-level differences between the 1998 and 2020 cohorts using ordered logistic regression, including the same set of predictors in each model as for the factor-level mean difference tests. We used an ordered logistic regression modeling approach because items were rated on a four-point scale (i.e., restricted range) and item responses tended to be highly skewed and kurtotic, and therefore the ordered logistic model, which requires fewer underlying assumptions (e.g., monotonicity, proportional odds), was more appropriate for the item-level data. These analyses were conducted in STATA v17. For each statistically significant ordered logistic regression analysis, the logit coefficient, standard error, odds ratio, and Cohen's d effect size are reported in a supplemental table. Logit coefficients and odds ratios represent the log-odds and the proportional decrease (or increase) in the odds, respectively, of students in the 2020 cohort having a higher rating on the item compared to students in the 1998 cohort. An odds ratio of 1.0 would indicate that the odds of having a higher rating were equal between the two cohorts, and therefore the item ratings are equivalent for students in both cohorts. Odds ratios greater than 1.0 would indicate that the odds of having a higher rating for students in the 2020 cohort were greater than for students in the 1998 cohort. Odds ratios less than 1.0 would indicate that the odds of having a higher rating for students in the 2020 cohort were less than for students in the 1998 cohort. For example, an odds ratio of 1.20 would indicate that students in the 2020 cohort were 1.2 times more likely to have a higher rating on the item compared to students in the 1998 cohort.

Although the odds ratio can be considered an effect size measure within the context of ordered logistic regression, the interpretation of its magnitude is not entirely straightforward; therefore, we transformed the odds ratios into Cohen's d effect sizes, which are more familiar and readily interpretable, according to the formula in Borenstein

et al., (2009). The statistical significance of the item-level differences was evaluated after Bonferroni correction at the 0.0013 level to account for multiple comparisons (i.e., 0.05 alpha level/39 tests). Using this conservative level of alpha for each item comparison resulted in a 0.05 probability of making one Type I error and a 0.001 probability of making two Type I errors across the collection of 39 comparisons.

Moderation by Student Gender

To test the extent to which changes across cohorts differed between male and female students, we extended the SEM and ordered logistic regression models to include the interaction between cohort*gender. A statistically significant interaction indicates that change over time was moderated by gender – that is, the regression slope for cohort (i.e., the degree of change between the 1998 and 2020 cohorts) was statistically different for male and female students. For example, a significant interaction might indicate that there was a statistical increase in Relationship Problems for male students between 1998 and 2020, but no statistical difference for female students between 1998 and 2020.

We tested moderation for both primary factor-level comparisons and supplemental item-level comparisons. The significance level for interpreting interactions were set at the same value used to interpret main effects – for factor-level comparisons, the significance level was set at 0.05 and for item-level comparisons, the significance level was set at 0.0013 after Bonferroni Correction. We computed simple effects, the effect of cohort *at a single level of gender*, for any statistically significant interaction in order to understand the differential pattern of change for male and female students. The simple effects were not interpreted if the interaction term was non-significant (because this would indicate that simple effects do not differ for male and female students). For the primary factor-level comparisons, we computed the simple effects and the significance of the simple effects using the approach reported by Aiken & West (1991), which utilizes unstandardized regression coefficients, unstandardized standard errors, and the covariance between the lower-order regression coefficient and the interaction coefficient (which was obtained from the asymptotic covariance matrix generated by the TECH3 command in Mplus). For supplemental item-level comparisons, we used the CONTRAST command in STATA to compute the simple effects and the significance of the simple effects.

Sensitivity Analysis

For the factor-level comparisons, we conducted a sensitivity analysis to evaluate the effect of using the matching procedure to select the analytic sample compared to retaining

Table 1 SAED Subscale Score Means [and 95% Confidence Intervals] by Cohort and Sex

	1998 Cohort			2020 Cohort		
	Male	Female	Overall	Male	Female	Overall
Inability to Learn	4.57 [4.11, 5.03]	3.03 [2.63, 3.43] ^b	3.82 [3.51, 4.13]	5.14 [4.66, 5.61]	3.57 [3.12, 4.02] ^c	4.38 [4.04, 4.71] ^a
Relationship Problems	1.59 [1.37, 1.81]	1.02 [0.84, 1.20] ^b	1.31 [1.17, 1.46]	1.69 [1.45, 1.93]	1.03 [0.84, 1.21] ^c	1.37 [1.22, 1.52]
Inappropriate Behaviors	2.95 [2.51, 3.40]	1.63 [1.30, 1.96] ^b	2.31 [2.03, 2.59]	2.89 [2.48, 3.30]	1.39 [1.12, 1.66] ^c	2.16 [1.91, 2.41]
Unhappiness/Depression	1.25 [1.04, 1.47]	1.13 [0.94, 1.33]	1.20 [1.05, 1.34]	1.25 [1.05, 1.45]	0.99 [0.82, 1.15] ^c	1.12 [0.99, 1.25]
Physical Symptoms and Fears	1.06 [0.89, 1.23]	1.14 [0.97, 1.31]	1.10 [0.98, 1.22]	1.51 [1.29, 1.73]	1.26 [1.06, 1.45]	1.39 [1.24, 1.53] ^a

