Running head: CONSEQUENCES OF BULLYING VICTIMIZATION

Quasi-experimental evidence on short and long-term consequences of bullying victimization: A meta-analysis

Tabea Schoeler, PhD [1*]

Lauren Duncan, BSc. [1]

Charlotte M. Cecil, PhD [2]

George B. Ploubidis, PhD [3]

Jean-Baptiste Pingault, PhD [1,4]

[1] Division of Psychology and Language Sciences, University College London, London, England

[2] Department of Psychology, King's College London, London, England

[3] Centre for Longitudinal Studies, University College London, London, England

[4] Social Genetic and Developmental Psychiatry Centre, King's College London, London, England

*Corresponding author: Department of Clinical, Educational and Health Psychology, Division of Psychology and Language Sciences, University College London, 26 Bedford Way, London WC1H 0AP, Email: t.schoeler@ucl.ac.uk

Abstract

Exposure to bullying victimization is associated with a wide-range of short and long-term adverse outcomes. However, the extent to which these associations reflect a causal influence of bullying victimization remains disputed. Here, we aimed to provide the most stringent evidence regarding the consequences of bullying victimization by meta-analysing all relevant Quasi-Experimental (QE) studies. Multilevel random effects models and meta-regression were employed to (i) estimate the pooled QE-adjusted effect size (Cohen *d*) for bullying victimization on outcomes and to (ii) evaluate potential sources of heterogeneity. A total of 16 studies were included. We derived 101 QE-estimates from three different methods (twin design, fixed effects analysis, and propensity score matching) for three pools of outcomes (internalizing symptoms, externalizing symptoms, academic difficulties). QE-adjusted effects were small for internalizing symptoms ($d_{adjusted}$=0.27, 95%CI 0.05;0.49), and smaller for externalizing symptoms ($d_{adjusted}$=0.15, 95%CI 0.10;0.21) and academic difficulties ($d_{adjusted}$=0.10, 95%CI 0.06; 0.13). Accounting for a shared rater effect between the exposure and the outcome further reduced the effect for internalizing ($d_{non\text{-}shared\ rater}$=0.14, 95%CI 0.05;0.23) and externalizing symptoms ($d_{non\text{-}shared\ rater}$=0.06, 95%CI 0.01;0.11). Finally, the adverse effects declined in the long-term, most markedly for internalizing symptoms ($d_{long\text{-}term}$=0.06, 95%CI -0.01;0.13). Based on the most stringent evidence available to date, findings indicate that bullying victimization may causally impact children's wellbeing in the short-term, especially anxiety and depression levels. The reduction of adverse effects over time highlights the potential for resilience in individuals who have experienced bullying. Secondary preventive interventions in bullied children should therefore focus on resilience and address children's pre-existing vulnerabilities.

Key words: Quasi-Experimental; bullying victimization; mental health; epidemiology; meta-analysis

Public Significance Statement

This meta-analysis of Quasi-Experimental studies suggests that bullying victimization leads to poorer developmental outcome in the short-term, including higher internalizing and externalizing symptoms and reduced academic achievement. These adverse effects diminish in the long-term, highlighting the potential for resilience in individuals that experienced bullying. In addition to tackling bullying, interventions should therefore address the immediate adverse consequences of bullying victimization, while fostering resilience in victimized children.

Bullying Victimization and Adverse Outcomes

Bullying victimization has commonly been defined as intentional and repeated direct (physical) or indirect (verbal/mental) aggression on an individual by peers, characterized by power imbalance (Dan, 1993). About 13% of 11 to 15 year olds worldwide are victims of bullying (Craig et al., 2009). High prevalence rates imply that bullying victimization may constitute a major public health concern, especially since the occurrence of bullying victimization may have increased over the past decades (Cosma, Whitehead, Neville, Currie, & Inchley, 2017). Extensive observational research reports widespread associations between bullying victimization and children's short and long-term outcomes [cf. meta-analytical evidence as summarized in (Hawker & Boulton, 2000; Holt et al., 2015; Moore et al., 2017; Nakamoto & Schwartz, 2010; Reijntjes et al., 2011; Reijntjes, Kamphuis, Prinzie, & Telch, 2010; Tsaousis, 2016; Ttofi, Farrington, Lösel, & Loeber, 2011; van Dam et al., 2012; van Geel, Goemans, & Vedder, 2016; Van Geel, Vedder, & Tanilon, 2014)]. For example, associations have been reported between bullying victimization and internalizing symptoms such as depression and anxiety (Brendgen & Poulin, 2018; Eastman et al., 2018; Lee & Vaillancourt, 2018), which may arise from increased emotional dysregulation, reduction in self-esteem, withdrawal from social contacts, loneliness and altered stress responses following the bullying experience (Hamilton, Newman, Delville, & Delville, 2008; Nansel et al., 2001; O'Moore & Kirkham, 2001). Associations have also been found between bullying victimization and externalizing symptoms, such as delinquency, substance use and conduct problems (Eastman et al., 2018; Evans, Smokowski, Rose, Mercado, & Marshall, 2018; Kretschmer et al., 2018; Quinn & Stewart, 2018). Such problems could occur as a result of higher states of arousal and feelings of anger, hostility and frustration induced by the experience of being bullied (Camodeca & Goossens, 2005; Sigfusdottir, Gudjonsson, & Sigurdsson, 2010; Woods & White, 2005). Those findings implicate bullying victimization as an adverse and stressful life event that may have long-lasting effects on well-being and developmental outcomes. However, the extent to which these associations truly reflect the

consequences of bullying victimization remains uncertain. Associations may partly or fully result from confounding. Quasi-Experimental (QE) designs constitute a powerful set of tools to strengthen conclusions on causality within observational designs (Bärnighausen et al., 2017; Jaffee, Strait, & Odgers, 2012).

In the following section, we introduce the counterfactual framework for causal inference and its application through QE-designs. For the purpose of this study, we defined QE-designs as those designs most commonly employed by observational studies using advanced methods of causal inference (Listl, Jürges, & Watt, 2016; Pingault et al., 2018; Rockers, Røttingen, Shemilt, Tugwell, & Bärnighausen, 2015; Stuart, 2010), including (a) difference studies, (b) propensity score matching, (c) regression discontinuity analysis, (d) instrumental variable analyses and (e) interrupted time series studies. We then present the results from a meta-analysis investigating the QE-adjusted effects of bullying victimization on developmental outcomes. We restricted our analysis to QE-studies to derive the most stringent estimates of the effects of bullying victimization.

Causal Inference and Bullying Victimization

Traditionally, causality has been evaluated by assessing several features of an identified association, such as dose-response relationships, directionality and confounding (Hill, 1965). For example, bullying victimization predicts subsequent adverse outcomes as a function of chronicity and severity (Baldwin et al., 2016; Bouffard & Koeppel, 2014; Esbensen & Carson, 2009; Schreier et al., 2009), implicating the presence of dose-response relationships. Scrutinizing directionality, prospective studies point either towards direct effects (bullying victimization as a predictor of subsequent mental health problems) or reciprocal effects (bullying victimization that is linked to both preceding and subsequent mental health problems) (Burke, Sticca, & Perren, 2017; Davis et al., 2017; Lester, Dooley, Cross, & Shaw, 2012; Sweeting, Young, West, & Der, 2006). While both pathways are consistent with the notion of causality, the main challenge – i.e. confounding –

remains. For example, there is consistent evidence that bullied children differ from their non-bullied counterparts. Namely, bullied children are more likely to be affected by personal and family risk factors for mental health, such as pre-existing mental health vulnerabilities, socioeconomic and migration status (Delprato, Akyeampong, & Dunne, 2017; Wong & Schonlau, 2013).

Studies have traditionally dealt with confounding by controlling for plausible pre-defined confounding variables in statistical models. However, statistical adjustment cannot account for unobserved sources of confounding, such as genetic factors. Genetically informative studies have found that children who experience bullying may indeed be more likely to carry genetic vulnerability for mental health problems, which, in turn, increase the likelihood of being bullied, thereby generating confounding (Ball et al., 2008; Shakoor et al., 2015). As an example, evidence suggests that underlying shared genetic factors may explain a substantial part of the association between bullying victimization and adverse outcomes, e.g. 93% of the association between bullying victimization and paranoia can be attributed to shared genetic aetiology (Shakoor et al., 2015). In addition to confounding, bias can also occur as a result of methodological caveats, such as shared method variance (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003), i.e. when the same method or source is used to assess both the predictor and the outcome variables (e.g. child-reported bullying victimization and child-reported depression). As such, shared rater effect may unduly inflate estimates. Research on child development has also emphasized the importance of assessing stability and change of the effects of early life adversity (Fearon, Bakermans-Kranenburg, van IJzendoorn, Lapsley, & Roisman, 2010; Fraley, Roisman, & Haltigan, 2013; Groh et al., 2014). For example, recent studies reported that the concurrent associations between bullying victimization and outcomes were stronger than the longitudinal associations (i.e. lagged outcome) (Lee & Vaillancourt, 2018; Singham et al., 2017), indicating the possibility of decreasing effects with time elapsed since exposure. This body of evidence suggests that the magnitude of the effects of bullying victimization may depend on several factors, including the level of adjustment of environmental and

genetic confounders, the shared rater effect, as well as the length of time elapsed between the exposure to bullying victimization and outcome. Whenever possible, epidemiological investigation aiming to examine the consequences of bullying victimization should therefore take into account such potential effect moderators.

Under ideal circumstances, a randomized controlled trial (RCT) is implemented to deal with confounding. However, randomly exposing children to bullying victimization is clearly unethical. As an alternative, methods to strengthen causal inference in observational studies can be implemented. Such methods are best understood within the counterfactual framework for causal inference (Höfler, 2005; Rubin, 1974, 1990; Rutter, 2007). According to this theoretical framework, the same individual should be both exposed and not exposed to a risk factor (e.g. bullying victimization) at the same time and then outcomes compared between the two situations. In this scenario, exposed and control individuals are literally one and the same, resulting in the elimination of all potential sources of observed and unobserved confounding. Hence, any difference in outcomes could solely be attributed to the effect of the risk factor. Naturally, this ideal scenario is impossible (i.e. counterfactual) in the real word, as the same individual cannot be at the same time exposed and non-exposed to a given risk factor. Quasi-Experimental designs (QE) aim to approximate this scenario, either by design or by statistical innovation. “Quasi” indicates the absence of randomization, while preserving an experiment-like scenario in which matched individuals should solely differ in the exposure of interest. Despite QE-designs being more powerful than mere correlational evidence in terms of causal inference, the majority of studies investigating consequences of bullying victimization have relied on conventional methods to adjust for confounding. This might reflect the more complex nature of QE-designs, which generally require large sample sizes, often multiple time points (e.g. for fixed effects analysis) or specific study populations (e.g. a sample of MZ twins for the discordant twin design). In the following section, we introduce some of the most commonly employed QE-designs in research on bullying

victimization, including difference studies and propensity score matching studies. Although other QE-methods exist (e.g. instrumental variable analysis/Mendelian randomization, regression discontinuity design, interrupted time series studies), they are not further described here as they have not yet been employed to study the impact of bullying victimization (see Supplementary Material for a more detailed introduction to these methods).

The Application of Quasi-Experimental Designs to Investigate the Consequences of Bullying Victimization

Two main types of QE-designs have been applied to study the consequences of bullying victimization: (i) difference studies, which rely on statistical innovation (fixed effect analysis) or specific sample features (twin design); and (ii) propensity score matching studies.

First, difference studies incorporate concepts of the counterfactual framework by comparing individuals to their genetically matched self, either within pairs of MZ twins (i.e. exposed twin vs. non-exposed co-twin) or within individuals over time (i.e. periods of exposure vs. periods of non-exposure). Statistically, this scenario is approximated when using panel data in fixed effects (FE) analysis, in which case both the exposure and the outcome of interest are measured at two or more time points in the same set of individuals. By comparing periods of exposure to periods of non-exposure within the same individual, such data structure enables the estimation of the adjusted effects of bullying victimization. By doing so, all time-invariant factors preceding the bullying experience, including genetic and environmental factors are controlled for. To illustrate, one study (McCuddy & Esbensen, 2017) reported that within-individual changes in bullying victimization over time were not significantly linked to subsequent within-individual changes in risk of substance

use (OR=1, $p>0.05$). In contrast, larger and significant effects were reported for between-individual differences on risk of substance use (OR=1.8, $p<0.05$) (McCuddy & Esbensen, 2017). This could indicate that a common underlying factor (e.g. pre-existing mental health problems) may bias the association.

