

RESEARCH ARTICLE

CACE closed: A multiverse examination of the influence of implementation variability on student outcomes in a randomised controlled trial of a universal, school-based social-emotional learning intervention

Annie O'Brien^{*}, Joao Santos^{ID}, Neil Humphrey, Margarita Panayiotou

Manchester Institute of Education, The University of Manchester, Manchester, United Kingdom

* annie.obrien@manchester.ac.uk



Abstract

Introduction

Amidst calls for more high-quality research to assess the influence of implementation variability on student outcomes in school-based trials using instrumental variable approaches (e.g., complier average causal effect; CACE), unreported researcher degrees of freedom can limit replicability and lead to uncertainty of conclusions. The current study aims to acknowledge and address these limitations using a multiverse framework. We investigate whether, and how, conclusions of a universal social-emotional learning intervention's efficacy are contingent on decisions about how compliance is defined and modelled.

Methods

Secondary analysis of data from a cluster randomised control trial of the intervention Passport was undertaken, with schools ($k=62$, $N=2,425$ children) randomly allocated to intervention ($k=33$; $N=1,264$) or control ($k=29$; $N=1,161$) conditions. Ten theoretically plausible specifications for the CACE model were identified and pre-registered, including five definitions of compliance (fidelity, dosage, quality, responsiveness, reach) and two compliance thresholds (50th and 75th percentile). Student relational outcomes (bullying, peer support, loneliness) were assessed pre- and post-intervention.

Results

Multilevel intent-to-treat analysis revealed null intervention effects. Applying a multiverse framework to CACE revealed variation in model results, manifest in entropy values, the precision of confidence intervals and the direction, size and statistical significance of CACE effects. A statistically significant and negative CACE effect was

OPEN ACCESS

Citation: O'Brien A, Santos J, Humphrey N, Panayiotou M (2026) CACE closed: A multiverse examination of the influence of implementation variability on student outcomes in a randomised controlled trial of a universal, school-based social-emotional learning intervention. PLoS One 21(6): e0349949. <https://doi.org/10.1371/journal.pone.0349949>

Editor: Nate Breznau, German Institute for Adult Education - Leibniz Center for Lifelong Learning, GERMANY

Received: November 18, 2025

Accepted: May 7, 2026

Published: June 2, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pone.0349949>

Copyright: © 2026 O'Brien et al. This is an open access article distributed under the terms

of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: The data for this study is available on the OSF: 10.17605/OSF.IO/S5PMW.

Funding: This work is part of the first author's PhD research, funded by the Kavli Trust (<https://kavlifondet.no/en/> [kavlifondet.no]), grant no. Kavli2021-0000000019. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors report no conflict of interest.

found for peer support when compliance was defined by reach using the 75th percentile ($\beta = -.38$, 95% CI $(-.68, -.08)$, $E = .71$, $d = -.21$), but non-statistically significant intervention effects were observed for the remaining CACE models.

Conclusions

A multiverse framework enables transparent reporting of analytic uncertainty in evaluations of implementation variability in school-based trials, thereby offering theoretical, methodological and empirical advancements for implementation science. In doing so, it enables us to move from a fragmented view towards a more coherent understanding of complex interventions in real-world settings.

Pre-registration

www.osf.io/s5pmw.

Introduction

Amidst rising mental health problems among young people [1], there is robust evidence that school-based social-emotional learning (SEL) programmes can improve children's social-emotional, behavioural, mental health and academic outcomes in the short and long-term [2–4]. However, smaller effect estimates are often reported when interventions are transported from controlled efficacy trials into real-world contexts [5] and may be explained by implementation variability [4,6–8].

Implementation variability is the difference in how a programme is theorised to how it is actioned in practice [6]. It is a multi-dimensional concept [6], such that delivery can vary across the following dimensions: dosage (the number of lessons delivered); fidelity (the key lesson objectives and core components according to the programme manual); quality (how well facilitators deliver programme components and engage recipients); responsiveness (recipient engagement and participation); adaptations (changes made during delivery), and reach (the proportion of recipients present). A review of prevention programmes reported that average effect size estimates were 2–3 times greater when delivered well, whereas under ideal conditions, effect sizes could be improved 12-fold, however full implementation (100%) is rarely reported and/or achieved [6].

The conclusion that “implementation matters” has been corroborated through recent systematic reviews that have revealed a relationship between implementation variability and student outcomes in universal school-based prevention interventions [9,10]. However, further research using high quality methods is required to further understanding of this relationship [9].

Capitalising on recent advancements in implementation science, we employ the instrumental variable approach CACE (Complier Average Causal Effect) to evaluate implementation variability of a universal, school-based SEL intervention ‘Passport: Skills for life’ (hereafter Passport) [11] on understudied relational outcomes (loneliness, bullying, peer social support). To address limitations of CACE, namely its

reliance on arbitrary decisions (detailed below), we use a multiverse approach to improve transparency and credibility of conclusions [12]. Thus, this research makes theoretical, methodological and empirical advancements to the fields of implementation science and SEL. We now present a three-part discussion of our aims.

Addressing social relationships through SEL interventions

Loneliness and childhood victimisation are well-established risk factors associated with problematic mental and physical health and academic functioning [13–15]. Loneliness can occur when there is a perceived difference between the actual and desired quantity and/or quality of one's relationships [16] and is associated with long-lasting negative health outcomes for youth [e.g., depression, anxiety; 17,18]. With experiences of loneliness found to peak in early adolescence and schools identified as critical environments for targeting and alleviating loneliness [17], the UK Government's loneliness strategy aims to prevent loneliness in youth by nurturing connected school communities [19].

Indeed, schools have a duty of care to children, to nurture student growth and to prevent bullying behaviours [20]. Bullying is intentional and reoccurring exclusionary or aggressive behaviour over time and in the presence of a power imbalance between perpetrator and victim [21]. It is associated with poorer academic performance and psychological, behavioural, and social maladjustment for both the bully and the victim [13,22,23]. Approximately 35% of 10–15-year-olds in England and Wales were bullied between 2022–23 [24] with a prevalence of 16% reported in Greater Manchester [25].

Peer social support is an important protective factor in child mental health [26], and moderates negative effects of stressful life events (e.g., bullying). It is defined as access to and receipt of emotional or physical support and resources through interactions with those in one's social network [27,28]. Peer relationships become increasingly important during early adolescence, and young people turn primarily to peers for advice or support when experiencing distress [29]. The UK Government thus aim to support the development of quality peer relationships in schools to improve children's mental health [30].

Children who have poor peer relationships (e.g., experience victimisation, social exclusion or lack close friends) are most likely to experience loneliness [17,31–33]. The 'friendship protection hypothesis' suggests that peer relationships protect against negative experiences and their negative outcomes [34] including loneliness, bullying perpetration and victimisation [35–37]. The importance of fostering caring school environments that promote positive peer relationships is therefore critical [17].

Although meta-analytic evidence demonstrates that SEL interventions improve social-emotional skills (effect size [ES] = 0.21–0.57), peer relationships (ES = 0.22) and reduce bullying behaviours (ES = 0.16–0.22) and victimisation (ES = 0.14–0.24) [2,3], the influence of implementation variability on loneliness and peer support remain underexplored [38,39]. Null intent-to-treat (ITT) effects were found when examining the impact of the PATHS SEL programme (Promoting Alternative THinking Strategies) on social support [40], however moderate intervention effects were revealed in moderate-high compliance classrooms (ES = 0.63) [41]. There is also emerging empirical support that school-based interventions targeting social-emotional competencies can alleviate loneliness [42] and the PATHS programme reduced experiences of loneliness among children in the intervention group relative to the control (ES = 1.35–1.68) [39]. The current study aims to contribute to this building evidence base by examining the impact of Passport on these relational outcomes. While a focus on SEL and relational outcomes in this study is therefore advantageous from a developmental perspective, this decision is also pragmatic (this is a secondary data analysis, as detailed in the Methods).

Underpinned by coping theory, Passport primarily aims to help children identify and use positive coping strategies and to offer, seek and accept help [11]. Lessons on 'Relationships' and 'Dealing with Conflict' aim to support children in making friends, dealing with bullying, and use adaptive coping skills to negate feelings of loneliness or rejection that can occur when a friendship ends. Passport is currently implemented in c.115 schools across the UK; evaluating its efficacy in building relation skills in this context is therefore timely [43]. No research to date has examined the effectiveness of SEL

interventions on loneliness and peer victimisation after accounting for implementation variability in a causal framework [9]. The current study aims to address this research gap.

Implementation matters in school-based trials

The standard approach to evaluating intervention effects is through applying an ITT framework which upholds the bias-protective principle of randomisation ("once randomised, always analysed"), yet accurate causal inference is limited by the assumption of full compliance (i.e., that the intervention group delivered the intervention in its entirety without deviating from the manual/protocol). Unlike pharmaceutical and medical trials where treatment compliance is closely monitored and/or controlled, the assumption does not hold in school-based trials amidst theoretical support for, and empirical evidence of, variation in intervention delivery [2]. Per-protocol and as-treated are often used alongside ITT to address implementation variability, yet these traditional methods produce biased effect estimates (e.g., by excluding data of non-compliers following randomisation and disrupting group equivalence).

A simplistic solution is to focus solely on the ITT effect [44], yet this can lead to biased estimates of true intervention effects [45], which can manifest in two ways. If participants experience adverse intervention effects and consequently discontinue treatment, but participant adherence is not addressed in the analysis, positive intervention effects are overestimated. Conversely, if participants fail to adhere to the protocol despite positive outcomes, intervention effects are underestimated. Thus, considering the average intervention effect in the context of compliance *alongside* the ITT effect (i.e., the average intervention effect across all participants offered the intervention) provides valuable insight into the true intervention impact among those who *received* the intervention.

There is a call for educational research to move "beyond ITT" and use instrumental variable approaches [46] such as propensity score analysis, treatment effect bounding, and CACE [45]. These approaches use a causal framework that specifically models participant compliance and are popular in the medical, pharmaceutical and psychological literature. Though increasingly employed, they are uncommon in school-based trials [47]. This study focuses on the use of CACE [48] which estimates the impact of the intervention among the subpopulation of compliers [45]. It does so by using random assignment as an instrumental variable and by applying assumptions, such as monotonicity and exclusion restrictions, to infer the proportion of compliers. Although CACE is a robust analytic choice [47], its application to education is not without limitations. Identifying and overcoming these limitations is an issue we now address.

Overcoming limitations to CACE using a multiverse framework

There are a number of limitations to CACE. First, school-based trials typically model compliance using dosage [49,50], as quantity of lessons delivered is considered an appropriate proxy for estimating the impact of both *offering* the intervention (i.e., ITT effect) and intervention *receipt* (i.e., CACE effect) when school attendance is mandatory. However, dose offered and dose received are conceptually distinct [51]. One could therefore argue that *delivering* an intervention session (i.e., dosage) does not guarantee intervention *receipt* if the programme objectives or core components are not delivered (i.e., poor fidelity); if lessons are delivered with poor quality; if students are not engaged or actively participating during lessons (i.e., low responsiveness), or if students are absent (i.e., low reach). The reliance on dosage as the sole indicator of intervention receipt in CACE models is criticised as arbitrary [52], arguably neglecting empirical evidence that interventions can yield positive outcomes even at low dosage levels [53,54]. Our recent systematic review revealed that multiple SEL implementation dimensions (fidelity, dosage, reach, responsiveness, quality) were associated with students outcomes and could therefore be considered as potential indicators of intervention receipt [9].

Second, compliance is modelled as a binary variable in a CACE model, yet defining 'full' compliance of a prevention intervention is significantly challenging. CACE is dependent on the exclusion restriction assumption (non-compliers derive no benefit from the intervention; see Data analysis) which is difficult to validate in the absence of theory-informed compliance thresholds [55] which are largely untested and unknown in psychological and educational interventions [56].