^a $p < .05$ for mean differences between cohort; ^b $p < .05$ for mean differences between sexes for the 1998 cohort; ^c $p < .05$ for mean differences between sexes for the 2020 cohort

the entire sample and using a statistical adjustment in the SEM to account for the effects of student sex, race, and school level. The sensitivity analysis was conducted with the full sample of 3,097 students – 1,189 students from the 1998 cohort and 1,908 students from the 2020 cohort. Otherwise, the analytic procedure was similar to the primary analysis: each latent factor was regressed on a dummy-coded variable representing the contrast between cohorts, with the 1998 cohort specified as the reference group, as well as on dummy-coded gender, school level, and race/ethnicity variables to adjust for differences in student characteristics between the two cohort groups. Coefficients are reported in text as “standardized on Y” coefficients, which are interpreted as the standardized mean difference between the cohorts, and standard errors were estimated using non-parametric bootstrapping based on 1,000 bootstrapped replications.

Results

Descriptive Statistics

Table 1 reports the descriptive statistics and t-tests for observed SAED-RS subscale score by cohort and by sex. As indicated, observed mean scores were substantively higher in the 2020 cohort for the Inability to Learn and Physical Symptoms and Fears subscale scores, but marginally lower in the 2020 cohort for the Relationship Problems, Inappropriate Behaviors, and Unhappiness/Depression subscale scores. Subscale scores for male students were consistently higher than scores for female students, regardless of cohort, for all problem areas with the one exception of Physical Symptoms and Fears in the 1998 cohort.

Table 2 CFA and Measurement Invariance Results

	χ^2	CFI	SRMR	RMSEA [90% CI ^a]	$\Delta\chi^2$
Five-Factor Model	5489.41	0.955	0.078	0.055 [0.054, 0.056]	–
Measurement Invariance					
Configural Model	5546.55	0.963	0.082	0.051 [0.050, 0.053]	–
Scalar Model	5767.87	0.962	0.082	0.050 [0.049, 0.051]	531.86*

Note. ^a CI = confidence interval; * $p < .001$

Confirmatory Factor Analysis and Measurement Invariance

Table 2 reports the goodness-of-fit indicators for the CFA models. The *initial* model included all 39 SAED-RS items across the five emotional and behavioral problems. This model demonstrated an acceptable fit to the data (CFI = 0.955, SRMR = 0.078, RMSEA = 0.055 [0.054, 0.056]). After establishing an acceptable overall measurement model, multiple-group CFA models were fit to the data to examine measurement invariance between the 1998 and 2020 cohorts. There were only minor differences in the measurement properties between the two cohorts of students, as indicated by the small change in CFI values ($\Delta\text{CFI} = 0.001$) and RMSEA values ($\Delta\text{RMSEA} = 0.001$) and the small, although statistically significant, chi-square difference test ($\Delta\chi^2_{(101)} = 531.86, p < .001$). The small differences in model fit indicated that the measurement parameters were substantively equivalent across the two cohorts, thereby providing evidence consistent with measurement invariance for the CFA model. CFA factor loadings are in Table 3.

Factor-Level Cohort Comparisons

Next, a SEM was specified to evaluate mean differences across the five factors between students in the 1998 and