To attain more stringent matching through the integration of unobserved environmental and genetic factors, difference studies have also made use of the twin design. The twin design capitalizes on the fact that MZ twins share 100% of their segregated genes and 100% of the shared environment (e.g. SES, household characteristics). Therefore, if bullying victimization has an effect independently of genetic and shared environmental confounding, we would expect than an exposed MZ twin will experience worse outcomes than their non-exposed twin. For example, in a study employing the discordant MZ twin design, MZ twin pairs discordant for the exposure of interest (i.e. one twin bullied vs. one twin not bullied) were selected and compared (Arseneault et al., 2008). Results indicated that the bullied twin had higher levels of internalizing symptoms than his or her co-twin in the year following the bullying experience. To summarize, the various types of difference studies embrace the counterfactual framework by creating groups of exposed and non-exposed individuals that are matched on a range of observed and unobserved confounders. Naturally, difference studies can only approximate the counterfactual scenario. Therefore such designs cannot control for the influence of all confounding variables, including all time-variant factors (in fixed effects analysis) or non-shared environmental influences (in the twin designs), even though additional analytical steps can mitigate this problem (e.g. by controlling for observed non-shared environmental risk factors in twins).

As a second QE-method that stems from the counterfactual framework, propensity score matching (PSM) (Rosenbaum & Rubin, 1983) has been employed to assess the developmental consequences of bullying victimization. As with difference studies, the overreaching goal of adjustment through PSM is to approximate the counterfactual scenario by creating statistically

matched groups of individuals that differ solely in their exposure to bullying victimization. Here, a PSM score for each individual is generated, reflecting the probability of being in the exposure group conditional on a set of observed variables (i.e. potential confounders). This score is then used to match an exposed group to a non-exposed group (for more details on the generation and implementation of PSM scores, see Supplementary Material). In the context of bullying victimization, PSM methods have been used to compare bullied children to their non-exposed counterparts (Connell, Morris, & Piquero, 2017), matching them on a range of pre-exposure variables such as gender, socioeconomic status, school performance, parental education and pre-existing behavioural (e.g. peer relations, fights with other children, irritability, restlessness) and emotional problems (e.g. fearful, worrisome). From a counterfactual perspective, PSM studies are conceptually distinct from difference studies since they do not account for unobserved factors. Given this lack of control for unobserved variables and the resulting reduction in the level of internal validity (Geldsetzer & Fawzi, 2017), PSM is not considered as a strictly QE-design. However, the creation of a pseudo-randomization through PSM has been shown to be superior to more conventional multivariate adjustment (Martens, Pestman, de Boer, Belitser, & Klungel, 2008; Stuart, 2010).

Summary and Aims

Strikingly, no study to date has systematically evaluated the existing QE evidence base regarding the impact of bullying victimization. This is surprising, considering the extensive number of meta-analyses investigating the concurrent or longitudinal associations between bullying victimization on developmental outcomes (Moore et al., 2017), including internalizing problems (Hawker & Boulton, 2000; Reijntjes et al., 2010; Ttofi et al., 2011; Yuchang, Junyi, Junxiu, Jing, & Mingcheng, 2017), externalizing problems (Nakamoto & Schwartz, 2010), academic achievement (Reijntjes et al., 2011; Ttofi, Farrington, & Lösel, 2012; Ttofi, Farrington, Lösel, Crago, &

Theodorakis, 2016), suicide attempts and suicidal ideation(Castellví et al., 2017; Holt et al., 2015; Van Geel et al., 2014), sleeping problems (van Geel et al., 2016), self-esteem (Tsaousis, 2016) and psychotic symptoms (van Dam et al., 2012). A possible reason for the exclusion of QE-evidence in those meta-analyses is that QE-studies employ a range of different statistical models, whose translation into a common metric can be challenging. However, ways of combining different estimates have been proposed to circumvent this restriction (Borenstein, Hedges, Higgins, & Rothstein, 2009), opening up the possibility of including QE-studies despite different underlying statistical models. Here, we set out a study to summarize all the available QE-evidence in order to deepen our understanding of the consequences of bullying victimization, focusing on the following questions:

1. Do the associations of bullying victimization with children's short and long-term outcomes hold true when more stringent causal inference methods (QE-designs) are applied?
2. Do the associations of bullying victimization with outcomes differ depending on (i) the level of adjustment (i.e. QE-adjustment vs. non-adjustment) (ii) a shared rater effect and (iii) the time elapsed between the exposure to bullying victimization and the outcome assessment (i.e. persistence of effects)?

Methods

Search Strategy

We searched the PubMed database and three databases through Ovid (PsycINFO, EMBASE, MEDLINE) without any initial language restriction for articles published up to September 13th, 2017 (cf. Supplementary Material for more information on the search procedure). In addition, the reference list of relevant articles and published meta-analyses were screened to identify articles that were missed by the search. To maximize the comparability with previous meta-analyses (Moore et al., 2017; Nakamoto & Schwartz, 2010; Reijntjes et al., 2010; van Dam et al., 2012; Van Geel et al.,

2014), the following search terms were used to index bullying victimization: victimi*, victim, bully*, bullie*, harass*, teas*. The study design of interest was defined as Quasi-Experiment, indexed by: Mendelian randomization, twin, twins, adoption, siblin*, propensity score, matching, experience sampling, ecological momentary assessment, difference in differences, instrumental variable, interrupted time series analysis, quasi-experimental, quasi-experiment, causal. To include all possible outcomes assessed to date, we did not restrict the search to any particular outcome. Relevant indexing terms (e.g. MeSH for PubMed) corresponding to each of the search terms were included in the search in all four databases. Initially, the titles and study abstracts were screened, resulting in the removal of non-relevant studies and duplicates. Full-text reading and assessment of eligibility of the studies were carried out by LD and TS and disagreement was resolved through discussion with the senior researcher (JBP). All authors of the study agreed on the final inclusion of N=16 studies.

Study Selection

We followed PRISMA guidelines (Moher, Liberati, Tetzlaff, & Altman, 2009) and more specific guidelines for meta-analysis of data from QE-studies (Aloe et al., 2017). Studies were included if:

1. they assessed bullying victimization in the general population (cf. Supplementary Material for excluded non-population based studies). The assessment of bullying/peer victimization had to be specific, i.e. studies that focused on other types of victimization (e.g. assault, sexual abuse, cybervictimization) or perpetrators other than peers (e.g. family, strangers, workplace contacts) were excluded (for details see Supplementary Material, sMethods)
2. they used any of the QE-methods described in this study (cf. Introduction and Supplementary Material) to estimate the effect of bullying victimization on outcomes
3. they reported estimates for bullying victimization on outcomes that reflected either (a) concurrent effects [e.g. victimization at age 12 years and psychological stress at age 12

(Ouellet-Morin et al., 2011)] or (b) subsequent effects [e.g. bullying victimization at age 11 years and outcomes at age 16 (Singham et al., 2017)].

Effect Size Calculation

We estimated the effect size (ES) Cohen *d* for each outcome reported, reflecting the standardized mean difference between victims of bullying and non-victims. The ES was directly estimated if the outcome variable was continuous and compared between exposed and unexposed individuals, using the reported means and standard deviations per group (Arseneault et al., 2008; Ouellet-Morin et al., 2011). For all other study designs, we transformed the reported statistics to ES by using the R package compute.es (Del Re, 2013), as follows: For studies that reported standardized correlational estimates between two continuous variables (Singham et al., 2017; Vitaro et al., 2011; Vitaro, Boivin, Brendgen, Girard, & Dionne, 2012), the coefficients were treated as correlation coefficients and converted to ES. For studies reporting effect estimates in proportions (McCuddy & Esbensen, 2017; Roh et al., 2015; Silberg et al., 2016), the chi-square statistics or odds ratio/log odd ratio were used. For regression estimates in which the predictor was binary and outcomes continuous (Delprato et al., 2017; Hoffman, Phillips, Daigle, & Turner, 2016; Kibriya, Xu, & Zhang, 2015; Wong & Schonlau, 2013), we used beta-coefficients and corresponding standard errors to calculate the Z-values and *p*-values in order to then estimate ES. Effect sizes were reversed when higher scores indicated better outcomes. Therefore, higher ES implicated worse outcomes related to bullying victimization across all analyses. Reversal was implemented for prosocial behaviour (Singham et al., 2017), academic performance (Delprato et al., 2017; Kibriya et al., 2015; Ponzo, 2013; Vitaro et al., 2012), socialising/sense of belonging (Delprato et al., 2017), age of first sexual intercourse (DeCamp & Newby, 2015), age of onset of alcohol use (DeCamp & Newby, 2015), cognitive development (Vitaro et al., 2012) and pubertal stage (Ouellet-Morin et al., 2011). A more detailed description of the measures, the coding procedures and the reported statistics per study can

be found in *Table 1.* and in the Supplementary Material (sMethods, *sTable 1.*). Whenever reported, we also extracted estimates from unadjusted models (i.e. non-quasi experimental), which was possible for most of the studies but three (M Brendgen et al., 2017; Hoffman et al., 2016; Ouellet-Morin et al., 2011). Authors were contacted if the reported data did not allow us to calculate the ES (DeCamp & Newby, 2015; Delprato et al., 2017; Roh et al., 2015; Silberg et al., 2016). Whenever estimates for nonsignificant effects were not available (Silberg et al., 2016), we assigned a *p*-value of 1 ($d = 0$) as the most conservative effect size estimate.

Multilevel Random Effects model

All analyses were conducted in R (R Core Team, 2015) and the package metafor (Viechtbauer, 2010). The meta-analytical models were conducted as random effects models (REM) in order to derive a pooled ES – Cohen *d* – assuming heterogeneity across the different outcomes. Since most of the included studies reported effect estimates for multiple outcomes and analysed data from overlapping cohorts (cf. *Table 1.*), the assumption of independence of effect sizes that underlies traditional fixed-effects models or two-level random effects models was violated. As outlined by a recent methodological article regarding meta-analyses (Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2015), the application of multilevel random effects models (MREM) allows us to address such dependence (i.e. correlation between effect sizes) by grouping together ES estimates based on higher order clustering. This is now commonly employed by more recent meta-analyses (Holt et al., 2015; Weisz et al., 2013; Zeegers, Colonnesi, Stams, & Meins, 2017) and is considered superior to previous methods, such as averaging effect sizes or selecting only one outcome per study (Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013). In this study, we tested a three-level MREM [see (Assink & Wibbelink, 2016) for a detailed description and implementation in R]. This model incorporates three sources of variation, namely variation in effect sizes due to random sampling of effect sizes (Level 1, variance that is unique for

each estimated ES per outcome), variation in ES between outcomes within a single cohort (Level 2, variance that is common to all outcomes within a single cohort) and variation in ES between different study cohorts (Level 3, variance that is common to all cohorts). Additional details on the definitions of the three levels included in our MREM models are provided in the Supplementary Material (cf. sMethods). In contrast to the MREM tested here, the traditional random effects model incorporates only two sources of variance (i.e. therefore called two-level random effects model), including the within-study variance and the between study/cohort variance. Hence, in our 3-level model, one additional level (Level 2) was integrated.

When non-independence was attributable to longitudinal data, e.g. in the form of multiple assessments of bullying victimization over time (Hoffman et al., 2016) or multiple time points for the same outcome measures (Silberg et al., 2016; Singham et al., 2017), we included data from the first wave only. This is because the first time point of assessment usually includes the largest number of participants. Late follow-up data was only excluded if the differences in follow-up times could not be used to create meaningful subgroups for the moderator analysis (e.g. short-term vs. long-term effects, cf. below). If studies reported data from several independent samples that were included in one paper (Ponzo, 2013) they were treated as separate cohorts. To reduce heterogeneity in outcomes and enable the meta-analysis, we classified the different outcomes listed in *Table 1* into three broad categories. For each category, a set of separate MREM models were tested to estimate the pooled ESs. More specifically, MREMs were tested for; (1) internalizing symptoms (i.e. symptoms that may be affective, emotional or psychological, e.g. depression, anxiety, stress), (2) externalizing symptoms (negative behaviours directed towards an individual's external environment, such as violence, misconduct, hyperactivity), and (3) academic difficulties (e.g. performances in school tests). One NOS-category (not otherwise specified) was created that included those outcomes that did not fall into any of the above categories (e.g. cortisol response, BMI, psychotic experiences). For this category, we did not estimate the pooled effects size since

outcomes were too heterogeneous. Instead, only individual effect sizes were computed to enable comparison with the other outcomes. We also tested whether heterogeneity of ESs in Level 2 (within-cohort heterogeneity) and Level 3 (between-cohort heterogeneity) was significant by conducting two separate one-sided log-likelihood-ratio tests (Assink & Wibbelink, 2016). Publication bias was tested visually by inspecting funnel plot asymmetry. More formally, Egger's linear regression test was used to assess bias using precision (sampling variance) to predict Cohen *d* effect size (Egger, Smith, Schneider, & Minder, 1997).