According to the developer of Passport, there is no theoretically informed threshold for optimal programme delivery and empirical research has not established a definition of compliance in the context of this intervention [e.g., 11]. An inductive approach to defining compliance is an alternative solution, but selecting the cut-off point for defining a complier and non-complier is “somewhat arbitrary along the percent compliance continuum” [46, p.22]. Researchers are encouraged to use multiple sensitivity analyses to observe any change across compliance thresholds [46]. Using the 50th and 75th percentile to define compliance is established practice in educational research [41,49,50,57,58].

The selection of a compliance definition and threshold can impact outcomes of the CACE model. Typically, methodological choices available to researchers range in quality, with strong theoretical and/or empirical support guiding selection of a superior approach. However, when there are numerous sound and theoretically justifiable decisions, selecting one over another is arbitrary [12]. Arbitrary data processing decisions can constrain a dataset and subsequently limit the validity of the conclusions drawn from it [59]. To increase transparency of these limitations, we constructed a multiverse of possible modelling choices for our CACE model.

Multiverse analysis systematically explores how different, equally defensible analytic choices impact research findings, to evaluate the robustness of results and transparently convey uncertainty [12]. A multiverse approach will reveal the instability or robustness of results that hinge on these data processing decisions and help to determine the moderating role of each implementation dimension on outcomes, and the level of implementation that yields desired intervention effects, thus addressing multiple and pertinent research priorities [7,8]. Crenshaw and colleagues [60] demonstrate the strength of this framework in a clinical trial for treatment of post-traumatic stress disorder amidst challenges with participant recruitment, differential attrition between conditions, and missing follow-up data. A multiverse approach facilitated examination of the robustness of results that were based upon pre-registered analytic decisions (e.g., ITT) versus alternative analytic choices that addressed unexpected challenges (e.g., as-treated), demonstrating the treatments’ efficacy across all analytic universes and thus ensuring confidence in conclusions drawn. Despite its strength in transparently revealing the fragility or robustness of conclusion, multiverse analysis remains underutilised in this field [61]. Our study extends the application of this framework to school-based intervention research.

The current study

This paper is a response to the call for increased methodological rigor when assessing implementation variability in education research [9]. We employ CACE to model implementation variability within a causal framework [45], a rigorous analytic method, however unreported researcher degrees of freedom can limit replicability and credibility of conclusions [62]. We therefore apply a multiverse framework to transparently report uncertainty [12].

This research thus aims to address the following research questions: (1) What is the impact of Passport on primary school-aged children’s relational outcomes (i.e., loneliness, peer social support, and victimisation)? (2) What is the impact of Passport on primary school-aged children’s relational outcomes when accounting for intervention compliance? (3) How does the complier average causal effect of Passport on pupil outcomes vary by compliance definition?

Method

Design

A two-arm (intervention vs usual practice) parallel cluster randomised controlled trial design was used. The Passport to Success trial took place in mainstream primary schools across Greater Manchester between September 2022–2025. Passport was implemented in Year 5 classrooms of participating schools in the intervention arm during the 2023/2024 academic year. Schools allocated to the control arm continued delivering the usual school provision. Ethical approval for the trial was obtained from the University of Manchester Research Ethics Committee [Ref: 2022-14050-24401]. The trial was prospectively registered with ISRCTN [ISRCTN12875599] and detailed methodology is

available in the published protocol [43]. This paper reports on methods pertaining to the current study only, which was preregistered on the OSF: www.osf.io/s5pmw.

Sample

Sixty-two schools participated in the current study. Eligible schools were mainstream, non-independent primary schools situated in Greater Manchester, delivering education to Year 4 – Year 6 students and had not previously delivered Passport. The final student sample consisted of 2425 pupils (50.1% male) in 79 classrooms. Students were eligible to take part in the trial if they were in a Year 4 classroom in a participating school in the 2022/2023 academic year, they had not been opted out of the study and they assented to participate. At baseline, the proportion of participating students that were eligible for free school meals (FSM; 34%) and had special educational needs (SEN; 18%) were greater than the national average [63]. Participant demographics are reported in Table 1. All variables, bar FSM, were balanced between trial arms [see 64].

Procedure

School recruitment commenced 20.09.2022 and concluded 21.03.2023. The trial manager (JS) contacted Local Authorities in Greater Manchester to nominate local schools and received signed Memorandum of Agreement and a Data Sharing Agreement from sixty-two schools.

Parents/carers of Year 4 children received an information sheet and the opportunity to opt their child out of participating in data collection. In line with schools' in loco parentis responsibilities, children opted out of the study in intervention

Table 1. Participant demographic information and descriptive statistics for study variables.

Variables	Mean (standard deviation)	
	Passport	Control
Student-level (N)	1264	1161
Sex % (male/female)	50.7/49.2	49.5/50.4
FSM % (no/yes)	70.6/38.2	67/32.9
SEN % (no/yes)	82.4/17.5	82.5/17.4
Peer support T0	49.75 (9.72)	48.75 (10.22)
Peer support T1	49.57 (9.80)	49.38 (10.19)
Victimisation T0	49.55 (10.36)	49.71 (10.50)
Victimisation T1	50.06 (10.26)	50.80 (9.41)
Loneliness T0*	4.96 (1.61)	5.13 (1.69)
Loneliness T1*	4.76 (1.64)	4.81 (1.64)
Teacher-level (N)	44	35
Gender (male/female)	16/28	11/24
Teaching years	10.8 (9.5)	12.6 (11.3)
Stress T0*	7.08 (2.2)	7.56 (1.56)
Coping with stress T0	6.39 (2.45)	6.03 (2.16)
Attitudes towards SEL T0	49.6 (4.7)	50.28 (4.7)
Dosage (%)	92.67 (19.28)	–
Fidelity (%)	87.99 (16.25)	–
Reach (%)	89.44 (16.05)	–
Responsiveness (%)	82.08 (19.43)	–
Quality of delivery (%)	88.38 (14.41)	–

Note. T0 = pre-intervention; T1 = post-intervention. *Higher scores indicate negative outcomes.

<https://doi.org/10.1371/journal.pone.0349949.t001>

schools still received the Passport programme, which was delivered as part of the school curriculum. Participating children assented prior to completing baseline outcome surveys (T0; April – July 2023), administered via Qualtrics.

Randomisation took place following baseline survey data collection by an independent statistician with schools as the unit of randomisation. The allocation sequence was generated using MinimPy software [65]. Minimisation was used to ensure balance across trial arms for school size and proportion of students eligible for FSM.

Year 5 teachers in intervention schools received mandatory 2-hour online training in delivering Passport from Partnership for Children in September 2023; a 1-hour optional booster training session was offered in February 2024 and aimed to address any challenges that arose during the initial months of programme delivery. Post-intervention data collection took place in April – July 2024 (T1).

Intervention

The reporting of the intervention is guided by the TIDieR checklist [66]. Passport follows the adventures of two children as they navigate a fantasy world with dragons and mythical creatures. There are 18 sequential lessons (approx. 55 minutes) delivered once weekly on topics of emotions, relationships and helping others, dealing with difficult situations (e.g., bullying), fairness and justice, and coping with change and loss. Passport's theory of change posits that fostering coping flexibility will help children self-regulate, inhibiting the onset of mental distress [11]. The programme's logic model is illustrated in Humphrey et al. [64].

Lessons are student-led and delivered during class time by Year 5 teachers using physical and digital resources [67], including comic strips, board games and role play. Teachers must adhere to the structure, sequence and content of the programme manual however they are encouraged to *augment* lessons, personalising them as necessary to meet the individual needs of their class.

Usual practice

Schools allocated to the control arm of the trial continued with usual practice. A usual practice survey administered to all participating classroom teachers revealed that approximately two thirds of the topics covered by Passport were taught across participating schools (e.g., identifying and managing emotions) and two thirds of classrooms delivered propriety SEL curricula (e.g., Jigsaw).

Outcome measures

Student outcomes. The self-report UCLA loneliness scale for children, recommended by the Office for National Statistics [68], captures loneliness via peer exclusion and isolation (e.g., 'How often do you feel that you have no one to talk to?') using three response categories (e.g., hardly ever or never, some of the time, often). The scale is validated among 10–15-year olds [68] and has strong internal reliability ($\alpha=0.87-0.89$) [69,70]. Alpha for the current sample at T0 was .71. Scores are summed, with higher scores indicating greater experiences of loneliness.

Peer victimisation was assessed using the 3-item social acceptance dimension of the KIDSCREEN-52 (KS52) health related quality of life (HRQoL) measure [e.g., 'Thinking about the last week, have other girls and boys made fun of you?'; 71]. Items are reverse scored, with higher scores indicating lower bullying incidence. The scale has strong internal reliability ($\alpha=0.77$) and convergent validity [Pearson's $r=0.29-0.32$; 71]. Internal consistency for the current sample at T0 was .75.

Peer support was measured with the 4-item Social Support and Peers dimension of the KIDSCREEN-27 (KS27) HRQoL [72]. The scale captures participant's relationships with other children (e.g., 'Thinking about the last week, have you been able to rely on your friends?'), has strong test-retest reliability (ICC=0.61), criterion validity ($r=0.94$) and convergent validity [$r=0.36$; 72]. Alpha for the current sample at T0 was .71. Scores are summed, and higher scores indicate positive peer relationships.

Both the KS27 and KS52 are self-report measures with 5 response items (e.g., Never, seldom, quite often, very often, always) and have been cross-culturally validated among young people aged 8–18-years [71,73]. Scores were adjusted for gender and age, following the manual guidelines.

Compliance definition. Implementation data was collected at a single time point (T1) using a teacher self-report survey (see [S1 File](#)). Previous conceptual and empirical work informed survey development [e.g., 6,38] which was reviewed by the research team prior to distribution. Survey items captured dosage (n=18), fidelity (n=4), quality of delivery (n=4), student responsiveness (n=3), and reach (n=1). Total scores for each dimension were multiplied by the number of sessions delivered by the respective staff to yield dosage-adjusted implementation scores for each dimension, thus accounting for programme exposure.

Covariates and compliance predictors. The approach to selecting the covariates and compliance predictors is detailed in [S2 File](#). Baseline scores across loneliness, peer support and victimisation were included as child-level covariates, and the following data provided by schools included as student-level compliance predictors in the CACE models: pupil FSM eligibility (yes/no), special educational needs status (SEN; yes/no) and sex (male/female).

A staff survey, issued at baseline, included single items to capture perceived stress ('How stressful is your job?') and coping with stress ('How well are you coping with the stress of your job right now?') using an 11-point Likert scale, ranging from not stressful/not well, to very stressful/very well, respectively. Both items have good concurrent and predictive validity [74]. Teacher attitudes towards SEL were measured using the Teacher SEL Beliefs Scale, which includes 12 items (e.g., 'I want to improve my ability to teach social and emotional skills to students') and a 5-point Likert scale (1 = strongly disagree; 5 = strongly agree).

Data analysis

CACE overview and assumptions. Angrist and colleagues [75] described 4 patterns of participant compliance behaviour in the context of binary condition assignment (T_i ; 0 = control, 1 = treatment) and treatment received (D_i ; 0 = not received, 1 = received) for an individual i . There are compliers (those who adhere to their condition assignment; $D_i(T_i = 1) = 1$, $D_i(T_i = 0) = 0$); defiers (those who do not adhere to their condition assignment; $D_i(T_i = 1) = 0$, $D_i(T_i = 0) = 1$); always-takers (i.e., those who seek out and receiving treatment despite condition assignment; $D_i(T_i = 1) = D_i(T_i = 0) = 1$) and never-takers (those who do not take the treatment regardless of condition assignment; $D_i(T_i = 1) = D_i(T_i = 0) = 0$). Satisfying the following four assumptions allow us to focus on two groups in the CACE model: compliers and non-compliers only.