Table 3 Standardized Factor Loadings by Cohort Group

SAED Item	1998 Cohort	2020 Cohort
Inability to Learn		
Writing is poor	0.84	0.90
Listening is weak	0.90	0.92
Does not complete work	0.91	0.91
Homework skills are weak	0.91	0.89
Math skills are weak	0.76	0.83
Reading skills are weak	0.80	0.86
Lacks interest/motivation	0.95	0.92
Gets distracted	0.92	0.91
Relationship Problems		
Has few friends	0.87	0.91
Rejected by peers	0.86	0.91
Feels picked on	0.86	0.85
Lacks social skills	0.85	0.92
Does not work in groups	0.95	0.96
Avoids interactions	0.70	0.77
Inappropriate Behaviors		
Disrespectful	0.93	0.93
Cruel to peers	0.89	0.91
Verbally abusive	0.90	0.92
Cheats, lies, steals	0.90	0.86
Makes threats	0.93	0.91
Disruptive	0.89	0.85
Physically assaults	0.89	0.90
Uses obscene language	0.83	0.76
Fails to consider consequences	0.95	0.94
Destroys things	0.90	0.85
Unhappiness or Depression		
Lacks self-confidence	0.85	0.88
Feels worthless	0.92	0.91
Little pleasure	0.85	0.89
Often sad	0.82	0.85
Pessimistic about future	0.93	0.89
Limited interaction w/ teacher	0.87	0.83
No longer interested	0.81	0.80
Physical Symptoms and Fears		
Physical discomfort	0.71	0.78
Anxious	0.73	0.85
Afraid of unlikely dangers	0.74	0.79
Overly sensitive feelings	0.90	0.85
Feels excessive guilt	0.79	0.84
Nervous habits	0.76	0.79
Harms self	0.76	0.76
Talks about suicide	0.73	0.73

Note. All factor loadings were statistically significant at $p < .001$

2020 cohorts (i.e., main effects of cohort) while accounting for the effects of gender, school level, and race/ethnicity of the students. This model also fit the data acceptably (CFI=0.957, SRMR=0.074, RMSEA=0.049 [0.048, 0.050]). Students across the two decades differed statistically at the 0.05 significance level on the Inability to Learn ($\beta=0.09$, $p=.024$) and Physical Symptoms and Fears

factors ($\beta=0.14$, $p=.005$); that is, students in the 2020 cohort had higher (more problematic) means on both factors compared to students in the 1998 cohort. Students across the two cohorts did not differ statistically on the Relationship Problems ($\beta=0.03$, $p=.551$), Inappropriate Behavior ($\beta=-0.05$, $p=.354$), or Unhappiness/Depression factors ($\beta=-0.03$, $p=.633$). All main effect standardized mean differences, which ranged from -0.03 to 0.14 , were considered *trivial* according to Cohen's (1988) guidelines for interpreting standardized mean differences but were comparable to effect sizes reported in Bor et al., (2014).

Factor-Level Moderation by Student Gender

To test the extent to which gender moderated the observed factor-level changes across cohorts, the SEM was extended to include a cohort*gender interaction term. When evaluated at the 0.05 significance level, the interaction was non-significant for all five of the behavioral and emotional problems. Statistically non-significant interactions suggest, but do not confirm, that changes across cohorts may be substantively equivalent for male and female students.

Sensitivity Analysis Testing for Factor-Level Cohort Differences Without Matching

The CFA model for the sensitivity analysis fit the data acceptably (CFI=0.955, SRMR=0.069, RMSEA=0.058 [0.056, 0.059]), and measurement invariance appeared to be a tenable assumption with only minor differences in the measurement properties between the two cohorts of students, as indicated by the small change in CFI values ($\Delta\text{CFI}<0.001$) and RMSEA values ($\Delta\text{RMSEA}=0.002$) and the small, although statistically significant, chi-square difference test ($\Delta\chi^2_{(101)}=462.66$, $p<.001$). Substantive results of the sensitivity analysis using the full sample of students (i.e., without using the matching procedure), indicated that there was a significant difference between the 1998 and 2020 cohort for the Inability to Learn factor ($\beta=0.09$, $p=.033$) but not for the Physical Symptoms and Fears factor ($\beta=0.07$, $p=.177$).

Supplemental Item-Level Comparisons

After evaluating factor-level mean differences between students in the 1998 and 2020 cohorts, item-level differences were evaluated between the two cohorts. Eight of the 39 items demonstrated statistically significant differences, when evaluated at the conservative 0.0013 level, between students in the 1998 and 2020 cohorts (see supplemental table). Of the eight items, five were from the Inability to Learn factor, one from the Relationship Problems factor, one from the Inappropriate Behavior factor, and one

from the Physical Symptoms and Fears factor. Five items that demonstrated a significant difference between the two cohorts indicated an increase in problem behavior in 2020, while the other three items indicated a decrease in problem behavior in 2020. Items on the Inability to Learn and Physical Symptoms and Fears factors tended to indicate an increase in problems from 1998 to 2020, while items on the Relationship Problems and Inappropriate Behaviors factors indicated a decrease in problems from 1998 to 2020. Cohen's d effect sizes ranged from -0.38 to 0.25 , indicating that differences were *small* in magnitude and on the higher end of the range of those reported by Bor et al., (2014).