Moderators

MREM models are also advantageous for testing moderating effects when effects sizes for different outcomes in each study are available. We tested the following moderators:

1. Level of adjustment. Most of the identified QE-studies reported separate estimates from traditional models (i.e. non-QE models) when examining the effect of bullying victimization on adverse outcomes. In our study, we used those results in order to compute the unadjusted effects of bullying victimization. The QE-unadjusted effects reflected estimates from models where no covariates were accounted for (except for one study that accounted for some observed covariates, cf. *sTable 1.*), including estimates from chi-square tests, t-tests, correlation coefficients or simple regression analyses. The QE-adjusted effects comprised only effect sizes estimated by QE-methods (cf. *sTable 1.*, column 'Statistics used to derive *d*' for an overview of the specific QE-methods used per study). This allowed us to compare how the level of adjustment (QE-unadjusted vs. QE-adjusted) affected effect sizes, in order to evaluate whether the level of adjustment had a significant impact on the magnitude of effects (see Moderator analysis in Supplementary Material for details).
2. Shared rater effect. Studies were classified based on whether the measurements for bullying victimization and outcomes were completed by the same person (i.e. shared source, e.g. both completed by the child) or whether the sources were different for bullying victimization and

outcomes (i.e. non-shared source, e.g. bullying reported by the teacher, outcomes reported by the child).

3. Persistence of effects. We tested a moderator defined as 'persistence of effects' to investigate whether QE-adjusted effect changed as a function of the time that elapsed following the exposure of bullying victimization. Based on the length of follow-up in available studies (cf. s*Table 1.*), we classified study outcomes depending on whether they assessed the short-term ($\leq$ 1 year of follow-up) or long-term effects (> 1 year of follow-up) of bullying victimization.

Results

Study Description

A total number of N=16 publications met the criteria for inclusion (cf. Flow chart, *Figure 1.*), all of which were published between 2008 and 2017 (cf. *Table 1.*). The studies comprised 13 distinct study cohorts. We derived k=101 QE-adjusted outcome estimates and k=80 unadjusted outcome estimates. Most of the unadjusted estimates (k=77) reflected statistics from simple models, i.e. models that did not control for any covariate. The most commonly reported QE-adjusted estimates were derived from twin designs [57.4% (k=58)], followed by propensity score matching analysis [39.6% (k=40)] and fixed effects regression analysis [3.0% (k=3)]. The majority of the outcomes [72.3% (k=73)] fell into one of our broader developmental outcome categories, including internalizing symptoms [23.8% (k=24)], externalizing symptoms [38.6% (k=39)] and academic difficulties [9.9% (k=10)]. For the NOS (not otherwise specified) category [27.7% (k=28)], we did not estimate the pooled effect size because outcomes were conceptually too distinct to be grouped together in a meaningful way (e.g. relationship quality, cortisol response, pubertal stage, BMI, psychotic experiences, school suspension). We therefore reported the individual standardized effect sizes in the Supplementary Material (cf. sResults and *sFigure 1.*) to help comparison with findings

for the three broader categories. Funnel plots (cf. *sFigure 2*, Supplementary Material) and Egger's test (cf. *Table 2.*) showed no evidence of publication bias in the QE-adjusted models for internalizing symptoms, externalizing symptoms and academic difficulties.

Multilevel random effects model: Effects of peer victimization

As shown in *Table 2* and *Figure 2*. to *Figure 4.*, the largest Cohen *d* effect sizes were identified in the unadjusted multilevel random effect models. In the adjusted models, the magnitude of adverse effects of bullying victimization dropped but remained significant for internalizing symptoms ($d_{adjusted}$=0.27, 95% CI 0.05; 0.49, *Figure 2.*), externalizing symptoms ($d_{adjusted}$=0.15, 95% CI 0.10; 0.21, *Figure 3.*) and academic difficulties ($d_{adjusted}$=0.10, 95% CI 0.06-0.13, *Figure 4.*). Significant unadjusted and adjusted effects were also present for some of the outcomes classified as NOS [not otherwise specified; $k_{adjusted}$=28 outcomes tested ($k_{adjusted}$= 8 outcomes were significant, e.g. psychotic symptoms)], which are reported in the Supplementary Material (cf. *sFigure 1.*). When evaluating heterogeneity among the effect sizes in multilevel random effects models, we found that most of the variance was due to between-cohort heterogeneity (Level 3) rather than within-cohort heterogeneity (Level 2), as evident for academic difficulties ($I^2_{Level\ 2}$<0.0001% vs. $I^2_{Level\ 3}$= 72.38%) and internalizing symptoms ($I^2_{Level\ 2}$= 12.30% vs. $I^2_{Level\ 3}$= 77.91%). This indicates that factors in which the cohorts may differ (e.g. mean age of the study population, length of follow-up) could account for some of the variation in effect sizes. For externalizing symptoms, a larger proportion of variance reflected within-cohort rather than between-cohort variation ($I^2_{Level\ 2}$=75.69% vs. $I^2_{Level\ 3}$= 7.27%), suggesting that within-cohort factors (i.e. differences in effect sizes between outcomes within the same cohort, such as hyperactivity and substance use) may account for some of the variations in effect sizes for externalizing symptoms.

Multilevel mixed effects model: Sources of heterogeneity

The results from the multilevel mixed models are displayed in *Table 3*. First, we tested the moderation effect of the variable 'level of adjustment' by comparing adjusted to unadjusted Cohen *d* effect sizes. Here, the unadjusted effects were visually larger than the QE-adjusted effects for all three outcome dimensions. The largest decrease in effect sizes due to adjustment was present for externalizing symptoms ($d_{unadjusted}$=0.34, 95% CI 0.11; 0.57, $d_{adjusted}$=0.15, 95% CI 0.10; 0.21), which was supported by a significant moderating effect ($p_{moderator}$=0.006). Smaller changes in effect sizes following QE-adjustment were present for internalizing symptoms ($d_{unadjusted}$=0.36, 95% CI 0.03; 0.69, $d_{adjusted}$=0.27, 95% CI 0.05; 0.49) and academic difficulties ($d_{unadjusted}$=0.12, 95% CI 0.08; 0.17, $d_{adjusted}$=0.10, 95% CI 0.06; 0.13), in which cases the moderating effect was non-significant ($p_{moderator}$>0.05).

Second, we tested whether a shared rater effect moderated effect sizes. To assess this, we compared estimates relying on a shared rater method (bullying victimization and outcomes reported by the same individual) to estimates relying on a non-shared rater method (bullying victimization and outcomes not reported by the same individual). The results from the mixed effects models implicated that the rater effect significantly impacted on the association between bullying victimization and internalizing symptoms (d_{shared}=0.37, 95% CI -0.03; 0.77, $d_{non\text{-}shared}$=0.14, 95% CI 0.05; 0.23, $p_{moderator}$<0.0001), implicating that stronger adverse effects were reported if the shared rater method was used. No significant moderation effect was present for externalizing symptoms, although there was a trend of reduction in effect sizes when relying on non-shared rater methods (d_{shared}=0.18, 95% CI 0.09; 0.27, $d_{non\text{-}shared}$=0.06, 95% CI 0.01; 0.11, $p_{moderator}$=0.12). Of note, all studies that relied on a shared rater method assessed both exposure and outcomes based on the child's self-report. A shared rater moderating effect could not be tested for academic difficulties, since all effect sizes for this outcome reflected non-shared rater estimates.

Finally, we tested whether the short-term effects of bullying victimization were different from long-term effects in QE-adjusted models (cf. Persistence of effect, *Table 3*.). Here, the results

indicated that the duration of the follow-up significantly altered the magnitude of the effects of bullying victimization on internalizing symptoms, pointing towards adverse short-term effects (i.e. 1 year or less of follow-up) that were no longer significant in the long-term ($d_{\text{short-term}}$=0.33, 95% CI 0.10; 0.57, $d_{\text{long-term}}$=0.06, 95% CI -0.01; 0.13, $p_{\text{moderator}}$=0.02). No significant moderation effect of the duration of the follow-up was present for academic difficulties and externalizing symptoms, although subgroup analysis indicated that detrimental effects of bullying victimization either did not persist in the long-term for academic difficulties ($d_{\text{long-term}}$=0.04, 95% CI -0.07; 0.16) or were only small in magnitude over the long-term for externalizing symptoms ($d_{\text{long-term}}$=0.13, 95% CI 0.09; 0.17).

Discussion

This meta-analysis set out to pool together the most stringent available evidence regarding the detrimental consequences of bullying victimization, by drawing on Quasi-Experimental (QE) studies. Overall, our findings indicate that bullying victimization has small causal adverse effects on a range of developmental outcomes, and most notably on internalizing behaviours; these adverse effects diminish in the long-term. Our QE-adjusted effects were smaller in magnitude for all outcomes than those reported by available non-QE meta-analyses. The largest adverse effects were identified for internalizing symptoms, but effect sizes were still small. Smaller effects were detected for externalizing symptoms and academic difficulties. For internalizing symptoms, the most impacted outcome, the harmful effects of bullying victimization decreased once the shared rater effect was accounted for, indicating that study estimates relying on shared raters (i.e. self-reported bullying victimization and self-reported developmental outcomes) may inflate effects of bullying victimization. Further decreases in effects of bullying victimization occurred as time elapsed, as indicated by the finding that short-term adverse effects for internalizing problems no longer remained significant in the long-term. In the following sections, we discuss in turn: (i) the detrimental impact of bullying victimization, (ii) the effect of stringent adjustment for confounding,

(iii) the impact of the shared rater effect, (iv) the potential for resilience and (v) implications for prevention.

Quasi-Experimental Evidence of Main Effects of Bullying Victimization

When pooling together evidence from Quasi-Experimental (QE) designs, significant QE-adjusted effects were present for all three types of outcomes: internalizing and externalizing symptoms as well as academic difficulties. This is in line with previous longitudinal (non-QE) studies, which reported significant adverse associations of bullying victimization and a range of outcomes (Arseneault et al., 2006; Stapinski et al., 2014; Takizawa, Maughan, & Arseneault, 2014). Our findings also support of the view that young victims show behavioural and emotional problems as symptoms of their psychological distress in response to the bulling experience, including internalizing symptoms such as depression, anxiety and suicidality (Arseneault et al., 2008; Roh et al., 2015; Singham et al., 2017), as well as externalizing symptoms such as conduct problems, hyperactivity, delinquency and substance use (Connell et al., 2017; DeCamp & Newby, 2015; McCuddy & Esbensen, 2017; Singham et al., 2017). The strength of effects were small for all three types of outcomes, with small effects corresponding to d=0.20 and medium effects to d=0.50 (Cohen, 1988). To put this into perspective, our QE-adjusted risk estimate can best be compared to risk factors for mental health issues that occur in a more random fashion, such as adverse natural events (e.g. tsunami, earthquake), which are less likely to be biased by potential confounders such as pre-existing mental health conditions. Estimates for such risk factors constitute a useful benchmark as they more likely reflect causal relationships. Our largest significant QE-adjusted risk estimate (d=0.27 for internalizing symptoms) was smaller or comparable to the effect of a variety of such 'naturally adjusted events' on post-traumatic stress symptoms, e.g., exposure to natural disaster (d=0.32, e.g. earthquake, nuclear waste disaster), man-made disaster (d=0.41, e.g. terrorism) or personal loss (d=0.32) (Furr, Comer, Edmunds, & Kendall, 2010). Similarly, our

largest QE-adjusted estimate was smaller than estimates reported in meta-analyses looking at the effects of bullying victimization on internalizing symptoms [*d*=0.37 (Reijntjes et al., 2010)]. Such results highlight the importance of rigorously addressing bias through the application of QE-designs, not just in empirical but also in future meta-analytical work. The outcome most affected by the application of QE-adjustment was externalizing symptoms, in which case adjustment halved the pooled effect size, implicating that shared genetic and/or environmental components may impact this particular association. This is in line with the wider literature, which suggests that up to 60% of the association between bullying victimization and externalizing outcomes could be due to genetic factors (Connolly & Beaver, 2016). The same study also reported that genetic factors played less of a role in the relationship between bullying victimization and internalizing symptoms, which is consistent with our finding that QE-adjustment had less of an impact on effect sizes for internalizing symptoms. Hence, affective conditions such as depression and anxiety may be more likely to be causally influenced by bullying victimization. Finally, only a small contribution of bullying victimization to academic difficulties was identified. This may be due to the fact that a diverse set of factors is likely to affect academic life, including genetic as well as social factors (Kiernan & Mensah, 2011; Krapohl et al., 2014) – some of which may be more influential than bullying victimization.