These assumptions are (1) independence of intervention assignment and outcomes; (2) individual outcomes are not impacted by the condition assignment of other participants (i.e., the Stable Unit Treatment Value Assumption [SUTVA]) [75]; (3) monotonicity, which assumes there are no defiers and there are at least some compliers, and (4) the exclusion restriction, which assumes that the intervention had zero effects for the non-compliers across the intervention and control arm.

The first two assumptions are satisfied through cluster randomisation, reducing the likelihood of participant interaction (or contamination) across condition assignments. Monotonicity is guaranteed in the current study as schools randomised to the control arm of the trial do not have access to the programme materials [45], and providing the opportunity to participate ensures that some participants will actually do so [44]. Satisfying monotonicity allows clear classification of those who do not adhere to condition assignment in the intervention arm as never-takers, and the control arm as always-takers thus removing 'defiers'.

The exclusion restriction is difficult to guarantee in school-based trials, as non-compliance is rarely zero, and partial implementation may still yield benefits. This contrasts with medical trials, such as vaccine studies, where the exclusion restriction is more plausible since there is a clear binary distinction; participants in the intervention arm who refuse a vaccine (non-compliers) cannot experience immunological benefit and the assumption of zero effect for non-compliers is more credible. Methods to overcome potential violation of this assumption include relaxing of assumption parameters

using strong compliance predictors [76] and using sensitivity analyses when dichotomising a continuous measure of compliance to facilitate comparison of compliance thresholds [45]. With no defiers and no intervention effect for never- or always-takers, participants can be dichotomised into compliers vs non-compliers.

Structural equation mixture modelling (SEMM) provides a probabilistic framework to estimate the CACE effect. SEMM assumes that the sample is made up of unobservable latent subgroups (compliers, non-compliers). Compliance behaviour (captured via the self-report implementation survey) is observed in the intervention arm of an RCT and is used, alongside information from the outcome variables and compliance predictors, to estimate the probability of each participant in the control group belonging to a latent class; this framework identifies those in the control arm who would be compliers (had they been offered the intervention) [46]. In other words, the two latent classes include participants from both the intervention (based on observed compliance) and control (based on probabilistically estimated compliance) groups to allow for the comparison of intervention compliers vs. control “would be” compliers. Assuming exclusion restriction, the comparison between intervention non-compliers and control would be non-compliers is set to zero (no intervention effects assumed for those that did not comply).

Selecting a compliance definition and threshold. Construction of the multiverse data set is detailed in [S2 File](#) and visually illustrated in [Fig 1](#). Ten theoretically plausible specifications were identified and pre-registered (<https://osf.io/z62wa>). Skewed dosage data resulted in 9 specifications (see Compliance Thresholds) and thus 27 models were analysed. The first author was responsible for data analysis and accessed the data following publication of the pre-registration.

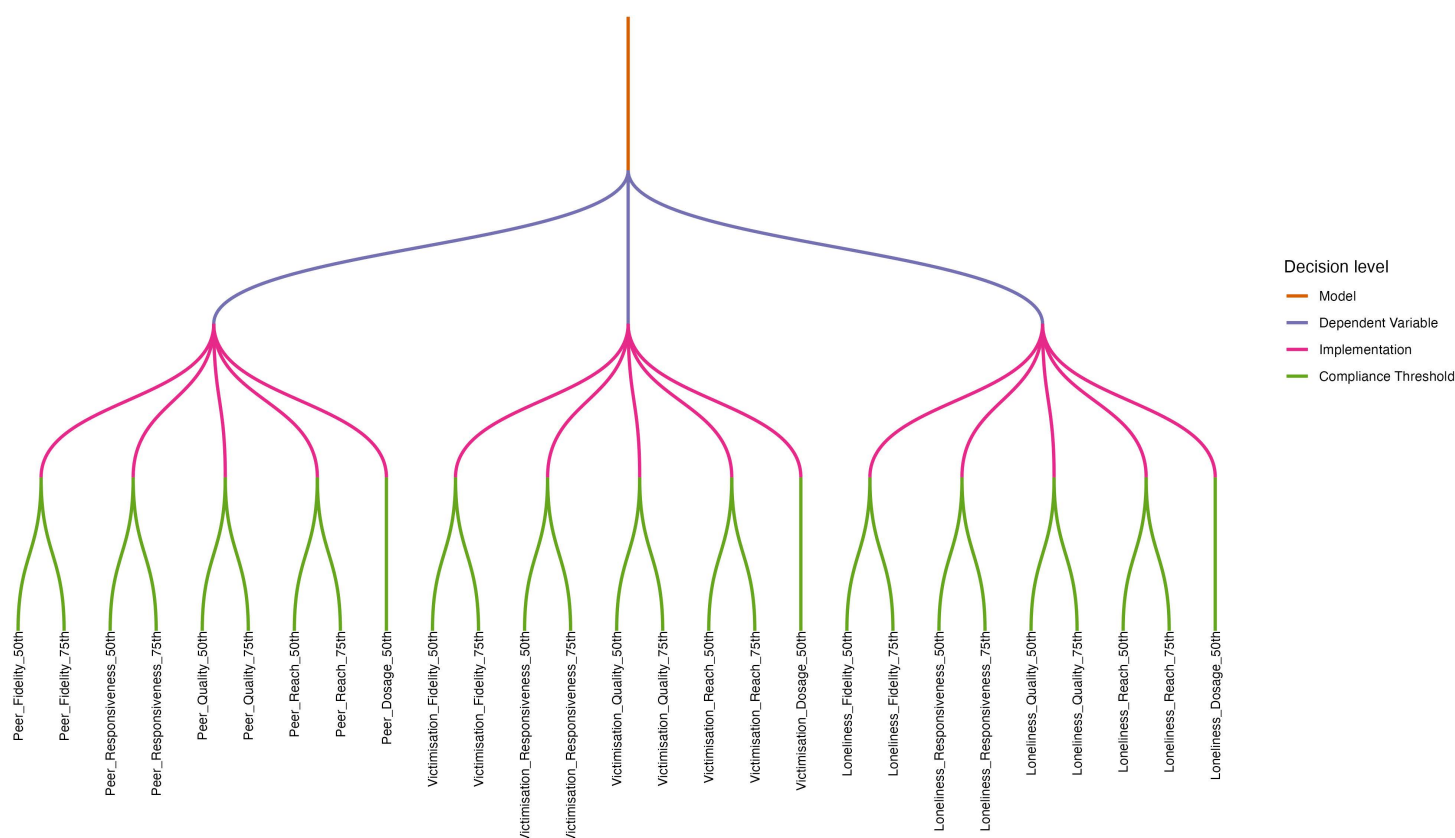


Fig 1. Decision tree for multiverse specifications.

<https://doi.org/10.1371/journal.pone.0349949.g001>

Compliance definition. Five theoretically plausible approaches to defining compliance were identified: fidelity, dosage, quality, responsiveness and reach. Each were introduced in separate CACE models.

Compliance thresholds. In line with recommendations [46] and established practice [41,49,57,58], two compliance thresholds were used to define compliers and non-compliers: the 50th and 75th percentiles. This decision was also pragmatic, as each CACE model requires substantial computational time (>24 hours) and it will facilitate comparison with other intervention studies using identical thresholds, such as the ones cited above. Classrooms that fell below the 50th percentile were classified as non-compliers, whereas classrooms above the 50th percentile were defined as moderate compliers and those above the 75th percentile as high compliers. Due to high levels of dosage, the 50th percentile only was used for compliance defined by dosage.

Statistical analysis. Statistical analyses were performed using Mplus Version 8 [77] and results were summarised via the Mplus Automation and ggplot2 in R Version 4.0.5 [78]. For code, see <https://osf.io/z62wa>. While the intervention was randomised at the school level, three-level mixture modeling is currently not possible. We therefore regressed y on x using two-level mixture hierarchical linear models (level 1 = child; level 2 = classroom) with random intercept fixed slopes and standard errors (SE) adjusted to account for school-level clustering (TYPE = COMPLEX). Classroom, and not school, was chosen as the level 2, as programme delivery occurred at the classroom-level. The compliance latent variable was estimated with MLR estimation and expectation-maximisation (ML-EM) algorithm [79].

Student-level covariates (baseline scores, sex, SEN) were modelled at the within-level and teacher-level covariates (stress, coping, attitudes) at the between-level. Full information maximum likelihood (FIML) was used to handle missing data. A correction for multiplicity was not applied in the current study as the multiverse is exploratory, in contrast with specification curve analysis for hypothesis testing [80]. The goal is not to correcting for multiplicity (i.e., alpha adjustment) but to increase transparency by visualising the distribution of outcomes [12].

Results

Descriptive statistics for study variables are reported in Table 1. Psychometric properties for all variables are reported in S3 File. Model fit was acceptable across all variables however the teacher sample size was small, and psychometric testing should be replicated with a larger sample.

ITT analyses

After controlling for student and classroom-level covariates, null ITT effects were found for peer support, victimisation and loneliness (see Table 2).

CACE analyses

The CACE effects for each analytic universe are reported in Table 3 (for individual model results, including entropy values, class count, and strength of predictors, see S4 File). The entropy value (E; range 0–1) indicates how well a model distinguishes between latent classes, with high values indicating stronger class separation between compliers and non-compliers [81]. Entropy values varied across models and only four met the pre-registered threshold of .76 [82]. The four

Table 2. Standardised ITT effect for relational outcomes (N=2,425).

Predictor	Outcome variables β (SE)		
	Peer support	Victimisation	Loneliness
Passport vs Control	-.02 (.16)	-.21 (.15)	.02 (.16)
	95% CI (-0.29, 0.25)	95% CI (-0.44, 0.03)	95% CI (-0.23, 0.28)

<https://doi.org/10.1371/journal.pone.0349949.t002>

Table 3. Standardised CACE effect for relational outcomes (N=2425).

Predictor	Compliance definition	Outcome variables β (SE) <i>Cohen's d</i>					
		Peer support		Victimisation		Loneliness	
Passport vs Usual Provision							
	Moderate fidelity	.32 (.3)	.13	.18 (.68)	.08	−.07 (.36)	−.02
	High fidelity	−.49 (1.25) ⁺	−0.26	.24 (.21)	.12	−.39 (.27)	−.14
	Moderate dosage	.22 (.24)	.09	.23 (5.6)	.08	−.15 (.21)	−.05
	Moderate responsiveness	−.33 (2.7) ⁺	−.14	.38 (.24)	.14	−.04 (.38)	−.02
	High responsiveness	−.38 (.24)	−.19	−.31 (.12)	−.18	−.08 (.23)	−.04
	Moderate quality	.31 (.29)	.13	.35 (.24)	.12	−.05 (.23)	−.02
	High quality	−.38 (.24)	−.23	.4 (.29)	.13	−.12 (3.7)	−.05
	Moderate reach	.4 (.31)	.17	.36 (.27)	.13	−.08 (.32)	−.02
	High reach	−.38 (.18)*	−.21	−.41 (.38)	−.3	.69 (.44)	.32

Note. These are the standardised model results. 50th percentile = moderate compliance; 75th percentile = high compliance. Positive values indicate improved scores for peer support and victimisation while negative values indicate poorer scores; the opposite is true of loneliness.

* $p < 0.05$.

+ Models without statistically significant compliance predictors.

<https://doi.org/10.1371/journal.pone.0349949.t003>

optimal-entropy models examined intervention effects on victimisation with compliance modelled using moderate reach, engagement, adherence, and dosage. Compliance predictors were weak across multiple CACE models (see [S4 File](#)).