Supplemental Moderation by Student Gender Item-Level Comparisons

In order to test the extent to which gender moderated the change across cohorts at the item-level, the ordered logistic regression models were extended to include a cohort*gender interaction term. When evaluated at the 0.0013 significance level, one item from the Physical Symptoms and Fears factor (*complains of physical discomfort*) exhibited statistically significant moderation by gender. Specifically, the simple effect for male students indicated a statistically significant increase in problems between 1998 and 2020 ($\chi^2_{(1)}=22.93$, $p<.0001$, $OR=2.34$, $d=0.47$), while the simple effect for female students indicated a non-significant slight increase between 1998 and 2020 ($\chi^2_{(1)}=0.15$, $p=.696$, $OR=1.06$, $d=0.03$). The effect sizes indicate that the increase for male students represented a small magnitude (albeit approaching a medium magnitude), while the increase for female students was trivial.

Discussion

Levels of emotional and behavioral problems in children and adolescents demonstrate secular changes over time, warranting ongoing investigation. In this study, comparisons on emotional and behavioral problems across two cohorts of students in the U.S. assessed over a 22-year time period revealed that students in the 2020 cohort (data collected from 2016 to 2018) had higher average scores on both Inability to Learn and Physical Symptoms and Fears compared to students in the 1998 cohort (data collected from 1996 to 1997), suggesting increases in these two types of problems over the time frame of this study. By contrast, there were no statistically significant cohort differences on the remaining factors examined, including Relationship Problems, Inappropriate Behavior, and Unhappiness/Depression, for either girls or boys.

A number of prior studies examining secular trends in a range of emotional and behavioral problems among general population children and adolescents have been conducted outside of the U.S. (e.g., Collishaw et al., 2010; Duinhof et al., 2015; Henriksen et al., 2012; Smart & Sanson, 2008; Sourander et al., 2012; Sweeting et al., 2009). Similar studies conducted in the U.S. tend to have a narrower focus, addressing only particular internalizing problems (e.g., Bitsko et al., 2018; Keyes et al., 2019; Mojtabai et al., 2016; Twenge et al., 2019) or only particular externalizing problems (e.g., Borodovsky et al., 2019; Grucza et al., 2018), which has been informative but has not provided a broad snapshot of multiple problems simultaneously. Taken together, the patterns of findings across these studies indicate increasing emotional problems and stable or slightly decreasing behavioral problems since the turn of the 21st century (for reviews, see Bor et al., 2014; Maughan et al., 2005). Results from this study partly conform to and partly diverge from these general trends. No cohort differences were observed for either the Unhappiness/Depression factor or for any of its component items in either girls or boys. This is in contrast to research conducted in the U.S. demonstrating increases in depression, especially for adolescent girls, in recent years (e.g., Keyes et al., 2019).

The current study relied on teacher reports of unhappiness and depression, which represent emotional states that can be less observable than other types of conditions, whereas most prior studies in the U.S. context have relied on self-reports (e.g., Keyes et al., 2019; Mojtabai et al., 2016; Twenge et al., 2019) and sometimes parent-reports (e.g., Bitsko et al., 2018) of youth depression. Future studies could benefit from obtaining multi-rater data from teachers, parents, and youth, although obtaining such data likely is not feasible for routine large-scale surveillance efforts due to the cost and burden. The studies reviewed above documenting secular increases in emotional problems, particularly in girls, typically have involved adolescents. By contrast, the majority of the students (63%) in the current study were at the elementary school level. The lack of evidence for increases and gender differences in emotional problems may have been due to the relatively young ages of most students in the cohorts.

Physical symptoms and fears showed statistically significant cohort differences that indicated increases over time, which includes certain types of more observable internalizing-related concerns, such as complaints about physical discomfort, showing nervous habits, and talking about suicide. This may be reflective of the increases in youth internalizing symptoms that have been reported previously (Bitsko et al., 2018; Keyes et al., 2019; Mojtabai et al., 2016; Twenge et al., 2019), perhaps resulting from the social pressures engendered by the proliferation of social media and other societal

changes. It is unclear why increases in physical discomfort complaints were more common among boys, which seems to contradict the finding that internalizing-type problems have increased more dramatically among girls; although, it is noteworthy that physical discomfort is a non-specific symptom that could result from a variety of health concerns.