Sources of Heterogeneity in Quasi-Experimental Studies

Mixed effects models indicated that several factors moderated the harmful effects of bullying victimization. First, when comparing the unadjusted models to QE-adjusted models, our moderator analysis confirmed that QE-adjustment significantly reduced the magnitude of effects of bullying victimization on externalizing symptoms. Second, we found that the strength of adverse effects of bullying victimization was influenced by the shared rater effect. This rater effect was significant and strong for internalizing symptoms but not externalizing symptoms. This implies that, in the

presence of shared rater variance, the effect of bullying victimization on internalizing symptoms is likely to be overestimated when compared to estimates from studies that make use of different sources of informants. For example, those individuals who report adverse outcomes and high levels of bullying victimization may do so because of an underlying tendency to report negative feelings about aspects of one's life. Hence, future investigations should include multiple rather than single informants in order to strengthen finding validity. Finally, we found that adverse effects were stronger in the short-term than in the long-term following the exposure to bullying victimization. In particular, internalizing symptoms were no longer significantly affected in the long-term. Encouragingly, this finding emphasizes the potential for resilience in bullied children. This is similar to a previous observation, in which it was reported that cessation of exposure to bullying was significantly linked to reductions in psychotic experiences over a three-month period (Kelleher et al., 2013). Similarly, in a 35-year follow-up study that controlled for childhood behavioural problems, the apparent link between bullying and psychosis dissipated over time (Boden, van Stockum, Horwood, & Fergusson, 2016). While it is not possible to draw more definite conclusion on the mechanisms underlying the process of resilience based on the current results, our findings highlight the need for future studies to explore the timing and pathways to resilience more thoroughly (Singham et al., 2017).

Taken together, these findings add to the current knowledge base by better characterizing the adverse consequences of bullying victimization on key developmental outcomes. Our findings highlight that it is essential to (1) implement causal inference designs, since widespread associations between bullying victimization and outcomes can partially or totally arise from observed and unobserved confounders; (2) adopt multi-method assessment approaches as effects depend on the methods of measurement used to assess both bullying. victimization and outcomes, and (3) more finely characterize to what extent and for how long do adverse effects of bullying victimization persist.

Implications for Intervention

Prevention of bullying victimization remains crucial, since children exposed to victimization are likely to exhibit difficulties in the short-term, such as increased levels of anxiety and depression. Given the high prevalence rates of emotional problems among children [e.g. 1 in 4 girls displays symptoms of depression at the age of 14 (Patalay & Fitzsimons, 2017)], it is important to pursue anti-bullying initiatives to reduce the occurrence of bullying victimization. Importantly, our results indicate that even if anti-bullying initiatives succeeded in fully eliminating bullying in school, only small changes in outcomes can be expected. The upper limit of the beneficial effects of such interventions can be expected to be the highest causal estimates of the effect of bullying victimization (e.g. $d = 0.27$ for internalizing symptoms). Any program efficient in reducing bullying victimization should result in a proportional decrease in its indirect impact on outcomes. This may explain why the KIVA program, which was successful in that it reduced bullying perpetration by about 60% (Kärnä et al., 2011), did not have a significant impact on depression (Williford et al., 2012). More encouragingly, however, diminishing detrimental effects following exposure to bullying over time highlight the potential for resilience in bullied children. Therefore, besides primary interventions aiming to eliminate bullying, secondary interventions should implement strategies to foster the process of resilience in bullied children. Interventions similar to the 'Resilience Triple P' program (Healy & Sanders, 2014) may help bullied children to build resilience and to deal with the associated stress. Our findings suggest that such interventions may be most effective if implemented immediately following the occurrence of bullying, with a particular focus on resulting internalizing symptoms. Where possible, such interventions should also address risk factors preceding bullying victimization, in order to reduce the risk of vulnerable individuals to experience further bullying, while improving long-term prospect by addressing common causes of victimization and negative outcomes (e.g. pre-existing mental health vulnerabilities that lead to both

a greater likelihood of being victimized in the first instance and to long-term mental health problems).

Limitations

Despite our attempt to strengthen causal inference, limitations that relate to the nature of meta-analytical procedures should be considered. First, to be able to pool together all available QE-evidence, our analysis included studies with different statistical designs. Since we transformed all estimates to a common metric, our reported Cohen *d* reflects different underlying mathematical models. Hence, our effect estimates should be considered as imperfect approximations of causal effects. Nevertheless, the alternative option, namely the systematic exclusion of QE-evidence, would constitute a greater threat to the validity of the results. In this context, it is also important to acknowledge that the QE-methods included in our analysis are not immune to bias (cf. Supplementary Material, where the approach-specific limitations are highlighted), so that we cannot rule out the presence of residual confounding. Second, we included only one broad dimension of bullying victimization and did not distinguish between different forms of bullying (e.g. physical or social). This was not feasible because only few studies looked at different forms of bullying in isolation (Delprato et al., 2017; Roh et al., 2015; Singham et al., 2017). Such evidence indicates that the consequences of bullying are similar across different forms of bullying, but future QE-studies should examine this question more systematically. Similarly, other relevant moderating factors such as gender were not examined in our study, since details on demographic variables were generally unavailable for the study samples used for QE-analysis. Yet, when tested more formally in a twin difference study, gender did not moderate the effects of bullying victimization on mental health outcomes (Singham et al., 2017). However, we were not in a position to detect any adverse effects exhibited only by a subcategory of individuals. The adverse effects of bullying victimization may depend on the age at victimization (Cook, Williams, Guerra, Kim, & Sadek, 2010; Yen et al., 2013).

However, we were unable to test whether age at victimization moderated the effect of bullying victimization on outcomes, as age was imprecisely reported in most studies, with wide age intervals (cf. *sTable 1.*, Supplementary Material). We are also unable to conclude whether bullying victimization has adverse effects on other outcomes not examined by QE-studies (e.g. eating disorders), although the dimensions of our developmental outcomes dimensions reflect those that are most commonly examined in the wider literature. Due to a lack of available data, it was also not possible to investigate more thoroughly dose-response patterns linked to severity of exposure or chronicity over time. Little QE-evidence in this regard is available – for instance, it was reported that the adverse effects on developmental outcomes increased if bullying was experienced recurrently (Hoffman et al., 2016) or increased in its severity (Connell et al., 2017). Finally, one potential source of confounding in twin designs is the impact of pre-existing mental health issues on the association between bullying victimization and developmental outcomes. However, most of the twin studies in our meta-analysis either controlled for pre-existing conditions in their models (Vitaro et al., 2012) or their results were not substantially affected by the inclusion of pre-existing mental health issues as reported in sensitivity analyses (Arseneault et al., 2008; Singham et al., 2017).

Conclusions

We pooled together the most stringent Quasi-Experimental evidence available to date on the consequences of bullying victimization. Our results indicate that being exposed to bullying negatively impact a range of outcomes including internalizing and externalizing symptoms, as well as academic difficulties. Effect sizes are smaller than what would be expected based on previous (non-Quasi-Experimental) research, accentuating the need to acknowledge the potential for pre-existing risk factors confounding the association between bullying victimization and outcomes. Detrimental effects decreased as time elapsed from the exposure to bullying, highlighting the

potential for resilience in bullied children. To reduce negative outcomes in the long-term, we propose that, in addition to anti-bullying programmes, interventions focusing on resilience and addressing pre-existing vulnerabilities in bullied children may be beneficial.

Acknowledgements

We thank all the authors of the included studies, especially Dr Delprato and Dr DeCamp for kindly sharing their unreported data for inclusion in this meta-analysis. This research was supported by a MQ Transforming Mental Health fellowship awarded to JBP (MQ16IP16).

References

Aloe, A. M., Becker, B. J., Duvendack, M., Valentine, J. C., Shemilt, I., & Waddington, H. (2017). Quasi-experimental study designs series—paper 9: collecting data from quasi-experimental studies. *Journal of Clinical Epidemiology*, *89*, 77–83. https://doi.org/10.1016/j.jclinepi.2017.02.013

* Arseneault, L., Milne, B. J., Taylor, A., Adams, F., Delgado, K., Caspi, A., & Moffitt, T. E. (2008). Being bullied as an environmentally mediated contributing factor to children's internalizing problems: a study of twins discordant for victimization. *Archives of Pediatrics & Adolescent Medicine*, *162*(2), 145–150. https://doi.org/10.1001/archpediatrics.2007.53

Arseneault, L., Walsh, E., Trzesniewski, K., Newcombe, R., Caspi, A., & Moffitt, T. E. (2006). Bullying victimization uniquely contributes to adjustment problems in young children: a nationally representative cohort study. *Pediatrics*, *118*(1), 130–138.

Assink, M., & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *The Quantitative Methods for Psychology*, *12*(3), 154–174. https://doi.org/10.20982/tqmp.12.3.p154

Baldwin, J. R., Arseneault, L., Odgers, C., Belsky, D. W., Matthews, T., Ambler, A., … T.E., M. (2016). Childhood bullying victimization and overweight in young adulthood: a cohort study. *Psychosomatic Medicine, 78*(9), 1094–1103. https://doi.org/10.1097/PSY.0000000000000388

Ball, H. A., Arseneault, L., Taylor, A., Maughan, B., Caspi, A., Moffitt, T. E., … A., C. (2008). Genetic and environmental influences on victims, bullies and bully-victims in childhood. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 49*(1), 104–112. https://doi.org/10.1111/j.1469-7610.2007.01821.x

Bärnighausen, T., Tugwell, P., Røttingen, J.-A., Shemilt, I., Rockers, P., Geldsetzer, P., … Brown, A. (2017). Quasi-experimental study designs series–Paper 4: uses and value. *Journal of Clinical Epidemiology, 89*, 21–29. https://doi.org/10.1016/j.jclinepi.2017.03.012

Boden, J. M., van Stockum, S., Horwood, L. J., & Fergusson, D. M. (2016). Bullying victimization in adolescence and psychotic symptomatology in adulthood: evidence from a 35-year study. *Psychological Medicine, 46*(6), 1311–1320. https://doi.org/10.1017/S0033291715002962

Borenstein, M., Hedges, L. V, Higgins, J. P. T., & Rothstein, H. R. (2009). Converting Among Effect Sizes. In *Introduction to Meta-Analysis* (pp. 45–49). London: John Wiley & Sons.

Bouffard, L. A., & Koeppel, M. D. H. (2014). Understanding the potential long-term physical and mental health consequences of early experiences of victimization. *Justice Quarterly*, *31*(3), 568–587. https://doi.org/10.1080/07418825.2012.734843

* Brendgen, M., Ouellet-Morin, I., Lupien, S. J., Vitaro, F., Dionne, G., & Boivin, M. (2017). Environmental influence of problematic social relationships on adolescents' daily cortisol secretion: a monozygotic twin-difference study. *Psychological Medicine*, *47*(03), 460–470. https://doi.org/10.1017/S003329171600252X

Brendgen, M., & Poulin, F. (2018). Continued bullying victimization from childhood to young adulthood: a longitudinal study of mediating and protective Factors. *Journal of Abnormal Child Psychology*, *46*(1), 27–39. https://doi.org/10.1007/s10802-017-0314-5

Burke, T., Sticca, F., & Perren, S. (2017). Everything's gonna be alright! The longitudinal interplay among social support, peer victimization, and depressive symptoms. *Journal of Youth and Adolescence*, 1–16. https://doi.org/10.1007/s10964-017-0653-0

Camodeca, M., & Goossens, F. A. (2005). Aggression, social cognitions, anger and sadness in bullies and victims. *Journal of Child Psychology and Psychiatry*, *46*(2), 186–197. https://doi.org/10.1111/j.1469-7610.2004.00347.x

Castellví, P., Miranda-Mendizábal, A., Parés-Badell, O., Almenara, J., Alonso, I., Blasco, M. J., … Alonso, J. (2017). Exposure to violence, a risk for suicide in youths and young adults. a meta-analysis of longitudinal studies. *Acta Psychiatrica Scandinavica, 135*(3), 195–211. https://doi.org/10.1111/acps.12679

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale: Lawrence Erlbaum Associates.