The multiverse analyses revealed a statistically significant and *negative* intervention effect for peer support when compliance was defined by high reach ($\beta = -.38$, 95% CI $(-.68, -.08)$, $E = .71$, $d = -.21$). The remaining 26 CACE models were not statistically significant, indicating that the intervention did not improve student self-report victimisation, loneliness or peer support relative to the control, irrespective of implementation variability. Additionally, several models revealed substantial effect sizes with corresponding unstandardised models considered statistically significant, explained by differences in the sampling distribution between standardised and non-standardised coefficients. Our primary interpretation and reporting are based on standardised results as per our pre-registration, and standardised estimates are reported in [S4 File](#). The results of the multiverse are illustrated by [Figs 2–4](#) for victimisation, loneliness and peer support, respectively.

Victimisation. The strongest entropy values were observed across models for victimisation (.7 – .78). CACE effects ranged between -0.41 and 0.4 , with seven of nine estimates above zero and all with confidence intervals including zero. Results for moderate compliers were relatively consistent across implementation dimensions, with estimates clustering between 0.18 and 0.38 . For high compliers, results showed greater variability depending on how compliance was defined; students in high compliance classrooms were more likely to report poorer scores when compliance was defined by reach ($\beta = -.41$, 95% CI $(-1.03, .22)$, $E = .73$) or engagement ($\beta = -.31$, 95% CI $(-.63, .01)$, $E = .7$), while high quality resulted in the largest positive effect ($\beta = .4$, 95% CI $(-.06, .86)$, $E = .76$). Confidence intervals are wide for dosage compared to other implementation dimensions ($\beta = .23$, 95% CI $(-8.91, 9.37)$, $E = .78$), indicating greater uncertainty in this estimate.

Loneliness. Entropy values were lowest for loneliness models (.59 – .69). Results for moderate and compliers were relatively consistent across all but one model, with estimates ranging between $-.39$ and $-.08$, indicating small but non-statistically significant improvements in scores. Students in classrooms with high levels of reach however were more likely to report higher levels of loneliness ($\beta = .69$, 95% CI $(-.04, 1.42)$, $E = .69$, $d = .32$). Confidence intervals are wide when compliance is defined by quality using the 75th percentile ($\beta = -.11$, 95% CI $(-6.21, 6)$, $E = .65$) compared to other implementation dimensions.

Peer support. The greatest variation in results is observed for peer support, with one model considered statistically significant when compliance was defined using high reach ($\beta = -.38$, 95% CI $(-.68, -.08)$, $E = .71$, $d = -.21$). Across

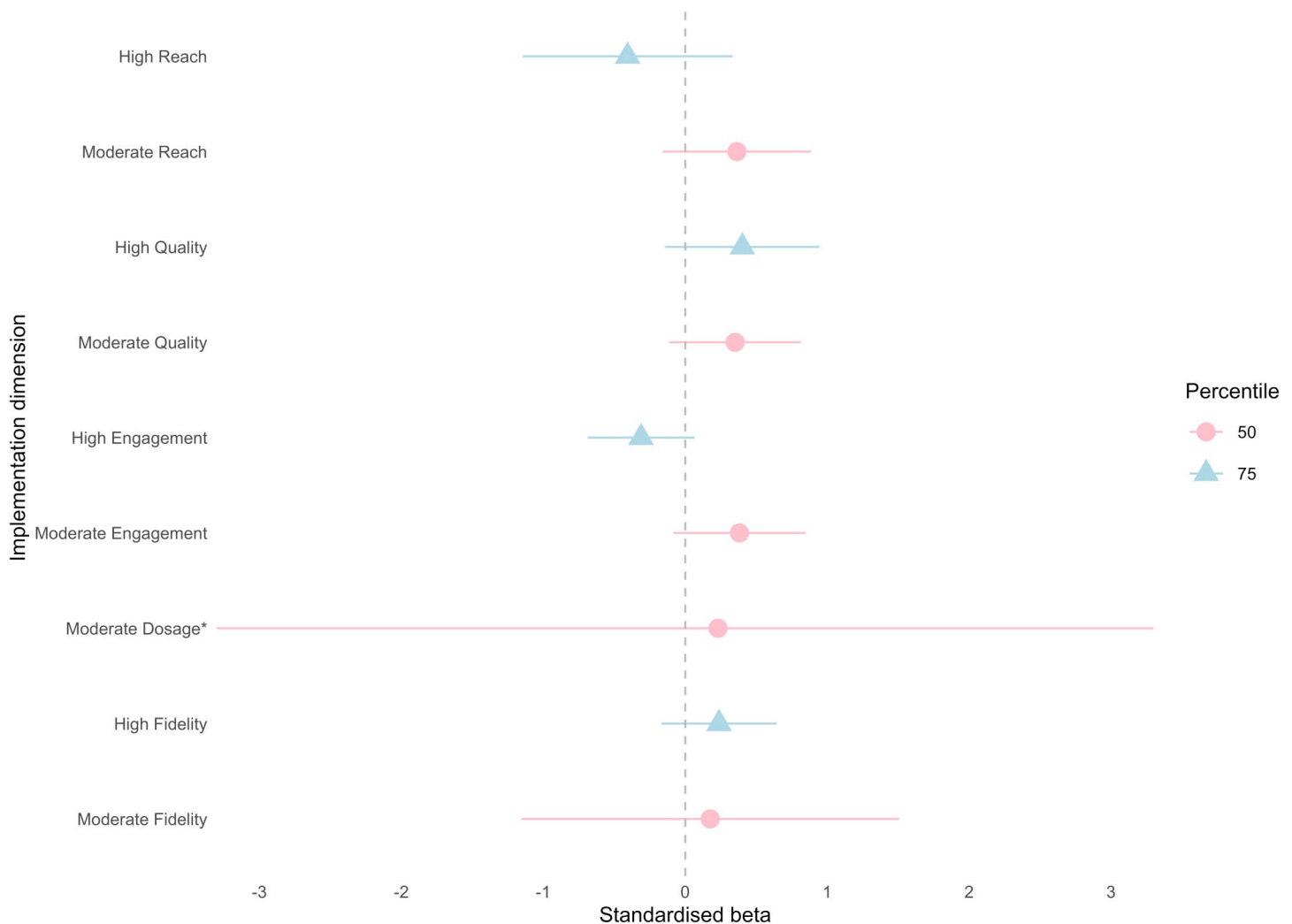


Fig 2. CACE estimates (x-axis) across five compliance definitions (y-axis) for victimisation outcome. Compliance definition is along the y-axis and point estimates are plotted as standardised beta coefficients on the x-axis, with pink circles representing the 50th percentile (moderate compliance) and blue triangles representing the 75th percentile (high compliance) across model specifications. Horizontal lines indicate 95% confidence intervals. An asterisk (*) indicates a confidence interval that extends beyond the x-axis limits shown and is truncated for visualisation.

<https://doi.org/10.1371/journal.pone.0349949.g002>

models, estimates varied considerably based on compliance levels and dimensions. For high compliers, estimates cluster between -0.49 and -0.38 while moderate compliers showed greater range of estimates (-0.33 and 0.4); when the estimate for moderate engagement (the only negative effect; $\beta = -0.33$) was removed, all estimates were above zero. Within compliance levels (moderate, high), estimates were relatively consistent except the outlier of moderate engagement, however meaningful differences are observed between compliance levels. Negative estimates were observed for high compliers while moderate compliers showed positive estimates, indicating the direction of intervention effects vary depending on the compliance threshold applied. When compliance was modelled by engagement, the difference between high ($\beta = -0.38$, 95% CI $(-0.77, 0.01)$, $E = 0.74$) and moderate compliers ($\beta = -0.33$, 95% CI $(-4.88, 4.23)$, $E = 0.6$) became negligible, although the width of the confidence intervals increased. Similarly, wide intervals are observed for high adherence ($\beta = -0.49$, 95%

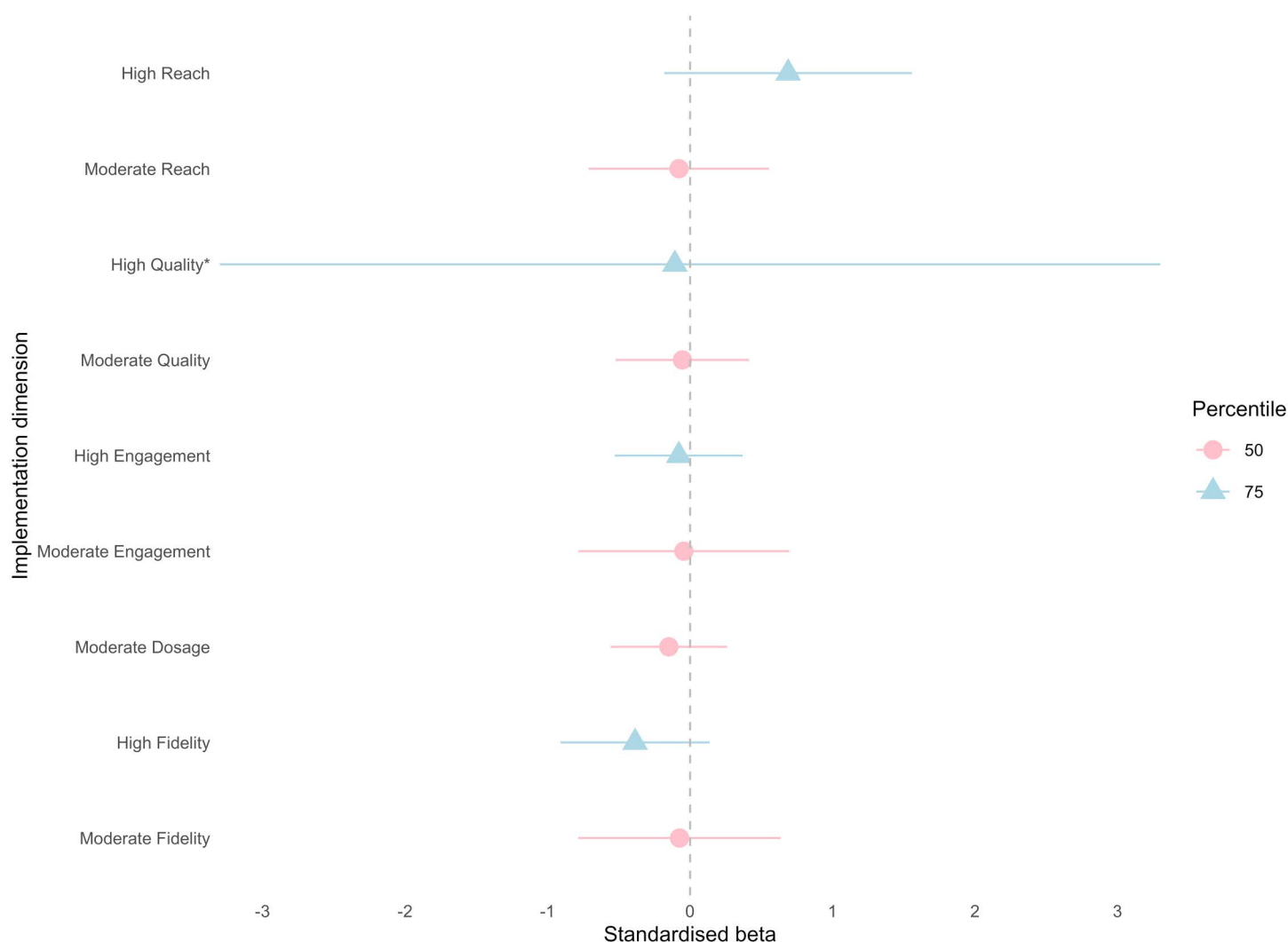


Fig 3. CACE estimates (x-axis) across five compliance definitions (y-axis) for loneliness outcome. Compliance definition is along the y-axis and point estimates are plotted as standardised beta coefficients on the x-axis, with pink circles representing the 50th percentile (moderate compliance) and blue triangles representing the 75th percentile (high compliance) across model specifications. Horizontal lines indicate 95% confidence intervals. An asterisk (*) indicates a confidence interval that extends beyond the x-axis limits shown and is truncated for visualisation.