The teacher-report modality of the SAED-RS provided a unique opportunity to examine problems associated with the inability to learn among students. To our knowledge, no prior studies of secular trends in child and adolescent emotional and behavioral problems among general population students have examined this type of factor. Yet, poor functioning in school is highly correlated with a range of emotional and behavioral difficulties (Bradley et al., 2008; Wagner et al., 2005), and schools provide an ideal context in which to routinely surveil students for early-emerging and persistent concerns. Results from this study revealed cohort differences in student learning challenges, suggesting increases over the 22-year observation period. Over the time frame of this study, attempts have been made in the U.S. to raise educational standards for students through such efforts as the No Child Left Behind Act of 2001, the Common Core State Standards Initiative of 2010, the emphasis on STEM (Science, Technology, Engineering and Math) education, and the increase in Charter Schools. Perhaps the most impactful change has been the widespread use of high stakes testing where student performance on select tests are used to make key decisions about students, teachers, school administrators, schools and school districts (Cizek, 2001). These efforts may have resulted in increased scrutiny on the part of teachers in regard to academic performance, contributing to the judgment that students increasingly are having learning difficulties.

It is noteworthy that other expected factor-level cohort differences related to relationship problems and inappropriate behaviors, in addition to unhappiness and depression, were not observed. Although certain item-level (e.g., picked on, fights) cohort differences appear to be consistent with the larger trend of decreasing conduct problems (e.g., bullying, fighting) reported in other studies conducted in the U.S. (e.g., Borodovsky et al., 2019), these results lacked a coherent theme in light of other statistically non-significant item-level differences (e.g., rejected, makes threats, destroys things) and should be interpreted with caution.

The matter of interpreting observed secular changes in child and adolescent emotional and behavioral problems is not without challenges. Our rigorous design, which involved implementing the same measure in two matched cohorts while providing evidence consistent with measurement equivalence, rules out some of the methodological explanations that otherwise might prevail in explaining trends in emotional and behavioral problems over time. As a result,

it is possible that the statistically significant cohort differences observed here reflect true increases in these concerns among students. Still, it is also possible that other factors, such as potential larger-scale changes in the composition of the general student population on characteristics not measured in this study (e.g., earlier puberty, increased screen time), might explain those differences. Further studies better designed to disentangle the potential long-term causal influences driving secular changes in emotional and behavioral problems are needed.

Limitations

There are a few noteworthy limitations of this study. First, recruitment of teachers (and their students) into the norming studies that provided the data for the current analyses was not based on random selection, and the resulting matched samples in the two cohorts are not nationally representative. As previously mentioned, the matching process used to create the sample from the two cohorts resulted in a greater number of elementary school students (63%) than secondary school students. Moreover, only 8.5% of the sample were Hispanic students which did not represent the change in Hispanic students in U.S. schools from 8.8% to 2016 to 17.9% in 2016 U.S. Census Bureau, 2017). However, the data collection efforts did cover schools across all regions of the U.S. Nonetheless, findings still might not be generalizable to general population students nationwide. Second, on a related point, teachers were given the option of rating only a portion of the students in their class in order to reduce the administrative burden. This may have introduced biases related to selection (e.g., the desire to highlight good outcomes in targeted students) and reporting (e.g., heightened awareness of emotional and behavioral problems) that likely have changed over the time frame of this study. Third, although the two cohorts were matched on age, gender, and race/ethnicity, they may have differed on other unmeasured characteristics, such as family socioeconomic status. Such differences represent confounding factors that could serve as alternative explanations for the cohort differences reported herein. Moreover, our matched analyses may further limit generalization to the population, and it is important to highlight that unmatched sensitivity analyses failed to reproduce factor-level cohort differences in physical symptoms and fears (although the pattern of differences was similar). Fourth, neither the 1998 nor the 2020 norming effort permitted linking students with teachers or linking teachers with schools; therefore, the nesting of students within classrooms and classrooms within schools could not be accounted for in the analyses. This precluded the ability to study teacher and school factors, and may have led to underestimated standard

errors and increased Type I errors. Finally, effect sizes for the statistically significant secular changes were trivial to small in magnitude according to Cohen's guidelines. However, similar effect sizes have been reported in studies of secular changes in emotional and behavioral problems among children and adolescents (e.g., Bor et al., 2014). There are currently no accepted standards for determining what might be a clinically meaningful degree of change in long-term trends and such a determination was beyond the scope of this study.