* Connell, N. M., Morris, R. G., & Piquero, A. R. (2017). Exploring the link between being bullied and adolescent substance use. *Victims & Offenders*, *12*(2), 277–296. https://doi.org/10.1080/15564886.2015.1055416

Connolly, E. J., & Beaver, K. M. (2016). Considering the genetic and environmental overlap between bullying victimization, delinquency, and symptoms of depression/anxiety. *Journal of Interpersonal Violence*, *31*(7), 1230–1256. https://doi.org/10.1177/0886260514564158

Cosma, A., Whitehead, R., Neville, F., Currie, D., & Inchley, J. (2017). Trends in bullying victimization in Scottish adolescents 1994–2014: changing associations with mental well-being. *International Journal of Public Health*, 1–8. https://doi.org/10.1007/s00038-017-0965-6

Craig, W., Harel-Fisch, Y., Fogel-Grinvald, H., Dostaler, S., Hetland, J., Simons-Morton, B., … Due, P. (2009). A cross-national profile of bullying and victimization among adolescents in 40 countries. *International Journal of Public Health*, *54*, 216–224. https://doi.org/10.1007/s00038-009-5413-9

Dan, O. (1993). *Bullying at school: what we know and what we can do*. Malden: Blackwell Publishing Inc.

Davis, J. P., Dumas, T. M., Berey, B. L., Merrin, G. J., Cimpian, J. R., & Roberts, B. W. (2017). Effect of victimization on impulse control and binge drinking among serious

juvenile offenders from adolescence to young adulthood. *Journal of Youth and Adolescence*, *46*(7), 1515–1532. https://doi.org/10.1007/s10964-017-0676-6

* DeCamp, W., & Newby, B. (2015). From bullied to deviant: The victim-offender overlap among bullying victims. *Youth Violence and Juvenile Justice*, *13*(1), 3–17. https://doi.org/10.1177/1541204014521250

Del Re, A. C. (2013). compute. es: Compute effect sizes. *R Package Version*, 0–2.

* Delprato, M., Akyeampong, K., & Dunne, M. (2017). The impact of bullying on students' learning in Latin America: a matching approach for 15 countries. *International Journal of Educational Development*, *52*, 37–57. https://doi.org/10.1016/j.ijedudev.2016.10.002

Eastman, M., Foshee, V., Ennett, S., Sotres-Alvarez, D., Reyes, H. L. M., Faris, R., & North, K. (2018). Profiles of internalizing and externalizing symptoms associated with bullying victimization. *Journal of Adolescence*, *65*, 101–110. https://doi.org/10.1016/j.adolescence.2018.03.007

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Esbensen, F.-A., & Carson, D. C. (2009). Consequences of Being Bullied. *Youth & Society*, *41*(2), 209–233. https://doi.org/10.1177/0044118X09351067

Evans, C. B. R., Smokowski, P. R., Rose, R. A., Mercado, M. C., & Marshall, K. J. (2018). Cumulative bullying experiences, adolescent behavioral and mental health, and academic achievement: an integrative model of perpetration, victimization, and

bystander Behavior. *Journal of Child and Family Studies*. https://doi.org/10.1007/s10826-018-1078-4

Fearon, R. P., Bakermans-Kranenburg, M. J., van IJzendoorn, M. H., Lapsley, A., & Roisman, G. I. (2010). The significance of insecure attachment and disorganization in the development of children's externalizing behavior: a meta-analytic study. *Child Development*, *81*(2), 435–456. https://doi.org/10.1111/j.1467-8624.2009.01405.x

Fraley, R. C., Roisman, G. I., & Haltigan, J. D. (2013). The legacy of early experiences in development: Formalizing alternative models of how early experiences are carried forward over time. *Developmental Psychology*, *49*(1), 109–126. https://doi.org/10.1037/a0027852

Furr, J. M., Comer, J. S., Edmunds, J. M., & Kendall, P. C. (2010). Disasters and youth: A meta-analytic examination of posttraumatic stress. *Journal of Consulting and Clinical Psychology*, *78*(6), 765–780. https://doi.org/10.1037/a0021482

Geldsetzer, P., & Fawzi, W. (2017). Quasi-experimental study designs series—paper 2: complementary approaches to advancing global health knowledge. *Journal of Clinical Epidemiology*. https://doi.org/10.1016/j.jclinepi.2017.03.015

Groh, A. M., Fearon, R. P., Bakermans-Kranenburg, M. J., van IJzendoorn, M. H., Steele, R. D., & Roisman, G. I. (2014). The significance of attachment security for children's social competence with peers: a meta-analytic study. *Attachment & Human Development*, *16*(2), 103–136. https://doi.org/10.1080/14616734.2014.883636

Hamilton, L. D., Newman, M. L., Delville, C. L., & Delville, Y. (2008). Physiological stress response of young adults exposed to bullying during adolescence. *Physiology & Behavior*, *95*(5), 617–624. https://doi.org/10.1016/j.physbeh.2008.09.001

Hawker, D. S. J., & Boulton, M. J. (2000). Twenty years' research on peer victimization and psychosocial maladjustment: a meta-analytic review of cross-sectional studies. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *41*(4), 441–455. https://doi.org/10.1111/1469-7610.00629

Healy, K. L., & Sanders, M. R. (2014). Randomized controlled trial of a family intervention for children bullied by peers. *Behavior Therapy*, *45*(6), 760–777. https://doi.org/10.1016/j.beth.2014.06.001

Hill, A. B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, *58*(5), 295. https://doi.org/10.1177/003591576505800503

* Hoffman, C. Y., Phillips, M. D., Daigle, L. E., & Turner, M. G. (2016). Adult consequences of bully victimization: are children or adolescents more vulnerable to the victimization experience? *Youth Violence and Juvenile Justice*, *15*(4), 441–464. https://doi.org/10.1177/1541204016650004

Höfler, M. (2005). Causal inference based on counterfactuals. *BMC Medical Research Methodology*, *5*(1), 28. https://doi.org/10.1186/1471-2288-5-28

Holt, M. K., Vivolo-Kantor, A. M., Polanin, J. R., Holland, K. M., DeGue, S., Matjasko, J. L., … Reid, G. (2015). Bullying and suicidal ideation and behaviors: a meta-analysis. *PEDIATRICS*, *135*(2), e496–e509. https://doi.org/10.1542/peds.2014-1864

Jaffee, S. R., Strait, L. B., & Odgers, C. L. (2012). From correlates to causes: Can quasi-experimental studies and statistical innovations bring us closer to identifying the causes of antisocial behavior? *Psychological Bulletin, 138*(2), 272–295. https://doi.org/10.1037/a0026020

Kärnä, A., Voeten, M., Little, T. D., Poskiparta, E., Kaljonen, A., & Salmivalli, C. (2011). A large-scale evaluation of the KiVa antibullying program: Grades 4–6. *Child Development, 82*(1), 311–330. https://doi.org/10.1111/j.1467-8624.2010.01557.x.

Kelleher, I., Keeley, H., Corcoran, P., Ramsay, H., Wasserman, C., Carli, V., … Cannon, M. (2013). Childhood trauma and psychosis in a prospective cohort study: cause, effect, and directionality. *American Journal of Psychiatry, 170*(7), 734–741. https://doi.org/10.1176/appi.ajp.2012.12091169

* Kibriya, S., Xu, Z. P., & Zhang, Y. (2015). The impact of bullying on educational performance in Ghana: a bias-reducing matching approach. In *2015 AAEA & WAEA Joint Annual Meeting, July 26-28, San Francisco, California*. Agricultural and Applied Economics Association & Western Agricultural Economics Association.

Kiernan, K. E., & Mensah, F. K. (2011). Poverty, family resources and children's early educational attainment: the mediating role of parenting. *British Educational Research Journal, 37*(2), 317–336. https://doi.org/10.1080/01411921003596911

Krapohl, E., Rimfeld, K., Shakeshaft, N. G., Trzaskowski, M., McMillan, A., Pingault, J.-B., … Dale, P. S. (2014). The high heritability of educational achievement reflects many

genetically influenced traits, not just intelligence. *Proceedings of the National Academy of Sciences*, *111*(42), 15273–15278. https://doi.org/10.1073/pnas.1408777111

Kretschmer, T., Veenstra, R., Branje, S., Reijneveld, S. A., Meeus, W. H. J., Deković, M., … Oldehinkel, A. J. (2018). How competent are adolescent bullying perpetrators and victims in mastering normative developmental tasks in early adulthood? *Journal of Abnormal Child Psychology*, *46*(1), 41–56. https://doi.org/10.1007/s10802-017-0316-3

Lee, K. S., & Vaillancourt, T. (2018). Longitudinal associations among bullying by peers, disordered eating behavior, and symptoms of depression during adolescence. *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2018.0284

Lester, L., Dooley, J., Cross, D., & Shaw, T. (2012). Internalising symptoms: an antecedent or precedent in adolescent peer victimisation. *Australian Journal of Guidance and Counselling*, *22*(02), 173–189. https://doi.org/10.1017/jgc.2012.25

Listl, S., Jürges, H., & Watt, R. G. (2016). Causal inference from observational data. *Community Dentistry and Oral Epidemiology*, *44*(5), 409–415. https://doi.org/10.1111/cdoe.12231

Martens, E. P., Pestman, W. R., de Boer, A., Belitser, S. V, & Klungel, O. H. (2008). Systematic differences in treatment effect estimates between propensity score methods and logistic regression. *International Journal of Epidemiology*, *37*(5), 1142–1147. https://doi.org/10.1093/ije/dyn079

* McCuddy, T., & Esbensen, F.-A. (2017). After the bell and into the night: the link between delinquency and traditional, cyber-, and dual-bullying victimization. *Journal of*

Research in Crime and Delinquency, *54*(3), 409–441. https://doi.org/10.1177/0022427816683515

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-Aanalyses: The PRISMA statement. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Moore, S. E., Norman, R. E., Suetani, S., Thomas, H. J., Sly, P. D., & Scott, J. G. (2017). Consequences of bullying victimization in childhood and adolescence: a systematic review and meta-analysis. *World Journal of Psychiatry*, *7*(1), 60. https://doi.org/10.5498/wjp.v7.i1.60

Nakamoto, J., & Schwartz, D. (2010). Is peer victimization associated with academic achievement? A meta-analytic review. *Social Development*, *19*(2), 221–242. https://doi.org/10.1111/j.1467-9507.2009.00539.x

Nansel, T. R., Overpeck, M., Pilla, R. S., Ruan, W. J., Simons-Morton, B., & Scheidt, P. (2001). Bullying behaviors among US youth. *JAMA*, *285*(16), 2094. https://doi.org/10.1001/jama.285.16.2094

O'Moore, M., & Kirkham, C. (2001). Self-esteem and its relationship to bullying behaviour. *Aggressive Behavior*, *27*(4), 269–283. https://doi.org/10.1002/ab.1010

* Ouellet-Morin, I., Danese, A., Bowes, L., Shakoor, S., Ambler, A., Pariante, C. M., … Arseneault, L. (2011). A discordant monozygotic twin design shows blunted cortisol reactivity among bullied children. *Journal of the American Academy of Child & Adolescent Psychiatry*, *50*(6), 574–582.e3. https://doi.org/10.1016/j.jaac.2011.02.015

Patalay, P., & Fitzsimons, E. (2017). Mental ill-health among children of the new century: trends across childhood with a focus on age 14. *Centre for Longitudinal Studies: London*.