<https://doi.org/10.1371/journal.pone.0349949.g003>

CI (-2.54, 1.56), $E = .67$). Despite the lack of statistical significance, models showed non-trivial effect sizes for high adherence ($d = -.26$) and quality ($\beta = -.38$, 95% CI (-.77, .01), $E = .74$, $d = -.23$).

Discussion

Amidst calls to use instrumental variable approaches, such as CACE, to address implementation variability in school-based trials, arbitrary analytic decisions can influence results and subsequent conclusions [9]. The current study provides a first step in addressing these limitations using a multiverse framework. We investigated whether, and how, conclusions of an intervention's effectiveness were contingent on decisions about how compliance was defined and modelled. We recommend caution when interpreting the results of the multiverse, which does not identify superior analytic decisions

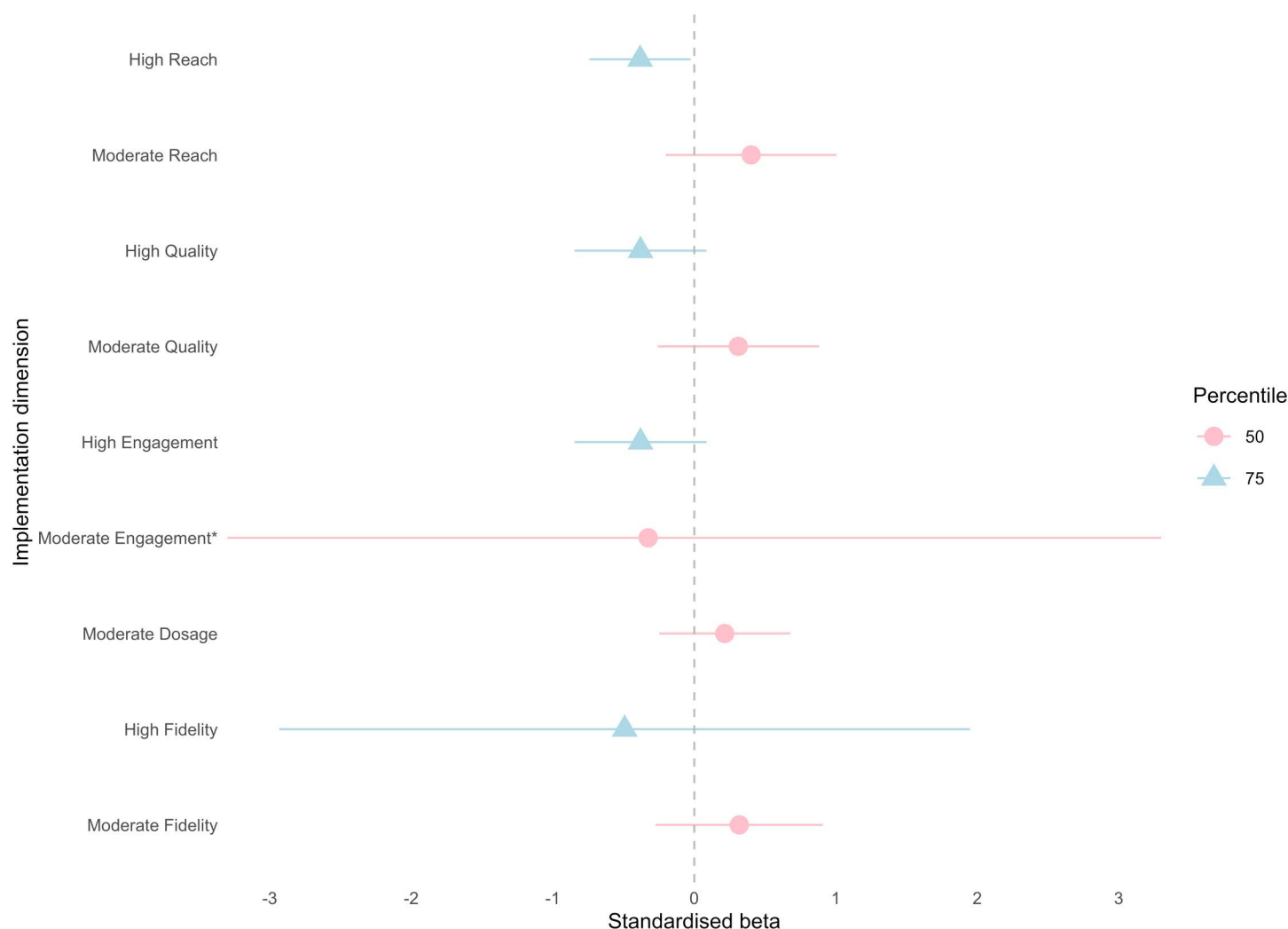


Fig 4. CACE estimates (x-axis) across five compliance definitions (y-axis) for peer social support outcome. Compliance definition is along the y-axis and point estimates are plotted as standardised beta coefficients on the x-axis, with pink circles representing the 50th percentile (moderate compliance) and blue triangles representing the 75th percentile (high compliance) across model specifications. Horizontal lines indicate 95% confidence intervals. An asterisk (*) indicates a confidence interval that extends beyond the x-axis limits shown and is truncated for visualisation.

<https://doi.org/10.1371/journal.pone.0349949.g004>

based on, for example, statistical significance, rather it provides a descriptive report of the robustness and replicability of results [12]. In other words, its purpose is not to discover which model conditions result in desirable outcomes (i.e., cherry picking), but transparently report whether results are robust to these variations.

The multiverse analysis revealed consistent results when examining the impact of intervention compliance on loneliness. However, results for victimisation were somewhat inconsistent across analytic universes depending on the compliance definition used. Results for peer support showed the greatest variability across universes, with effects dependent on both the moderator used and compliance threshold applied. Across the multiverse, variation manifested in entropy values, the direction and strength of effects, and CI precision. While CACE effects were sensitive to our choice of compliance definition and threshold, we can confidently conclude that the Passport intervention did not improve students' relational

outcomes relative to the control group, even when accounting for implementation variability. To explore the implications of this variability for intervention research, we examine each in turn.

Evaluating entropy values

Entropy values ranged between .59–.78 across models with most (85%) below the recommended threshold of .76 [82]. We initially hypothesised that suboptimal entropy values could be explained by negatively skewed dosage data resulting in low score variance across compliance definitions which were weighted by dosage. Transforming continuous implementation dimensions into a binary variable weighted by dosage may result in equifinality [83], whereby different combinations (e.g., low attendance and high dosage vs high attendance and low dosage) produce the same overall compliance score. This is in contrast to a ‘clean’ binary variable for dosage. However, the multiverse revealed variation across all compliance definitions and thresholds bar dosage, making this theory less viable. A clear pattern emerged when looking at the outcome variables, with the strongest entropy values observed for victimisation models and the weakest for loneliness. Although this hypothesis is speculative, this trend may be explained by the compliance predictors in the CACE models, which included the baseline score of the outcome under investigation. It is possible that bullying victimisation (and by proxy, perpetration) is a stronger predictor of teacher compliance as an externalising behaviour relative to feelings of loneliness which are internalising [84].

Evaluating precision and confidence intervals (CI)

The precision of point estimates varied across implementation dimensions and outcomes, as reflected in the differing widths of CI. Specifically, wide CI were observed when modelling dosage (victimisation), high quality (loneliness) and moderate engagement (peer support), indicating greater uncertainty around the estimated effects in these models. To a lesser extent, this was also observed for high adherence in relation to peer support.

These patterns may result from a number of potential factors. First, weak predictor variables may contribute to this imprecision. Indeed, some model results should be interpreted with a degree of caution, as the multiverse revealed differences in the strength and specificity of the compliance predictors across CACE models, which also likely impacted the specificity of class separation [85]. Strong compliance predictors are essential to enhancing the reliability and robustness of CACE models, as they directly influence class assignment precision and, consequently, the validity of estimated intervention effects [44]. Model results with weak compliance predictors potentially violate the exclusion restriction assumption and should therefore be interpreted with caution [86].

Despite theory and evidence informing covariate selection, this secondary data analysis was limited to available predictors. This issue of low predictive power of carefully chosen compliance predictors is not unique to the current study [87,88]. Predictors were consistent between models to limit inflating the number of models in the multiverse, nonetheless reviews suggest that contextual and individual factors may differentially influence implementation behaviours in school-based social-emotional and mental health interventions [89,90].

Furthermore, three models with low entropy values had comparatively larger CI: loneliness moderated by high quality ($E = .65$) and social support moderated by moderate engagement ($E = .6$) and high adherence ($E = .67$). However, wide CI for a model with strong entropy (victimisation moderated by dosage; $E = .78$) requires consideration of an alternative explanation. Ninety-two percent of all lessons were delivered on average resulting in negatively skewed dosage data and a small number of non-compliers (38%, $n = 921$), reducing power to detect moderation effects. Wide CI therefore reflect limited variability in the moderator, rather than outcome variability. Finally, implementation data may have been subject to measurement error, which is explored further below. Taken together, caution is warranted in interpreting these specific findings as the low precision limits confidence in the estimated effects, discussion of which we now turn.

Evaluating the direction and size of effects

The CACE effects were largely consistent with the ITT effect, indicating that there was no statistically significant difference between Passport and usual provision for student's relational outcomes for all but one model. Additionally, the multiverse revealed differences in the direction of point estimates between compliance thresholds. When compliance was modelled using the 50th percentile as the cut-off point between compliers and non-compliers, the direction of effects was mainly positive, but direction of effects was mainly negative for high compliance. An interesting trend was also observed for the size of effects: negligible effects were observed for loneliness across all models, except for high reach which had the largest effect size of all 27 models ($d = .32$). The second largest effect for high reach and victimisation, but effects were negative ($d = -.3$). Similar effects were observed for peer social support in classrooms with high adherence ($d = -.26$), high quality ($d = -.23$) and high reach ($d = -.21$, $p < .05$).

Opposing directional effects between high and moderate compliers indicates that there may be an implementation 'sweet spot' such that more is not always better [7]. For instance, high compliance to activity structure and timing (i.e., adherence) may require teachers to go against their professional judgement to adapt lesson delivery to their unique class context [91]. Teachers who prioritised adherence to activity timing over student-led discussion thought that Passport lessons were less effective [92]. Indeed, such a top-down prescriptive approach to teaching is criticised as a 'democratic deficit' in education [93]. Moderate compliers that adhere comparatively less to a top-down delivery model, and in turn remain responsive to student-led discussion, may more successfully foster an environment conducive to SEL [91].

The finding that high quality delivery is associated with negative effects for peer social support is harder to explain; Passport may inadvertently nurture emotional and social awareness without providing students with skills to seek support or to managing complex social situations. Findings from the process evaluation support this hypothesis, as some teachers described adapting programme content to their class (an indicator of quality) which resulted in inadvertently removing essential activities that award help-seeking (e.g., the Help Thermometer) and the identification of coping skills (the Dragon's Path); indeed, some students struggled to apply strategies in practice, particularly in the context of persistent bullying [94]. It is possible that greater exposure to lesson content, without sufficient structural support to practice skills, prevented development of targeted skills required to navigate social challenges.