Implications

Notwithstanding these limitations, the current study advances the literature by providing a comprehensive assessment of long-term changes in a range of emotional and behavioral problems, including learning challenges, among children and adolescents in the U.S. over a 22-year time period. Results have policy and practical implications. Schools provide a natural setting for conducting surveillance efforts to monitor trends in emotional and behavioral problems. Although such surveillance efforts do exist, such as the Washington State Healthy Youth Survey initiative, they still do not represent a normative practice in schools across the country and often focus on grades that cover the adolescent years. Routine surveillance using, for example, the SAED-RS could provide an efficient and cost-effective way to monitor the emergence of growing student concerns on a larger scale. Of note, particular attention might be directed toward monitoring the learning challenges and physical symptoms or fears of students. For instance, students might be screened for these challenges in the school setting, which can provide the first point of contact for the referral and care of students needing additional supports in these areas. As ongoing societal changes as well as historical stressors, such as the COVID-19 pandemic, continue to place demands on children and adolescents, routine screening and referral efforts hold particular promise for guiding policies and practices that nimbly respond to emerging trends and provide tailored supports that promote the health and well-being of youth.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10802-022-01014-z>.

Compliance with Ethical Standards

The third author is one of the developers of the SAED, which was used in this research study. He receives royalties from sales of the assessment. The remaining authors have no financial or non-financial interests to disclose. No funding was received to assist with the preparation of this manuscript.

References

- Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms and Profiles*. University of Vermont Research Center for Children, Youth, and Families.
- Achenbach, T. M., Dumenci, L., & Rescorla, L. A. (2002). Ten-year comparisons of problems and competencies for national samples of youth: self, parent, and teacher reports. *Journal of Emotional and Behavioral Disorders, 10*(4), 194–203. <https://doi.org/10.1177/10634266020100040101>.
- Achenbach, T. M., Dumenci, L., & Rescorla, L. A. (2003). Are american children's problems still getting worse? A 23-year comparison. *Journal of Abnormal Child Psychology, 31*(1), 1–11. <https://doi.org/10.1023/A:1021700430364>.
- Aiken, L. S., & West, S. G. (1991). *Multiple regression: testing and interpreting interactions*. Newbury Park, CA: Sage Publications.
- Anderson, E. R., & Mayes, L. C. (2010). Race/ethnicity and internalizing disorders in youth: a review. *Clinical Psychology Review, 30*, 338–348. <https://doi.org/10.1016/j.cpr.2009.12.008>.
- Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Press.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238–246. <https://doi.org/10.1037/0033-2909.107.2.238>.
- Bentler, P. M., & Yuan, K. H. (2000). On adding a mean structure to a covariance structure model. *Educational and Psychological Measurement, 60*(3), 326–339. <https://doi.org/10.1177/00131640021970574>.
- Bitsko, R. H., Holbrook, J. R., Ghandour, R. M., Blumberg, S. J., Visser, S. N., Perou, R., & Walkup, J. T. (2018). Epidemiology and impact of health care provider-diagnosed anxiety and depression among US children. *Journal of Developmental and Behavioral Pediatrics, 39*, 395–403. <https://doi.org/10.1097/DBP.0000000000000571>.
- Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling: A Multidisciplinary Journal, 7*(3), 461–483. https://doi.org/10.1207/S15328007SEM0703_6.
- Bor, W., Dean, A. J., Najman, J., & Hayatbakhsh, R. (2014). Are child and adolescent mental health problems increasing in the 21st century? A systematic review. *The Australian and New Zealand journal of psychiatry, 48*, 606–616. <https://doi.org/10.1177/0004867414533834>.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.
- Borodovsky, J. T., Krueger, R. F., Agrawal, A., & Grucza, R. A. (2019). A decline in propensity toward risk behaviors among U.S. adolescents. *The Journal of adolescent health, 65*, 745–751. <https://doi.org/10.1016/j.jadohealth.2019.07.001>.
- Bradley, R., Doolittle, J., & Bartolotta, R. (2008). Building on the data and adding to the discussion: the experiences and outcomes of students with emotional disturbance. *Journal of Behavioral Education, 17*, 4–23. <https://doi.org/10.1007/s10864-007-9058-6>.
- Brannick, M. T. (1995). Critical comment on applying covariance structure modeling. *Journal of Organizational Behavior, 16*(3), 201–213. <https://doi.org/10.1002/job.4030160303>.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Sage.
- Centers for Disease Control and Prevention. (2013). Mental health surveillance among children — U.S., 2005–2011. *Morbidity and Mortality Weekly Report, 62*(2), 1–35.
- Chen, F., Curran, P., Bollen, K., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models.