Pingault, J.-B., O'Reilly, P. F., Schoeler, T., Ploubidis, G. B., Rijsdijk, F., & Dudbridge, F. (2018). Using genetic data to strengthen causal inference in observational research. *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-018-0020-3

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879. https://doi.org/0.1037/0021-9010.88.5.879

* Ponzo, M. (2013). Does bullying reduce educational achievement? An evaluation using matching estimators. *Journal of Policy Modeling*, *35*(6), 1057–1078. https://doi.org/10.1016/j.jpolmod.2013.06.002

Quinn, S. T., & Stewart, M. C. (2018). Examining the long-term consequences of bullying on adult substance use. *American Journal of Criminal Justice*, *43*(1), 85–101. https://doi.org/10.1007/s12103-017-9407-5

R Core Team. (2015). R: A Language and Environment for Statistical Computing. Retrieved from https://cran.r-project.org/doc/manuals/fullrefman.pdf

Reijntjes, A., Kamphuis, J. H., Prinzie, P., Boelen, P. A., Van der Schoot, M., & Telch, M. J. (2011). Prospective linkages between peer victimization and externalizing problems in

children: a meta-analysis. *Aggressive Behavior*, *37*(3), 215–222. https://doi.org/10.1002/ab.20374

Reijntjes, A., Kamphuis, J. H., Prinzie, P., & Telch, M. J. (2010). Peer victimization and internalizing problems in children: a meta-analysis of longitudinal studies. *Child Abuse & Neglect*, *34*(4), 244–252. https://doi.org/10.1016/j.chiabu.2009.07.009

Rockers, P. C., Røttingen, J.-A., Shemilt, I., Tugwell, P., & Bärnighausen, T. (2015). Inclusion of quasi-experimental studies in systematic reviews of health systems research. *Health Policy*, *119*(4), 511–521. https://doi.org/10.1016/j.healthpol.2014.10.006

* Roh, B.-R., Yoon, Y., Kwon, A., Oh, S., Lee, S. I., Ha, K., … Hong, H. J. (2015). The structure of co-occurring bullying experiences and associations with suicidal behaviors in korean adolescents. *PLOS ONE*, *10*(11), e0143517. https://doi.org/10.1371/journal.pone.0143517

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Rubin, D. B. (1990). [On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.] Comment: Neyman (1923) and Causal Inference in

Experiments and Observational Studies. *Statistical Science*, *5*(4), 472–480. https://doi.org/10.1214/ss/1177012032

Rutter, M. (2007). *Identifying the Environmental Causes of Disease: How Should We Decide what to Believe and when to Take Action?: Report Synopsis*. Academy of Medical Sciences.

Schreier, A., Wolke, D., Thomas, K., Horwood, J., Hollis, C., Gunnell, D., … Harrison, G. (2009). Prospective study of peer victimization in childhood and psychotic Ssymptoms in a nonclinical population at age 12 years. *Archives of General Psychiatry*, *66*(5), 527. https://doi.org/10.1001/archgenpsychiatry.2009.23

Shakoor, S., McGuire, P., Cardno, A. G., Freeman, D., Plomin, R., Ronald, A., … R., P. (2015). A shared genetic propensity underlies experiences of bullying victimization in late childhood and self-rated paranoid thinking in adolescence. *Schizophrenia Bulletin*, *41*(3), 754–763. https://doi.org/http://dx.doi.org/10.1093/schbul/sbu142

Sigfusdottir, I. D., Gudjonsson, G. H., & Sigurdsson, J. F. (2010). Bullying and delinquency. the mediating role of anger. *Personality and Individual Differences*, *48*(4), 391–396. https://doi.org/10.1016/j.paid.2009.10.034

* Silberg, J. L., Copeland, W., Linker, J., Moore, A. A., Roberson-Nay, R., & York, T. P. (2016). Psychiatric outcomes of bullying victimization: a study of discordant monozygotic twins. *Psychological Medicine*, *46*(09), 1875–1883. https://doi.org/10.1017/S0033291716000362

* Singham, T., Viding, E., Schoeler, T., Arseneault, L., Ronald, A., Cecil, C. M., … Pingault, J.-B. (2017). Concurrent and longitudinal contribution of exposure to bullying in childhood to mental health: the role of vulnerability and resilience. *JAMA Psychiatry*, *74*(11), 1112–1119. https://doi.org/10.1001/jamapsychiatry.2017.2678

Stapinski, L. A., Bowes, L., Wolke, D., Pearson, R. M., Mahedy, L., Button, K. S., … Araya, R. (2014). Peer victimization during adolescence and risk for anxiety disorders in adulthood: a prospective cohort study. *Depression and Anxiety*, *31*(7), 574–582. https://doi.org/10.1002/da.22270

Stuart, E. A. (2010). Matching methods for causal inference: a review and a look forward. *Stat. Sci. a Rev. J. Inst. Math. Stat.*, *25*(1), 1. https://doi.org/10.1214/09-STS313

Sweeting, H., Young, R., West, P., & Der, G. (2006). Peer victimization and depression in early–mid adolescence: a longitudinal study. *British Journal of Educational Psychology*, *76*(3), 577–594. https://doi.org/0.1348/000709905X49890

Takizawa, R., Maughan, B., & Arseneault, L. (2014). Adult health outcomes of childhood bullying victimization: evidence from a five-decade longitudinal British birth cohort. *The American Journal of Psychiatry*, *171*(7), 777–784. https://doi.org/10.1176/appi.ajp.2014.13101401

Tsaousis, I. (2016). The relationship of self-esteem to bullying perpetration and peer victimization among schoolchildren and adolescents: a meta-analytic review. *Aggression and Violent Behavior*, *31*, 186–199. https://doi.org/10.1016/j.avb.2016.09.005

Ttofi, M. M., Farrington, D. P., & Lösel, F. (2012). School bullying as a predictor of violence later in life: a systematic review and meta-analysis of prospective longitudinal studies. *Aggression and Violent Behavior*, *17*(5), 405–418. https://doi.org/10.1002/yd.20009

Ttofi, M. M., Farrington, D. P., Lösel, F., Crago, R. V, & Theodorakis, N. (2016). School bullying and drug use later in life: a meta-analytic investigation. *School Psychology Quarterly : The Official Journal of the Division of School Psychology, American Psychological Association*, *31*(1), 8–27. https://doi.org/10.1037/spq0000120

Ttofi, M. M., Farrington, D. P., Lösel, F., & Loeber, R. (2011). Do the victims of school bullies tend to become depressed later in life? A systematic review and meta-analysis of longitudinal studies. *Journal of Aggression, Conflict and Peace Research*, *3*(2), 63–73. https://doi.org/https://doi.org/10.1108/17596591111132873

van Dam, D. S., van der Ven, E., Velthorst, E., Selten, J.-P., Morgan, C., & de Haan, L. (2012). Childhood bullying and the association with psychosis in non-clinical and clinical samples: a review and meta-analysis. *Psychological Medicine*, *42*(12), 2463–2474. https://doi.org/10.1017/S0033291712000360

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45*(2), 576–594. https://doi.org/10.3758/s13428-012-0261-6.

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2015). Meta-analysis of multiple outcomes: a multilevel approach. *Behavior Research Methods*, *47*(4), 1274–1294. https://doi.org/10.3758/s13428-014-0527-2.

van Geel, M., Goemans, A., & Vedder, P. H. (2016). The relation between peer victimization and sleeping problems: A meta-analysis. *Sleep Medicine Reviews*, *27*, 89–95. https://doi.org/10.1016/j.smrv.2015.05.004

Van Geel, M., Vedder, P., & Tanilon, J. (2014). Relationship between peer victimization, cyberbullying, and suicide in children and adolescents: a meta-analysis. *JAMA Pediatrics*, *168*(5), 435–442. https://doi.org/10.1001/jamapediatrics.2013.4143

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

* Vitaro, F., Boivin, M., Brendgen, M., Girard, A., & Dionne, G. (2012). Social experiences in kindergarten and academic achievement in grade 1: a monozygotic twin difference study. *Journal of Educational Psychology*, *104*(2), 366–380. https://doi.org/http://dx.doi.org/10.1037/a0026879

* Vitaro, F., Brendgen, M., Boivin, M., Cantin, S., Dionne, G., Tremblay, R. E., … Pérusse, D. (2011). A monozygotic twin difference study of friends' aggression and children's adjustment problems. *Child Development*, *82*(2), 617–632. https://doi.org/10.1111/j.1467-8624.2010.01570.x

Weisz, J. R., Kuppens, S., Eckshtain, D., Ugueto, A. M., Hawley, K. M., & Jensen-Doss, A. (2013). Performance of evidence-based youth psychotherapies compared with usual clinical care: a multilevel meta-analysis. *JAMA Psychiatry*, *70*(7), 750–761. https://doi.org/10.1001/jamapsychiatry.2013.1176.

Williford, A., Boulton, A., Noland, B., Little, T. D., Kärnä, A., & Salmivalli, C. (2012). Effects of the KiVa anti-bullying program on adolescents' depression, anxiety, and perception of peers. *Journal of Abnormal Child Psychology*, *40*(2), 289–300. https://doi.org/10.1007/s10802-011-9551-1

* Wong, J. S., & Schonlau, M. (2013). Does bully victimization predict future delinquency? A propensity score matching approach. *Criminal Justice and Behavior*, *40*(11), 1184–1208. https://doi.org/http://dx.doi.org/10.1177/0093854813503443

Woods, S., & White, E. (2005). The association between bullying behaviour, arousal levels and behaviour problems. *Journal of Adolescence*, *28*(3), 381–395. https://doi.org/10.1016/j.adolescence.2004.09.002

Yuchang, J., Junyi, L., Junxiu, A., Jing, W., & Mingcheng, H. (2017). The differential victimization associated with depression and anxiety in cross-cultural perspective: a meta-analysis. *Trauma, Violence, & Abuse*, 1524838017726426. https://doi.org/10.1177/1524838017726426

Zeegers, M., Colonnesi, C., Stams, G.-J., & Meins, E. (2017). Mind matters: a three-level meta-analysis on parental mentalization and sensitivity as predictors of infant-parent attachment. *Psychological Bulletin*, *143*(12), 1245–1272. https://doi.org/10.1037/a0028566

Table 1. Quasi-Experimental studies investigating developmental outcome of bullying victimization

Cohort	Publication	Outcomes	Analytical design
National Longitudinal Survey of Youth 1997 (NLSY97), United States	(Wong & Schonlau, 2013)	Carry handgun	Propensity score
		Selling drugs	
		Other property crime	
	(Hoffman et al., 2016)	Academic competence	
		Depression score	
		Sleeping difficulties	
		Conviction	
		Substance use	
		Violence	
	(DeCamp & Newby, 2015)	Vandalism	
		Theft	
		Assault	
		Gang membership	
		Lie/cheat	
		Runaway	
		Sexual intercourse	
		Age of first sexual intercourse	
		Number of sexual partners	
		Alcohol use	
		Age of onset substance use	
		Cannabis use	
		Suspension from school	
		Arrest	
Trends in International Mathematics and Science Study (2011-TIMSS, 8th grade), Ghana	(Kibriya et al., 2015)[1]	Math performance	
Progress in International Reading Literacy Study (2006-PIRLS), Italy, Age 9 cohort	(Ponzo, 2013) [1][2]	Reading literacy	

Trends in International Mathematics and Science Study (2007-TIMSS), Italy, Age 9 cohort	(Ponzo, 2013) [2][2]	Math score	
		Science score	
Trends in International Mathematics and Science Study (2007-TIMSS), Italy, Age 13 cohort	(Ponzo, 2013) [3][2]	Math score Science score	
Third Regional Comparative and Explanatory Study (TERCE), Latin America (15 countries)	(Delprato et al., 2017)	Math score Reading score Sense of belonging Study at home Socialising	
Simmons Longitudinal Study (SLS), United States	(Connell et al., 2017)	Substance use (any) Cigarette use Cannabis use Alcohol use	
Children and Adolescents' Mental Health Promotion Project (CAMHP), Korea	(Roh et al., 2015)	Suicidal ideation Suicide attempt	
Environmental Risk Longitudinal Twin Study (E-Risk), UK	(Arseneault et al., 2008)	Internalizing problems	MZ Discordant twin design
	(Ouellet-Morin et al., 2011)	Perceived stress Cortisol response Body Mass index Pubertal maturity Bullying perpetration Negative affective Scale	
Virginia Twin Study of Adolescent Behavioral Development (VTSABD), United States	(Silberg et al., 2016)	Major depression Social anxiety Separation anxiety Suicidal ideation Overanxious disorder	

		Oppositional defiant disorder Conduct disorder ADHD Generalized anxiety Panic attacks Antisocial personality	
Quebec Newborn Twin Study (QNTS), Canada	(Vitaro et al., 2011) (Vitaro et al., 2012) (M Brendgen et al., 2017)	Aggressive behavior Depressive symptoms Friendship quality Relationship with teacher Externalizing problems Cognitive development Academic achievement Body mass index Pubertal status Physical health problems Relationship quality with mother Relationship quality with father Relationship quality with friend Cortisol awakening response Cortisol level Change in cortisol level	Twin difference design
Twins Early Development Study (TEDS), UK	(Singham et al., 2017)	Anxiety Depression Hyperactivity Conduct Peer problems Prosocial Paranoia Hallucinations Grandiosity Disorganisation Anhedonia Negative symptoms	
National Evaluation of the Gang Resistance Education		Substance use	Fixed effects model

and Training (GREAT) program, America	(McCuddy & Esbensen, 2017)	Violent delinquency Non-violent delinquency

Note. Further details on study characteristics and effect estimates are provided in sTable 1. (Supplementary Material)