This theory can be further interrogated in the context of an unexpected trend towards poorer scores for high attending classrooms (i.e., reach). The effectiveness of Passport may be compromised in large classroom settings if students have fewer opportunities to engage meaningfully in discussions or practice-based activities. Teacher interviews suggest that managing activities became increasingly challenging with over 30 students, particularly as Passport is student-led and centres around peer group discussions [92]. Passport was initially designed and piloted in schools with an average class size of 20 pupils [11] yet the programme is now delivered in English classrooms [67] which have on average 26 pupils [95], as reflected in the current trial's sample. This 30% increase in class size may disrupt the programme's core components and prevent students from deriving the intended benefits from guided practice elements (e.g., role play, games) that are essential to building confidence and competency to navigate such social challenges successfully.

The unexpected finding that high reach is paradoxically related to poorer outcomes ($d = .21 - .32$), compared to moderate reach, could be further explained by drawing upon the economic theory of opportunity cost, or the value of the next-best alternative that must be forgone when making a choice [96]. Delivering a universal programme like Passport requires schools to make resource allocation decisions within finite timetables and staffing constraints [97] – in other words, Passport must replace another lesson [98]. Teachers who report high attendance may choose to deliver a universal programme like Passport as an alternative to targeted support, which is more costly [99]; in contrast, moderate attendance may reflect a strategic approach, whereby an intervention like Passport is delivered when a portion of students who require targeted support leave the classroom to engage in those activities. Although critical to informing economic evaluations of school-based interventions [100], it's unclear how opportunity cost plays out in practice [101]. Further qualitative interviews were

conducted with school staff to understand how perceptions of an intervention's value and cost informs decisions to adopt and implement a universal intervention like Passport [92].

Finally, this trend towards poorer outcomes among high compliers may be explained by a complementary phenomenon of 'response shift' [102], whereby an intervention that aims to improve student outcomes (e.g., emotional literacy) improves measurement accuracy post-intervention. We recommend that future studies consider using differential item functioning analyses to examine whether control and intervention participants interpret measure items differently [103]. While previous research found no difference in teachers' interpretation of bullying behaviours following delivery of the PATHS programme [104], to the author's knowledge, the possibility of response shift has not yet been explored among children participating in school-based SEL interventions.

Considering measurement error

As per our trial protocol, implementation data was collected using a cross-sectional teacher-report survey to minimise burden and mitigate potential attrition compared to, for example, teacher observations [43]. This approach required teachers to provide a global rating across the 18-week intervention period, rather than session-specific assessments. This methodological constraint, though necessary, may have resulted in several limitations.

First, this approach limits identification and examination of within-programme fluctuation across implementation dimensions [7]. Second, the retrospective design introduces potential for recall bias, particularly recency effects, whereby scores may have been disproportionately influenced by the final 'Celebration' session (which was thoroughly enjoyed by staff and students) [94]. Third, self-report implementation is subject to social desirability bias, whereby responses are inflated to align with perceived expectations or professional standards [105,106]. Accordingly, the consistency of results across conceptually distinct implementation dimensions may result from reliance on single-reporter implementation data, inflating intercorrelations among dimensions.

The optimal approach to assessing implementation in school-based interventions is third-party observations [9,107] with an increasing number of tools available to researchers [108,109]. Observations offer value in terms of objectivity and rigour, but can be resource intensive and burdensome on school staff [7] and are not immune to bias (e.g., observer effect) [107]. Additionally, the inter-rater reliability between teacher and observer fidelity measures tends to be low [105,110]. It is therefore essential to consider approaches to optimising self-report measures, such as assessing student responsiveness using student-report surveys (rather than teacher-report) [9] and involving participants in survey design and development [111]. One such approach is ecological momentary assessment (EMA) to capture participant implementation behaviours (e.g., student responsiveness, teacher quality) during programme delivery. EMA capture variation in behaviours and attitudes across time and contexts, providing insight into whether and how responses vary between and across participants groups, contexts and programme sessions [112].

These suggestions align with calls to embed implementation considerations during programme design rather than an "added and separate effort" [113], and are important given evidence that researchers and practitioners may conceptualise implementation constructs differently [e.g., fidelity; 114]. This was observed in the current trial when teachers simultaneously reported high fidelity (i.e., adherence to lesson learning objectives, activity sequence and structure, and programme manual for core activities), while detailing lesson adaptations, modifications, and omissions [94]. This conceptual distinction could be explained by whether teachers are considering and reporting pedagogical adherence (*how* it is taught, e.g., role play) or practice-based adherence (*what* is taught, e.g., identifying emotions) [115]. Teachers draw upon their pedagogical knowledge to adapt an SEL intervention [116] and may therefore still report high fidelity if they believe they are adhering to the programme *ethos* despite pedagogical deviations [117]. The construct validity of implementation measures is rarely assessed or reported in school-based SEL research, raising concerns about the accuracy with which current tools capture the intended construct [9]. For example, the current measure of quality asked teachers if they felt able to engage

pupils and well-prepared to teach lessons; one could argue that these items are indicators of programme design rather than quality of teacher delivery.

We urge researchers to carefully consider best practice when assessing implementation variability in school-based trials [106]. Ideally, implementation data is collected from multiple informants (e.g., student responsiveness self-reported by students) and at multiple timepoints. Moving forwards, researchers using multi-method multi-informant approaches (e.g., event-contingent EMA) would benefit from extending the multiverse framework to explore whether seemingly homogeneous results in a dataset are the result of measurement issues outlined above or a well-delivered intervention [118]. Using a study within a trial (SWAT) design can provide valuable insights for advancing school-based process evaluations [119], such as the validity of approaches to collecting implementation data (e.g., weekly, cross-sectional). Our Implementation Quality Appraisal Checklist provides tiered guidance to improve research practice, which can be useful when managing resource constraints [9].

Moving from individual parts to a whole picture

The multiverse reveals variability in results depending on how we define and model intervention compliance, which manifested in entropy values, precision of confidence intervals and the strength and direction of intervention effects. Despite this variability, clear patterns emerged across the multiverse: certain outcomes (e.g., victimisation) produced stronger entropy values than others (e.g., loneliness) and moderate compliance was more often associated with positive outcomes compared to high compliance.

When considering the results of the multiverse as a whole, it instils confidence in findings that were not variable across analytic universes: all but one of the 27 CACE models were not statistically significant, indicating that the CACE effects are largely consistent with the ITT results and Passport was no better than usual practice in improving children's relational outcomes, however it is delivered. This further supports the hypothesis that null intervention effects are not a result of implementation failure but can be explained by low programme differentiation [64], as approximately 60% of teachers reported implementing similar universal SEL programmes that overlapped substantially with Passport content. This finding is not unexpected given that schools are encouraged to prioritise SEL due to robust evidence of its efficacy [120].

Our findings reinforce the importance of considering implementation behaviours as distinct constructs and thus shaped by different determinants [6,89]. Future research would benefit from further empirical work examining student, teacher and classroom characteristics that predict implementation behaviours [121]. This would not only result in more robust modelling but have important practical implications; identifying individuals or contexts at greater risk of non-compliance can inform targeted allocation of resources to promote equitable implementation. For researchers utilising the current paper to guide implementation and process evaluation, we recommend using theory and evidence to prospectively identify appropriate predictors to enhance model robustness, for example, using qualitative and quantitative evidence of factors influencing implementation behaviours from systematic reviews. Additionally, involving teachers at participating schools via participatory co-design could inform the final selection of compliance predictors ensuring contextual relevancy. These recommendations have value beyond CACE; while its critics propose using a propensity score approach as a more robust alternative [52], this approach similarly relies on good covariates to predict compliance [45].

Furthermore, the findings from the multiverse challenge the assumption that more is always better when delivering a universal SEL programme like Passport. Considering that time is a major barrier to programme delivery in schools [90], our results support the continued calls to empirically establish intervention core components. This could be achieved using a factorial trial design [115,122] or bottom-up network approaches, whereby variation in intervention outcomes are mapped onto the associated active ingredients delivered that week [123]. This suggestion may draw scepticisms due to the subsequent burden on intervention participants, particularly amidst continued challenges with recruiting schools to trials [124,125]. Subsequently, secondary data analysis may be an appropriate solution [e.g., 126]. Such insights could

refine intervention models, removing redundant activities to produce a streamlined intervention (i.e., how much is good enough) [7,127].

Finally, we are encouraged to practice caution when null and/or negative intervention effects juxtapose participants' experiences of a programme's effectiveness [128]. The qualitative strand of the Passport trial found that students thoroughly enjoyed Passport and many teachers expressed an interest in continuing programme delivery for future year groups [94]. The programme manual provided teachers with clear structure and guidance for teaching SEL, and supported strong implementation, as indicated by the negatively skewed implementation data. This aligns with teacher preferences for clear, structured and time-efficient SEL curricula [129]. It is therefore possible that a manualised SEL programme like Passport, with minimal planning and preparation time, provides additional benefits for *teachers*, rather than students, when compared to the usual curriculum, given that educators' time is one of the most valuable resources available to schools [99]. However, our findings highlight a critical gap in our understanding of how schools navigate the opportunity cost, or trade-off, when deciding whether to adopt and deliver a universal intervention like Passport. These decision-making processes that guide resource allocation are poorly understood, particularly in the context of competing demands and resource constraints [97,98,129]. Further research will explore what drives school's programme preferences, and how this underpins resource allocation, which is essential to understanding and predicting programme adoption, implementation and sustainability patterns [92].

Conclusions

Evaluating implementation variability in school-based trials is not easy, but it is essential [8]. This paper demonstrates the use of multiverse analysis to systematically assess, and transparently report, how arbitrary decisions in our CACE model can influence results and subsequent conclusions. Importantly, it highlights how reliance on dosage to define compliance in education may obscure critical information about an intervention's effectiveness, particularly if an intervention has few sessions. By applying a multiverse framework to CACE, this paper employs a new approach to address an old problem; specifically, it provides an analytic framework for modelling implementation variability for school-based trials. In doing so, we move from a fragmented view of the implementation-outcomes relationship, towards a more coherent understanding of the complexity of examining complex interventions in real-world settings, seeing not just the parts, but the whole.

Supporting information

S1 File. Outcome surveys.

(DOCX)

S2 File. Multiverse decision making.

(DOCX)

S3 File. Psychometric properties.

(DOCX)

S4 File. Model results.

(XLSX)

Acknowledgments

We extend our sincerest gratitude to schools, teachers and students who participated in the Passport to Success trial.

Author contributions

Conceptualization: Annie O'Brien, Margarita Panayiotou.

Data curation: Joao Santos.

Formal analysis: Annie O'Brien.

Funding acquisition: Neil Humphrey.

Investigation: Annie O'Brien.

Methodology: Annie O'Brien, Margarita Panayiotou.

Supervision: Joao Santos, Neil Humphrey, Margarita Panayiotou.

Visualization: Annie O'Brien.

Writing – original draft: Annie O'Brien.

Writing – review & editing: Annie O'Brien, Joao Santos, Neil Humphrey, Margarita Panayiotou.