- Sociological Methods & Research*, 36(4), 462–494. <https://doi.org/10.1177/0049124108314720>.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5.
- Cizek, G. L. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, 20, 19–27. <https://doi.org/10.1111/j.1745-3992.2001.tb00072.x>.
- Cohen, J. (1988). *Statistical power and analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Collishaw, S., Maughan, B., Goodman, R., & Pickles, A. (2004). Time trends in adolescent mental health. *Journal of child psychology and psychiatry and allied disciplines*, 45(8), 1350–1362. <https://doi.org/10.1111/j.1469-7610.2004.00842.x>.
- Collishaw, S., Maughan, B., Natarajan, L., & Pickles, A. (2010). Trends in adolescent emotional problems in England: a comparison of two national cohorts twenty years apart. *Journal of child psychology and psychiatry and allied disciplines*, 51, 885–894. <https://doi.org/10.1111/j.1469-7610.2010.02252.x>.
- Duinhof, E. L., Stevens, G. W., van Dorsselaer, S., Monshouwer, K., & Vollebergh, W. A. (2015). Ten-year trends in adolescents' self-reported emotional and behavioral problems in the Netherlands. *European child & adolescent psychiatry*, 24, 1119–1128. <https://doi.org/10.1007/s00787-014-0664-2>.
- Elander, J., & Rutter, M. (1996). Use and development of the Rutter parents' and teachers' scales. *International Journal of Methods in Psychiatric Research*, 6, 63–78. [https://doi.org/10.1002/\(SICI\)1234-988X\(199607\)6:2<63::AID-MPR151>3.3.CO;2-M](https://doi.org/10.1002/(SICI)1234-988X(199607)6:2<63::AID-MPR151>3.3.CO;2-M).
- Epstein, M. H., & Cullinan, D. (1998). Scale for Assessing Emotional Disturbance (SAED). PRO-ED.
- Epstein, M., Cullinan, D., Pierce, C., Huscroft-D'Angelo, J., & Wery, J. (2020). Scales for assessing emotional disturbance (3rd ed.). PRO-ED.
- Epstein, M. H., Cullinan, D., Lambert, M. C., Kauffman, J. M., Katsiyannis, A., & Mason, W. A. (2022). A 22-year cross-sectional cohort study of the emotional and behavioral characteristics of students with emotional disturbance. *The Journal of Special Education*. <https://doi.org/10.1177/00224669221109172>
- Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Journal of child psychology and psychiatry and allied disciplines*, 38(5), 581–586. <https://doi.org/10.1111/j.1469-7610.1997.tb01545.x>.
- Graham, J. M., Guthrie, A. C., & Thompson, B. (2003). Consequences of not interpreting structure coefficients in published CFA research: a reminder. *Structural Equation Modeling*, 10(1), 142–153. https://doi.org/10.1207/S15328007SEM1001_7.
- Gruza, R. A., Krueger, R. F., Agrawal, A., Plunk, A. D., Krauss, M. J., Bongu, J., Cavazos-Rehg, P. A., & Bierut, L. J. (2018). Declines in prevalence of adolescent substance use disorders and delinquent behaviors in the USA: a unitary trend? *Psychological Medicine*, 48, 1494–1503. <https://doi.org/10.1017/S0033291717002999>.
- Hagquist, C. (2010). Discrepant trends in mental health complaints among younger and older adolescents in Sweden: an analysis of WHO data 1985–2005. *The Journal of adolescent health*, 46, 258–264. <https://doi.org/10.1016/j.jadohealth.2009.07.003>.
- Hartung, C. M., & Lefler, E. K. (2019). Sex and gender in psychopathology: DSM-5 and beyond. *Psychological Bulletin*, 145, 390–409. <https://doi.org/10.1037/bul0000183>.
- Henriksen, J., Nielsen, P. F., & Bilenberg, N. (2012). New danish standardization of the child Behaviour Checklist. *Danish medical journal*, 59, A4462.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>.
- Keyes, K. M., Gary, D., O'Malley, P. M., Hamilton, A., & Schulenberg, J. (2019). Recent increases in depressive symptoms among US adolescents: trends from 1991 to 2018. *Social psychiatry and psychiatric epidemiology*, 54, 987–996. <https://doi.org/10.1007/s00127-019-01697-8>.
- Lambert, M. C., Cullinan, D., Epstein, M. H., & Martin, J. (2021a). Differences between students with emotional disturbance, learning disabilities, and without disabilities on the five dimensions of emotional disturbance. *Journal of Applied School Psychology*. <https://doi.org/10.1080/15377903.2021.1895399>