[1] The TIMSS data included in the analysis was collected from different (independent) cohorts: (a) 4th grade student that participated in TIMSS 2007 (Italy), (b) 8th grade student that participated in TIMSS 2007 (Italy),

[2] This study included 3 samples that were independent from each other and were therefore considered as separate studies

Table 2. Three-level random effects models: Effects of victimization on outcome

		Internalizing symptoms	Externalizing symptoms	Academic difficulties
Unadjusted MREM models	k_{cohort}	5	6	6
	$k_{outcome}$	13	31	9
	d_{pooled}	0.36	0.34	0.12
	$d_{95\% CI}$	0.03-0.69	0.11-0.57	0.08-0.17
Adjusted MREM models	k_{cohort}	6	7	7
	$k_{outcome}$	17	35	10
	d_{pooled}	0.27	0.15	0.10
	$d_{95\% CI}$	0.05-0.49	0.10-0.21	0.06-0.13
	$\sigma^2_{Level\ 2}$	χ^2=11.05, p=0.0009	χ^2=102.56, p<0.001	χ^2=0.00, p=1.00
	$\sigma^2_{Level\ 3}$	χ^2=9.25, p=0.0024	χ^2=9.25, p=0.81	χ^2=4.00, p=0.045
	$I^2_{Level\ 1}$	10.79%	17.04%	27.62%
	$I^2_{Level\ 2}$	12.30%	75.69%	<0.0001%
	$I^2_{Level\ 3}$	77.91%	7.27%	72.38%
	Publ. bias	t=0.496, p=0.69	t=1.104, p=0.28	t=0.519, p=0.62

I^2= % of the total variance accounted for by random sampling variance (Level 1), variation within cohorts (Level 2), variation between cohorts (Level 3); $\chi 2$ = Statistics from likelihood-ratio test to test within-cohort variance ($\sigma^2_{Level\ 2}$) and between-cohort variance ($\sigma^2_{Level\ 3}$) for significance; MREM= Multilevel random effects model

Table 3. Moderator analysis: sources of heterogeneity

	Subgroup	k_{cohort}	$k_{outcome}$	d_{MREM}	95% CI_{MREM}	$p_{moderator}$
Level of adjustment[a]						
All outcomes	Unadjusted	13	68	0.23	0.14-0.33	0.0002
	Adjusted	13	89	0.13	0.08-0.18	
Internalizing symptoms	Unadjusted	5	13	0.36	0.03-0.69	0.20
	Adjusted	6	17	0.27	0.05-0.49	
Externalizing symptoms	Unadjusted	6	31	0.34	0.11-0.57	0.006
	Adjusted	7	35	0.15	0.10-0.21	
Academic difficulties	Unadjusted	6	9	0.12	0.08-0.17	0.78
	Adjusted	7	10	0.10	0.06-0.13	
Rater-effect[b]						
All outcomes[c]	Shared	5	51	0.13	0.05-0.22	0.26
	Non-shared	11	43	0.10	0.07-0.13	
Internalizing symptoms	Shared	3	6	0.37	-0.03-0.77	<0.0001
	Non-shared	4	13	0.14	0.05-0.23	
Externalizing symptoms	Shared	3	24	0.18	0.09-0.27	0.12
	Non-shared	5	13	0.06	0.01-0.11	
Persistence of effect[b]						
All outcomes	Short-term	11	54	0.15	0.07-0.23	0.004
	Long-term	4	44	0.11	0.07-0.14	
Internalizing symptoms	Short-term	5	13	0.33	0.10-0.57	0.016
	Long-term	3	8	0.06	-0.01-0.13	
Externalizing symptoms	Short-term	5	11	0.17	0.08-0.27	0.28
	Long-term	4	26	0.13	0.09-0.17	
Academic difficulties	Short-term	6	9	0.10	0.06-0.14	0.42
	Long-term	1	1	0.04	-0.07-0.16	

[a] Test of moderation through subsets, including level of adjustment as a dichotomized variable [1: adjusted estimates (estimates derived from Quasi-Experimental models), 2: unadjusted models (estimates derived from uncontrolled models)]

[b] Test of moderation through subsets, including age as a dichotomized variable [Non-shared rater= bullying victimization and outcome assessed by different individuals; Shared=bullying victimization and outcome reported by the same individual]. The shared rater effect could not be tested for 'academic difficulties', since all studies for this outcome relied on non-shared rater methods.

[c] All outcomes including internalizing symptoms, externalizing symptoms, academic difficulties and all outcomes classified as 'NOS' (not otherwise specified)

[b] Test of moderation through subsets, including age as a dichotomized variable [long-term effects= > 1 years of follow up; short-term effect = 1 year or less of follow up]