References

- McGorry PD, Mei C, Dalal N, Alvarez-Jimenez M, Blakemore S-J, Browne V, et al. The Lancet Psychiatry Commission on youth mental health. *Lancet Psychiatry*. 2024;11(9):731–74. [https://doi.org/10.1016/S2215-0366\(24\)00163-9](https://doi.org/10.1016/S2215-0366(24)00163-9) PMID: 39147461
- Cipriano C, Strambler MJ, Naples LH, Ha C, Kirk M, Wood M, et al. The state of evidence for social and emotional learning: a contemporary meta-analysis of universal school-based SEL interventions. *Child Dev*. 2023;94(5):1181–204. <https://doi.org/10.1111/cdev.13968> PMID: 37448158
- Durlak JA, Weissberg RP, Dymnicki AB, Taylor RD, Schellinger KB. The impact of enhancing students' social and emotional learning: a meta-analysis of school-based universal interventions. *Child Dev*. 2011;82(1):405–32. <https://doi.org/10.1111/j.1467-8624.2010.01564.x> PMID: 21291449
- Durlak JA, Mahoney JL, Boyle AE. What we know, and what we need to find out about universal, school-based social and emotional learning programs for children and adolescents: a review of meta-analyses and directions for future research. *Psychol Bull*. 2022;148(11–12):765–82. <https://doi.org/10.1037/bul0000383>
- Wigelsworth M, Lendrum A, Oldfield J, Scott A, ten Bokkel I, Tate K, et al. The impact of trial stage, developer involvement and international transferability on universal social and emotional learning programme outcomes: a meta-analysis. *Cambridge J Educ*. 2016;46(3):347–76. <https://doi.org/10.1080/0305764x.2016.1195791>
- Durlak JA, DuPre EP. Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. *Am J Community Psychol*. 2008;41(3–4):327–50. <https://doi.org/10.1007/s10464-008-9165-0> PMID: 18322790
- Durlak JA. Programme implementation in social and emotional learning: basic issues and research findings. *Cambridge J Educ*. 2016;46(3):333–45. <https://doi.org/10.1080/0305764x.2016.1142504>
- Durlak JA. Studying program implementation is not easy but it is essential. *Prev Sci*. 2015;16(8):1123–7. <https://doi.org/10.1007/s11121-015-0606-3> PMID: 26399607
- O'Brien A, Panayiotou M, Santos J, Hamilton S, Humphrey N. A systematic review informing recommendations for assessing implementation variability in universal, school-based social and emotional learning interventions. *Soc Emot Learn: Res Pract Policy*. 2025;5:100112. <https://doi.org/10.1016/j.sel.2025.100112>
- Rojas-Andrade R, Bahamondes LL. Is implementation fidelity important? A systematic review on school-based mental health programs. *Contemp School Psychol*. 2018;23(4):339–50. <https://doi.org/10.1007/s40688-018-0175-0>
- Mishara BL, Dufour S. Randomized control study of the implementation and effects of a new mental health promotion program to improve coping skills in 9 to 11 year old children: passport: skills for life. *Front Psychol*. 2020;11. <https://doi.org/10.3389/fpsyg.2020.573342>
- Steege S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing transparency through a multiverse analysis. *Perspect Psychol Sci*. 2016;11(5):702–12. <https://doi.org/10.1177/1745691616658637> PMID: 27694465
- Moore SE, Norman RE, Suetani S, Thomas HJ, Sly PD, Scott JG. Consequences of bullying victimization in childhood and adolescence: A systematic review and meta-analysis. *World J Psychiatry*. 2017;7(1):60–76. <https://doi.org/10.5498/wjp.v7.i1.60> PMID: 28401049
- Park C, Majeed A, Gill H, Tamura J, Ho RC, Mansur RB, et al. The effect of loneliness on distinct health outcomes: a comprehensive review and meta-analysis. *Psychiatry Res*. 2020;294:113514. <https://doi.org/10.1016/j.psychres.2020.113514> PMID: 33130511
- Dunn C, Sicouri G. The relationship between loneliness and depressive symptoms in children and adolescents: a meta-analysis. *Behav change*. 2022;39(3):134–45. <https://doi.org/10.1017/bec.2022.13>
- Perlman D, Peplau LA. Toward a social psychology of loneliness. In: Gilmour R, editor. *Personal relationships in disorders*. London, UK: Academic Press; 1981. pp. 31–56.
- Jefferson R, Barreto M, Verity L, Qualter P. Loneliness during the school years: how it affects learning and how schools can help. *J Sch Health*. 2023;93(5):428–35. <https://doi.org/10.1111/josh.13306> PMID: 36861756

18. Matthews T, Qualter P, Bryan BT, Caspi A, Danese A, Moffitt TE, et al. The developmental course of loneliness in adolescence: Implications for mental health, educational attainment, and psychosocial functioning. *Dev Psychopathol*. 2023;35(2):537–46. <https://doi.org/10.1017/S0954579421001632> PMID: 35109947
19. Department for Culture, Media and Sport. A connected society: a strategy for tackling loneliness - laying the foundations for change. London, UK: HM Government; 2018.
20. Department for Education. Preventing and tackling bullying: Advice for headteachers, staff and governing bodies. London, UK: HM Government; 2017.
21. Gladden RM, Vivolo-Kantor AM, Hamburger ME, Lumpkin CD. Bullying surveillance among youths: uniform definitions for public health and recommended data elements, version 1.0. Atlanta, GA: National Center for Injury Prevention and Control, Centers for Disease Control and Prevention and U.S. Department of Education; 2014.
22. Ye Z, Wu D, He X, Ma Q, Peng J, Mao G, et al. Meta-analysis of the relationship between bullying and depressive symptoms in children and adolescents. *BMC Psychiatry*. 2023;23(1):215. <https://doi.org/10.1186/s12888-023-04681-4> PMID: 36997959
23. Vrijen C, Wiertsema M, Ackermans MA, van der Ploeg R, Kretschmer T. Childhood and adolescent bullying perpetration and later substance use: a meta-analysis. *Pediatrics*. 2021;147(3):e2020034751. <https://doi.org/10.1542/peds.2020-034751> PMID: 33597287
24. Office for National Statistics. Bullying and online experiences among children in England and Wales: year ending March 2023. ONS; 2024.
25. Thornton E, Panayiotou M, Humphrey N. Prevalence, inequalities, and impact of bullying in adolescence: insights from the #BeeWell Study. *Int J Bullying Prev*. 2024. <https://doi.org/10.1007/s42380-024-00244-7>
26. Bauer A, Stevens M, Purtscheller D, Knapp M, Fonagy P, Evans-Lacko S, et al. Mobilising social support to improve mental health for children and adolescents: a systematic review using principles of realist synthesis. *PLoS One*. 2021;16(5):e0251750. <https://doi.org/10.1371/journal.pone.0251750> PMID: 34015021
27. Sifers SK. Social Support. In: Levesque RJR, editor. *Encyclopedia of Adolescence*. Cham: Springer International Publishing; 2018. pp. 3708–14.
28. Cohen S. Measures and concepts of social support. *Social support and health*. 1985. pp. 83–108.
29. Department of Health. Future in mind: promoting, protecting and improving our children and young people's mental health and wellbeing. London, UK: HM Government; 2015. https://assets.publishing.service.gov.uk/media/5a80b26bed915d74e33fbc3c/Childrens_Mental_Health.pdf
30. Coleman N, Sykes W, Groom C. Peer support and children and young people's mental health. London, UK: Department of Education; 2017.
31. Matthews T, Caspi A, Danese A, Fisher HL, Moffitt TE, Arseneault L. A longitudinal twin study of victimization and loneliness from childhood to young adulthood. *Dev Psychopathol*. 2022;34(1):367–77. <https://doi.org/10.1017/S0954579420001005> PMID: 33046153
32. Qualter P, Vanhalst J, Harris R, Van Roekel E, Lodder G, Bangee M, et al. Loneliness Across the Life Span. *Perspect Psychol Sci*. 2015;10(2):250–64. <https://doi.org/10.1177/1745691615568999>
33. Sha S, Loveys K, Qualter P, Shi H, Krpan D, Galizzi M. Efficacy of relational agents for loneliness across age groups: a systematic review and meta-analysis. *BMC Public Health*. 2024;24(1):1802. <https://doi.org/10.1186/s12889-024-19153-x> PMID: 38971769
34. Boulton MJ, Trueman M, Chau C, Whitehand C, Amatya K. Concurrent and longitudinal links between friendship and peer victimization: implications for befriending interventions. *J Adolesc*. 1999;22(4):461–6. <https://doi.org/10.1006/jado.1999.0240> PMID: 10469510
35. Kendrick K, Jutengren G, Stattin H. The protective role of supportive friends against bullying perpetration and victimization. *J Adolesc*. 2012;35(4):1069–80. <https://doi.org/10.1016/j.adolescence.2012.02.014> PMID: 22464910
36. Kochel KP, Bagwell CL, Ladd GW, Rudolph KD. Do positive peer relations mitigate transactions between depressive symptoms and peer victimization in adolescence? *J Appl Dev Psychol*. 2017;51:44–54. <https://doi.org/10.1016/j.appdev.2017.04.003> PMID: 29104337
37. Yang K, Petersen KJ, Qualter P. Undesirable social relations as risk factors for loneliness among 14-year-olds in the UK: Findings from the Millennium Cohort Study. *Int J Behav Dev*. 2020;46(1):3–9. <https://doi.org/10.1177/0165025420965737>
38. Humphrey N, Hennessey A, Lendrum A, Wigelsworth M, Turner A, Panayiotou M. The PATHS curriculum for promoting social and emotional well-being among children aged 7–9 years: a cluster RCT. 2018. <https://doi.org/10.3310/phr06100>
39. Hennessey A, Qualter P, Humphrey N. The impact of promoting alternative thinking strategies (PATHS) on loneliness in primary school children: Results from a randomized controlled trial in England. *Frontiers in Education*. Frontiers Media SA; 2021.
40. Kusché CA, Greenberg MT, Anderson LA. The PATHS curriculum: Promoting alternative thinking strategies. Seattle, WA: Developmental Research & Programs; 1994.
41. Panayiotou M, Humphrey N, Hennessey A. Implementation matters: Using complier average causal effect estimation to determine the impact of the Promoting Alternative Thinking Strategies (PATHS) curriculum on children's quality of life. *Journal of Educational Psychology*. 2020;112(2):236–53. <https://doi.org/10.1037/edu0000360>
42. Eccles AM, Qualter P. Review: alleviating loneliness in young people - a meta-analysis of interventions. *Child Adolesc Ment Health*. 2021;26(1):17–33. <https://doi.org/10.1111/camh.12389> PMID: 32406165
43. O'Brien A, Hamilton S, Humphrey N, Qualter P, Boehnke JR, Santos J, et al. Examining the impact of a universal social and emotional learning intervention (Passport) on internalising symptoms and other outcomes among children, compared to the usual school curriculum: study protocol for a school-based cluster randomised trial. *Trials*. 2023;24(1):703. <https://doi.org/10.1186/s13063-023-07688-0> PMID: 37915094