- Lambert, M. C., Cullinan, D., Epstein, M. H., & Martin, J. (2021b). Factor structure of the Scales for Assessing Emotional Disturbance – 3 Rating Scale for student identified with emotional disturbance. *Behavioral Disorders*. <https://doi.org/10.1177/01987429211012328>
- Lambert, M. C., Martin, J., Epstein, M. H., & Cullinan, D. (2021). Differential item functioning of the Scales for Assessing Emotional Disturbance-3 for White and Hispanic students. *Journal of Emotional and Behavioral Disorders*. <https://doi.org/10.1177/1063426620988658>
- Lambert, M. C., Martin, J., Epstein, M. H., Cullinan, D., & Katsiyannis, A. (2021). Differential item functioning of the Scales for Assessing Emotional Disturbance-3 for White and African American students. *Psychology in the Schools*, 58(3)1–16. <https://doi.org/10.1002/pits.22463>
- Maughan, B., Iervolino, A. C., & Collishaw, S. (2005). Time trends in child and adolescent mental disorders. *Current opinion in psychiatry*, 18, 381–385. <https://doi.org/10.1097/01.yco.0000172055.25284.f2>.
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge/Taylor & Francis Group.
- Miron, O., Yu, K. H., Wilf-Miron, R., & Kohane, I. S. (2019). Suicide rates among adolescents and young adults in the U.S., 2000–2017. *Journal Of The American Medical Association*, 321, 2362–2364. <https://doi.org/10.1001/jama.2019.5054>.
- Mojtabai, R., Olfson, M., & Han, B. (2016). National Trends in the prevalence and treatment of Depression in Adolescents and Young adults. *Pediatrics*, 138, e20161878. <https://doi.org/10.1542/peds.2016-1878>.
- Muthén, L. K., & Muthén, B. O. (1998). – 2021). *Mplus user's guide* (8th ed.). Muthén & Muthén.
- Pew Research Center (2019). Two-thirds of Americans support marijuana legalization. Retrieved, & March (2020). 28, from <https://www.pewresearch.org/facttank/2019/11/14/americans-support-marijuana-legalization/>
- Rutter, M., & Smith, D. J. (1995). *Psychosocial disorders in young people: Time trends and their causes*. Chichester, West Sussex: John Wiley & Sons, Ltd.
- Smart, D., & Sanson, A. (2008). Do australian children have more problems than twenty years ago? *Family Matters*, 79, 50–57.
- Sourander, A., Koskelainen, M., Niemelä, S., Rihko, M., Ristkari, T., & Lindroos, J. (2012). Changes in adolescents mental health and use of alcohol and tobacco: a 10-year time-trend study of finnish adolescents. *European child & adolescent psychiatry*, 21, 665–671. <https://doi.org/10.1007/s00787-012-0303-8>.
- Sweeting, H., Young, R., & West, P. (2009). GHQ increases among Scottish 15 year-olds 1987–2006. *Social Psychiatry and Psychiatric Epidemiology: The International Journal for Research in Social and Genetic Epidemiology and Mental Health Services*, 44, 579–586. <https://doi.org/10.1007/s00127-008-0462-6>
- Twenge, J. M., Cooper, A. B., Joiner, T. E., Duffy, M. E., & Binau, S. G. (2019). Age, period, and cohort trends in mood disorder indicators and suicide-related outcomes in a nationally representative dataset, 2005–2017. *Journal of abnormal psychology*, 128, 185–199. <https://doi.org/10.1037/abn0000410>.

- U.S. Census Bureau, & Washington, D. C. (2017). *School enrollment of the hispanic population: two decades of growth*. I AM NOT CERTAIN ABOUT THE REFERENCE).
- Verhulst, F. C., van der Ende, J., & Rietbergen, A. (1997). Ten-year time trends of psychopathology in dutch children and adolescents: no evidence for strong trends. *Acta Psychiatrica Scandinavica*, 96, 7–13. <https://doi.org/10.1111/j.1600-0447.1997.tb09898.x>.
- Wagner, M., Kutash, K., Duchnowski, A. J., Epstein, M. H., & Sumi, C. (2005). The children and youth we serve: a national picture of the characteristics of students with emotional disturbance receiving special education. *Journal of Emotional and Behavioral Disorders*, 13, 79–96. <https://doi.org/10.1177/10634266050130020201>.
- Wang, K., Chen, Y., Zhang, J., & Oudekerk, B. A. (2020). *Indicators of School Crime and Safety: 2019* (NCES 2020-063/NCJ 254485).

National Center for Education Statistics, U. S. Department of Education, and Bureau of Justice Statistics, Office of Justice Programs, U.S. Department of Justice.

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.