Note: MREM = Multilevel random effects model

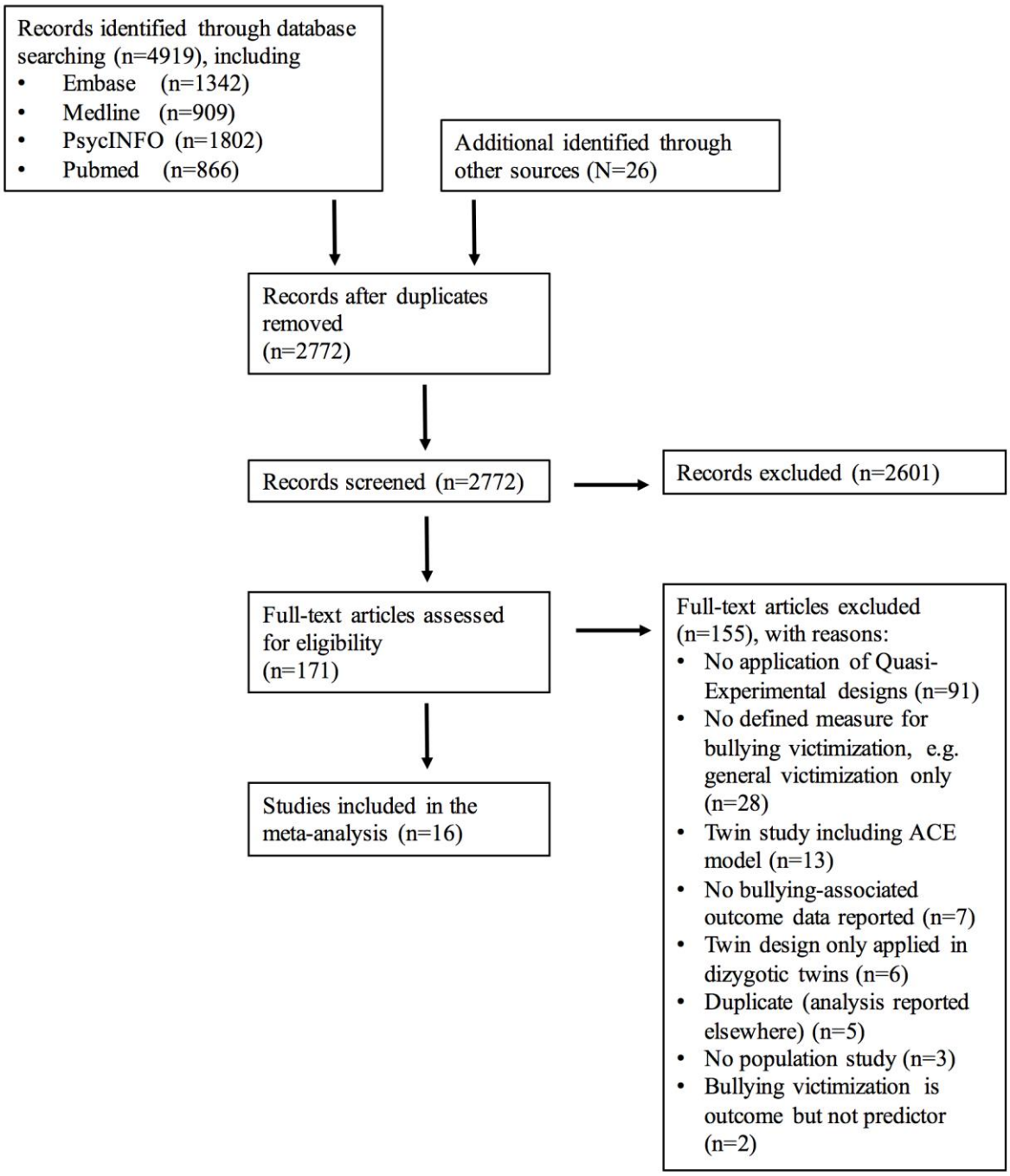

Figure 1. Flow chart

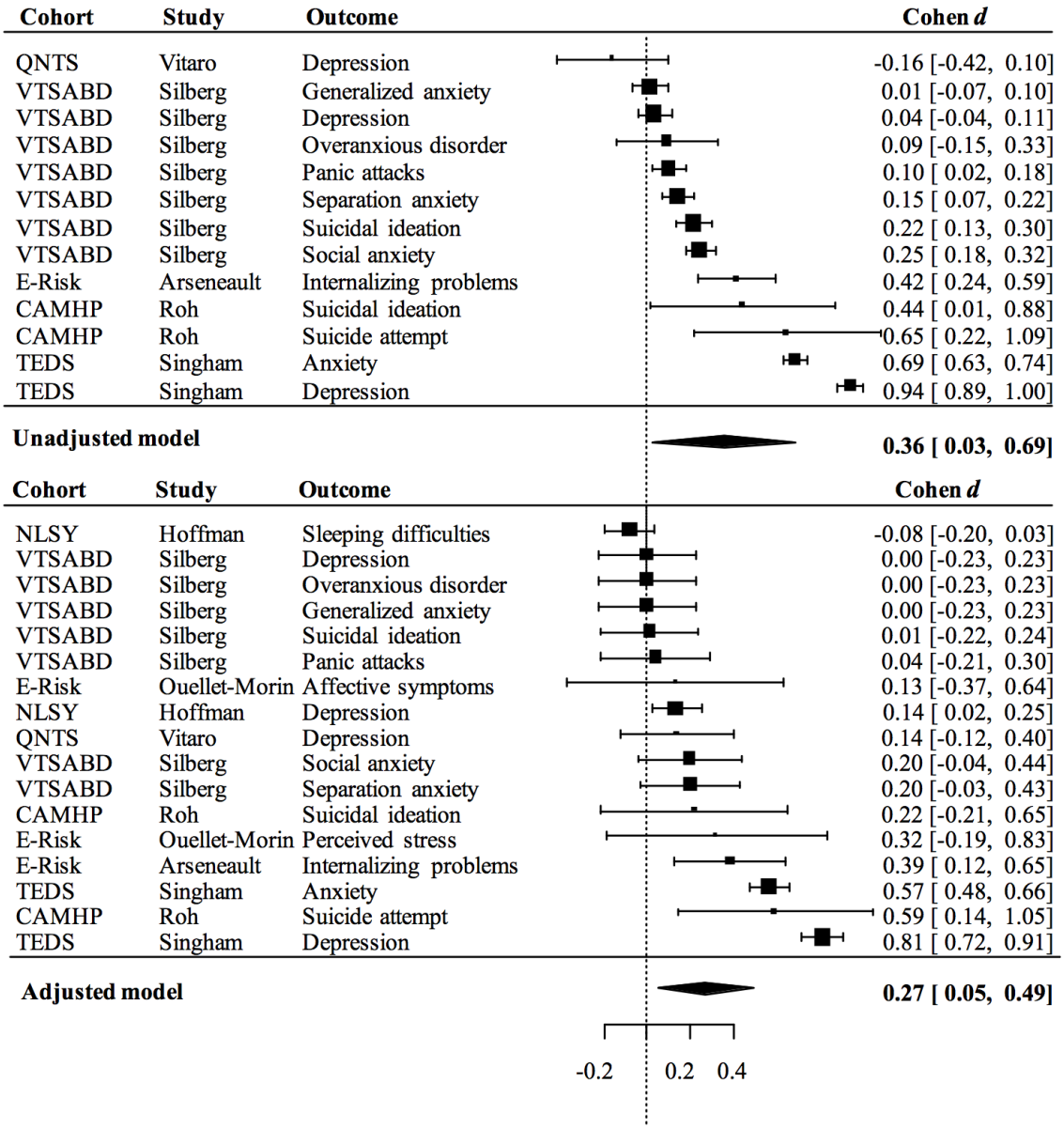

Figure 2. Multilevel random effects model for internalizing symptoms

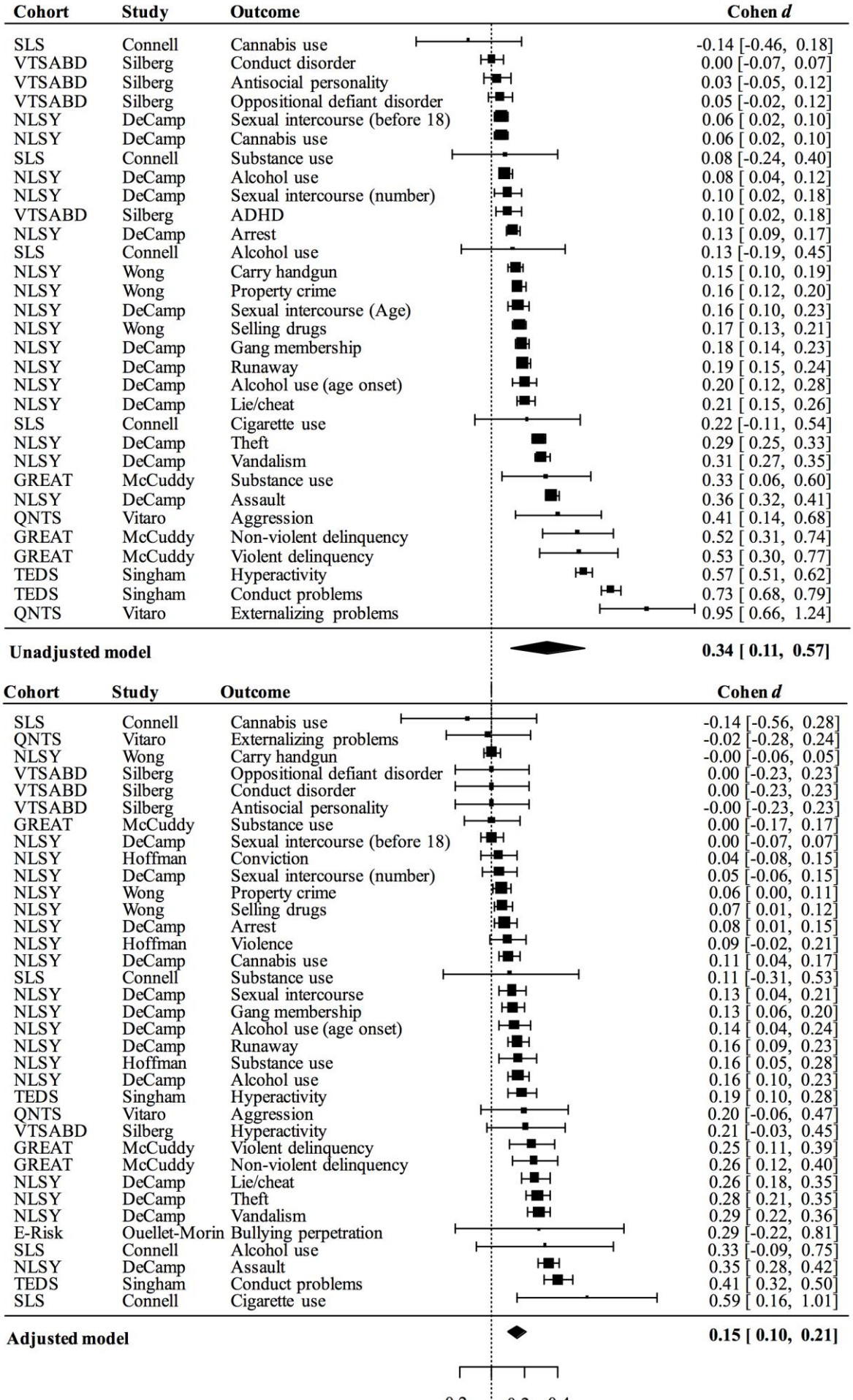
Cohort Study Outcome Cohen d
SLS Connell Cannabis use -0.14 [-0.46, 0.18]
VTSABD Silberg Conduct disorder 0.00 [-0.07, 0.07]
VTSABD Silberg Antisocial personality 0.03 [-0.05, 0.12]
VTSABD Silberg Oppositional defiant disorder 0.05 [-0.02, 0.12]
NLSY DeCamp Sexual intercourse (before 18) 0.06 [0.02, 0.10]
NLSY DeCamp Cannabis use 0.06 [0.02, 0.10]
SLS Connell Substance use 0.08 [-0.24, 0.40]
NLSY DeCamp Alcohol use 0.08 [0.04, 0.12]
NLSY DeCamp Sexual intercourse (number) 0.10 [0.02, 0.18]
VTSABD Silberg ADHD 0.10 [0.02, 0.18]
NLSY DeCamp Arrest 0.13 [0.09, 0.17]
SLS Connell Alcohol use 0.13 [-0.19, 0.45]
NLSY Wong Carry handgun 0.15 [0.10, 0.19]
NLSY Wong Property crime 0.16 [0.12, 0.20]
NLSY DeCamp Sexual intercourse (Age) 0.16 [0.10, 0.23]
NLSY Wong Selling drugs 0.17 [0.13, 0.21]
NLSY DeCamp Gang membership 0.18 [0.14, 0.23]
NLSY DeCamp Runaway 0.19 [0.15, 0.24]
NLSY DeCamp Alcohol use (age onset) 0.20 [0.12, 0.28]
NLSY DeCamp Lie/cheat 0.21 [0.15, 0.26]
SLS Connell Cigarette use 0.22 [-0.11, 0.54]
NLSY DeCamp Theft 0.29 [0.25, 0.33]
NLSY DeCamp Vandalism 0.31 [0.27, 0.35]
GREAT McCuddy Substance use 0.33 [0.06, 0.60]
NLSY DeCamp Assault 0.36 [0.32, 0.41]
QNTS Vitaro Aggression 0.41 [0.14, 0.68]
GREAT McCuddy Non-violent delinquency 0.52 [0.31, 0.74]
GREAT McCuddy Violent delinquency 0.53 [0.30, 0.77]
TEDS Singham Hyperactivity 0.57 [0.51, 0.62]
TEDS Singham Conduct problems 0.73 [0.68, 0.79]
QNTS Vitaro Externalizing problems 0.95 [0.66, 1.24]
Unadjusted model 0.34 [0.11, 0.57]
Cohort Study Outcome Cohen d
SLS Connell Cannabis use -0.14 [-0.56, 0.28]
QNTS Vitaro Externalizing problems -0.02 [-0.28, 0.24]
NLSY Wong Carry handgun -0.00 [-0.06, 0.05]
VTSABD Silberg Oppositional defiant disorder 0.00 [-0.23, 0.23]
VTSABD Silberg Conduct disorder 0.00 [-0.23, 0.23]
VTSABD Silberg Antisocial personality -0.00 [-0.23, 0.23]
GREAT McCuddy Substance use 0.00 [-0.17, 0.17]
NLSY DeCamp Sexual intercourse (before 18) 0.00 [-0.07, 0.07]
NLSY Hoffman Conviction 0.04 [-0.08, 0.15]
NLSY DeCamp Sexual intercourse (number) 0.05 [-0.06, 0.15]
NLSY Wong Property crime 0.06 [0.00, 0.11]
NLSY Wong Selling drugs 0.07 [0.01, 0.12]
NLSY DeCamp Arrest 0.08 [0.01, 0.15]
NLSY Hoffman Violence 0.09 [-0.02, 0.21]
NLSY DeCamp Cannabis use 0.11 [0.04, 0.17]
SLS Connell Substance use 0.11 [-0.31, 0.53]
NLSY DeCamp Sexual intercourse 0.13 [0.04, 0.21]
NLSY DeCamp Gang membership 0.13 [0.06, 0.20]
NLSY DeCamp Alcohol use (age onset) 0.14 [0.04, 0.24]
NLSY DeCamp Runaway 0.16 [0.09, 0.23]
NLSY Hoffman Substance use 0.16 [0.05, 0.28]
NLSY DeCamp Alcohol use 0.16 [0.10, 0.23]
TEDS Singham Hyperactivity 0.19 [0.10, 0.28]
QNTS Vitaro Aggression 0.20 [-0.06, 0.47]
VTSABD Silberg Hyperactivity 0.21 [-0.03, 0.45]
GREAT McCuddy Violent delinquency 0.25 [0.11, 0.39]
GREAT McCuddy Non-violent delinquency 0.26 [0.12, 0.40]
NLSY DeCamp Lie/cheat 0.26 [0.18, 0.35]
NLSY DeCamp Theft 0.28 [0.21, 0.35]
NLSY DeCamp Vandalism 0.29 [0.22, 0.36]
E-Risk Ouellet-Morin Bullying perpetration 0.29 [-0.22, 0.81]
SLS Connell Alcohol use 0.33 [-0.09, 0.75]
NLSY DeCamp Assault 0.35 [0.28, 0.42]
TEDS Singham Conduct problems 0.41 [0.32, 0.50]
SLS Connell Cigarette use 0.59 [0.16, 1.01]
Adjusted model 0.15 [0.10, 0.21]

Figure 3. Multilevel random effects model for externalizing symptoms

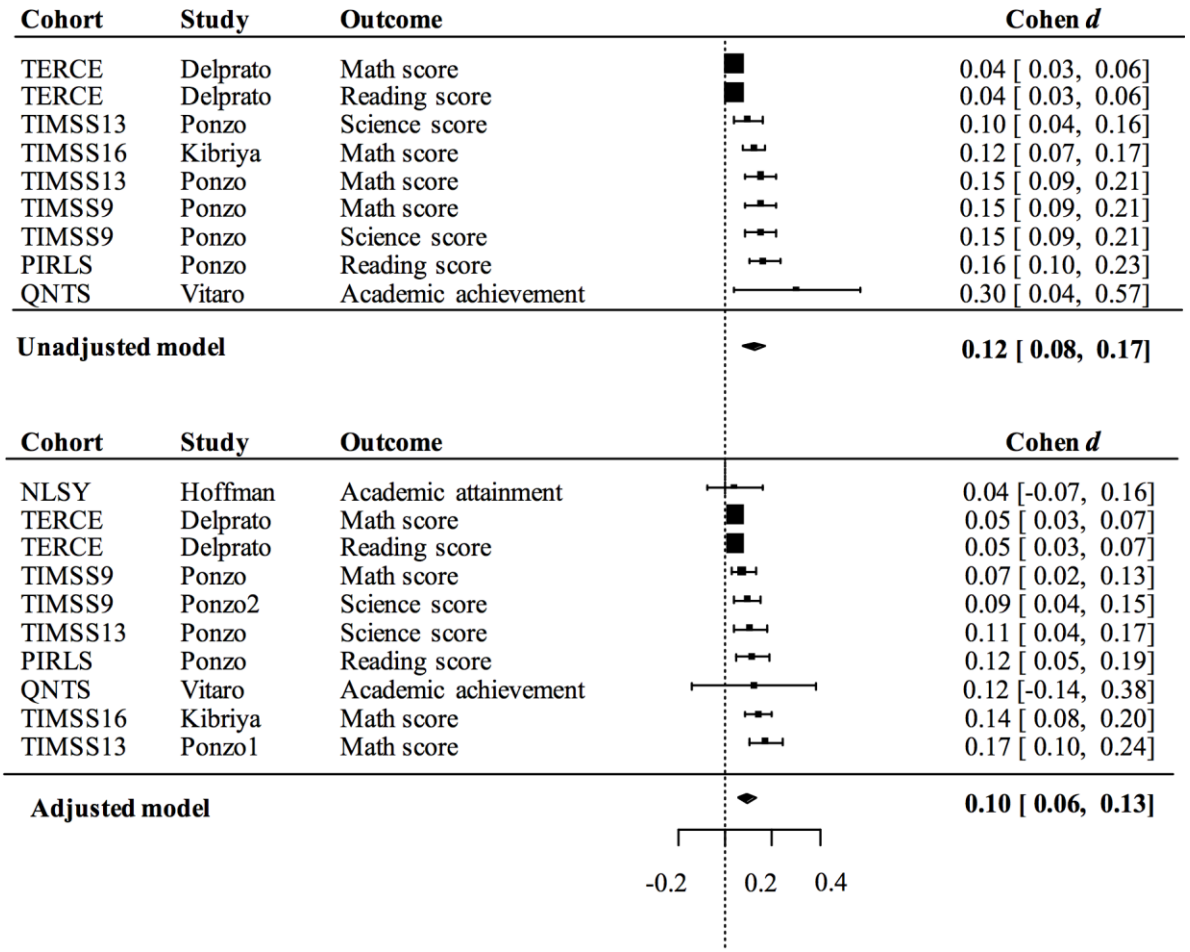

Figure 4. Multilevel random effects model for academic difficulties