44. Stuart EA, Perry DF, Le HN, Ialongo NS. Estimating intervention effects of prevention programs: accounting for noncompliance. *Prev Sci*. 2008;9(4):288–98. <https://doi.org/10.1007/s11121-008-0104-y>
45. Sagarin BJ, West SG, Ratnikov A, Homan WK, Ritchie TD, Hansen EJ. Treatment noncompliance in randomized experiments: statistical approaches and design issues. *Psychol Methods*. 2014;19(3):317–33. <https://doi.org/10.1037/met0000013> PMID: [24773358](#)
46. Peugh JL, Strotman D, McGrady M, Rausch J, Kashikar-Zuck S. Beyond intent to treat (ITT): a complier average causal effect (CACE) estimation primer. *J Sch Psychol*. 2017;60:7–24. <https://doi.org/10.1016/j.jsp.2015.12.006> PMID: [28164801](#)
47. Axford N, Berry V, Lloyd J, Hobbs T, Wyatt K. Promoting learning from null or negative results in prevention science trials. *Prev Sci*. 2022;23(5):751–63. <https://doi.org/10.1007/s11121-020-01140-4> PMID: [32748164](#)
48. Imbens GW, Rubin DB. Estimating outcome distributions for compliers in instrumental variables models. *Rev Econ Stud*. 1997;64(4):555–74. <https://doi.org/10.2307/2971731>
49. Bradshaw CP, Shukla KD, Pas ET, Berg JK, Ialongo NS. Using complier average causal effect estimation to examine student outcomes of the PAX good behavior game when integrated with the PATHS curriculum. *Adm Policy Ment Health*. 2020;47(6):972–86. <https://doi.org/10.1007/s10488-020-01034-1> PMID: [32297095](#)
50. Humphrey N, Panayiotou M, Hennessey A, Ashworth E. Treatment effect modifiers in a randomized trial of the good behavior game during middle childhood. *J Consult Clin Psychol*. 2021;89(8):668–81. <https://doi.org/10.1037/ccp0000673> PMID: [34472894](#)
51. Rowbotham S, Conte K, Hawe P. Variation in the operationalisation of dose in implementation of health promotion interventions: insights and recommendations from a scoping review. *Implement Sci*. 2019;14(1):56. <https://doi.org/10.1186/s13012-019-0899-x> PMID: [31171008](#)
52. Gómez JA, Brown JL, Downer JT. High quality implementation of 4Rs + MTP increases classroom emotional support and reduces absenteeism. *Front Psychol*. 2023;14:1065749. <https://doi.org/10.3389/fpsyg.2023.1065749> PMID: [37179887](#)
53. Humphrey N, Barlow A, Wigelsworth M, Lendrum A, Pert K, Joyce C. Promoting alternative thinking strategies (paths): Evaluation report and executive summary. London, UK: Education Endowment Foundation; 2015.
54. Conduct Problems Prevention Research Group. Initial impact of the fast track prevention trial for conduct problems: II. Classroom effects. *J Consult Clin Psychol*. 1999;67(5):648–57. <https://doi.org/10.1037/0022-006x.67.5.648>
55. Education Endowment Foundation. Statistical analysis guidance for EEF evaluations. London, UK: Education Endowment Foundation; 2018.
56. Durlak JA. The importance of doing well in whatever you do: A commentary on the special section, “Implementation research in early childhood education”. *Early Childhood Res Quart*. 2010;25(3):348–57. <https://doi.org/10.1016/j.ecresq.2010.03.003>
57. Berg JK, Bradshaw CP, Jo B, Ialongo NS. Using complier average causal effect estimation to determine the impacts of the good behavior game preventive intervention on teacher implementers. *Admin Policy Mental Health Mental Health Serv Res*. 2017;44(4):558–71. <https://doi.org/10.1007/s10488-016-0738-1>
58. Ashworth E, Panayiotou M, Humphrey N, Hennessey A. Game on—complier average causal effect estimation reveals sleeper effects on academic attainment in a randomized trial of the good behavior game. *Prev Sci*. 2020;21(2):222–33. <https://doi.org/10.1007/s11121-019-01074-6>
59. Crenshaw AO, Pukay-Martin N, Wagner AC, Schobitz RP, Yarvis JS, Young-McCaughan S. A multiverse approach to analyzing data from clinical trials. 2022. <https://doi.org/10.31219/osf.io/p6c5v>
60. Crenshaw AO, Pukay-Martin ND, Wagner AC, Schobitz RP, Yarvis JS, Young-McCaughan S, et al. Navigating Analytical Challenges in Clinical Trials Using the Multiverse Approach. *Collabra: Psychology*. 2026;12(1):1–12. <https://doi.org/10.1525/collabra.155619>
61. Böschen I. Changes in methodological study characteristics in psychology between 2010–2021. *PLoS One*. 2023;18(5):e0283353. <https://doi.org/10.1371/journal.pone.0283353> PMID: [37163505](#)
62. Del Giudice M, Gangestad SW. A traveler’s guide to the multiverse: promises, pitfalls, and a framework for the evaluation of analytic decisions. *Adv Methods Pract Psychol Sci*. 2021;4(1). <https://doi.org/10.1177/2515245920954925>
63. Department for Education. Schools, pupils and their characteristics: Academic year 2022/23. HM Government. 2023 [cited 2023 Jun 8]. Available from: <https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics/2022-23>
64. Humphrey N, Boehnke JR, Santos J, Alemdar M, Panayiotou M, O’Brien A. The effect of a universal, school-based social and emotional learning intervention (Passport: skills for life) on internalising symptoms and related outcomes during the transition from childhood to adolescence: A cluster-randomised controlled trial. *J Educ Psychol*. 2025;117(7):1095–114. <https://doi.org/10.1037/edu0000963>
65. Saghaei M, Saghaei S. Implementation of an open-source customizable minimization program for allocation of patients to parallel groups in clinical trials. *J Biomed Sci Eng*. 2011;04(11):734–9. <https://doi.org/10.4236/jbise.2011.411090>
66. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ*. 2014;348:g1687. <https://doi.org/10.1136/bmj.g1687> PMID: [24609605](#)
67. Partnership for Children. Passport: Teaching Coping and Social Skills for ages 9–11. Surrey: Partnership for Children. 2019. Available from: <https://www.partnershipforchildren.org.uk/what-we-do/programmes-for-schools/passport.html>
68. Office for National Statistics. Measuring loneliness: guidance for use of the national indicators on surveys Internet. 2018. Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/methodologies/measuringlonelinessguidanceforuseofthenationalindicatorsonsurveys>
69. Rotenberg KJ, MacDonald KJ, King EV. The relationship between loneliness and interpersonal trust during middle childhood. *J Genet Psychol*. 2004;165(3):233–49. <https://doi.org/10.3200/GNTP.165.3.233-249> PMID: [15382815](#)

70. Rotenberg KJ, Boulton MJ, Fox CL. Cross-sectional and longitudinal relations among children's trust beliefs, psychological maladjustment and social relationships: are very high as well as very low trusting children at risk? *J Abnorm Child Psychol*. 2005;33(5):595–610. <https://doi.org/10.1007/s10802-005-6740-9> PMID: 16195953
71. Ravens-Sieberer U, Gosch A, Rajmil L, Erhart M, Bruil J, Duer W, et al. KIDSCREEN-52 quality-of-life measure for children and adolescents. *Expert Rev Pharmacoecon Outcomes Res*. 2005;5(3):353–64. <https://doi.org/10.1586/14737167.5.3.353> PMID: 19807604
72. Ravens-Sieberer U, Auquier P, Erhart M, Gosch A, Rajmil L, Bruil J, et al. The KIDSCREEN-27 quality of life measure for children and adolescents: psychometric results from a cross-cultural survey in 13 European countries. *Qual Life Res*. 2007;16(8):1347–56. <https://doi.org/10.1007/s11136-007-9240-2> PMID: 17668292
73. Robitail S, Ravens-Sieberer U, Simeoni M-C, Rajmil L, Bruil J, Power M, et al. Testing the structural and cross-cultural validity of the KIDSCREEN-27 quality of life questionnaire. *Qual Life Res*. 2007;16(8):1335–45. <https://doi.org/10.1007/s11136-007-9241-1> PMID: 17668291
74. Eddy CL, Herman KC, Reinke WM. Single-item teacher stress and coping measures: concurrent and predictive validity and sensitivity to change. *J Sch Psychol*. 2019;76:17–32. <https://doi.org/10.1016/j.jsp.2019.05.001> PMID: 31759465
75. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables: rejoinder. *J Am Stat Assoc*. 1996;91(434):468. <https://doi.org/10.2307/2291634>
76. Jo B. Estimation of intervention effects with noncompliance: alternative model specifications. *J Educ Behav Stat*. 2002;27(4):385–409. <https://doi.org/10.3102/10769986027004385>
77. Muthén LK, Muthén BO. *Mplus User's Guide*. 8 ed. Los Angeles, CA: Muthén & Muthén; 2017. pp. 1998–2017.
78. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2021.
79. Muthén B, Muthén L. *Mplus. Handbook of item response theory*. Chapman and Hall/CRC; 2017. pp. 507–18.
80. Simonsohn U, Simmons JP, Nelson LD. Specification curve analysis. *Nat Hum Behav*. 2020;4(11):1208–14. <https://doi.org/10.1038/s41562-020-0912-z> PMID: 32719546
81. Grimm KJ, Ram N, Estabrook R. *Growth modeling: Structural equation and multilevel modeling approaches*. Guilford Publications; 2016.
82. Wang M-C, Deng Q, Bi X, Ye H, Yang W. Performance of the entropy as an index of classification accuracy in latent profile analysis: a Monte Carlo simulation study. *Acta Psychologica Sinica*. 2017;49(11):1473. <https://doi.org/10.3724/sp.j.1041.2017.01473>
83. Hinnant B, Schulenberg J, Jager J. Multifinality, equifinality, and fanning: developmental concepts and statistical implications. *Int J Behav Dev*. 2021;45(5):429–39. <https://doi.org/10.1177/01650254211020402>
84. Humphrey N, Hennessey A, Ashworth E, Frearson K, Black L, Petersen K, et al. *Good Behaviour Game: Evaluation Report and Executive Summary*. London, UK: Education Endowment Foundation; 2018.
85. Jo B. Model misspecification sensitivity analysis in estimating causal effects of interventions with non-compliance. *Stat Med*. 2002;21(21):3161–81. <https://doi.org/10.1002/sim.1267> PMID: 12375297
86. Gerber AS, Green DP. *Field experiments: design, analysis, and interpretation*. W.W. Norton; 2012.
87. Knowles C, Thornton E, Panayiotou M, Santos J, Ashworth E, Mason C, et al. Session delivery completion as a modifier of treatment effects of universal mental health literacy curricula on emotional difficulties and help-seeking in primary and secondary schools: complier average causal effect estimation in the AWARE and INSPIRE cluster randomised controlled trials. <https://doi.org/10.17605/OSF.IO/KXUG7>
88. Thornton E, Knowles C, Panayiotou M, Santos J, Ashworth E, Mason C, et al. Implementation dosage as a modifier of treatment effects of universal mindfulness and relaxation interventions on emotional difficulties in primary and secondary schools: complier average causal effect estimation in the INSPIRE cluster randomised trial. <https://doi.org/10.17605/OSF.IO/KXUG7>
89. Mason C, Mansfield R, Demkowicz O, Humphrey N. Factors that influence teachers' implementation of school-based mental health interventions: systematic review. *School Mental Health*. 2025;18(1):57–75. <https://doi.org/10.1007/s12310-025-09830-6>
90. Ulla T, Poom-Valickis K. Program support matters: A systematic review on teacher-and school related contextual factors facilitating the implementation of social-emotional learning programs. *Frontiers in Education*. Frontiers Media SA; 2023.
91. Lendrum A, Humphrey N, Greenberg M. Implementing for success in school-based mental health promotion: The role of quality in resolving the tension between fidelity and adaptation. In: Shute RH, Slee PT, editors. *Mental health and wellbeing through schools: the way forward*. Routledge/Taylor & Francis Group; 2016. pp. 53–63.
92. O'Brien A, Demkowicz O, Humphrey N, Thompson A. Beyond budgets: how opportunity cost frameworks shape schools' decision to adopt and implement interventions. <https://doi.org/10.21203/rs.3.rs-8758562/v1>
93. Biesta G. Why “what works” won't work: evidence-based practice and the democratic deficit in educational research. *Educ Theory*. 2007;57(1):1–22. <https://doi.org/10.1111/j.1741-5446.2006.00241.x>
94. Demkowicz O, O'Brien A, Hamilton S, Burke L, Alemdar M, Mason C, et al. Child and school staff perceptions and experiences of universal social and emotional learning curricula in context: A qualitative case study registered report examining “Passport Skills for Life”. *Br J Educ Psychol*. 2026;10.1111/bjep.70065. <https://doi.org/10.1111/bjep.70065> PMID: 41772807
95. Department for Education. *Schools, pupils and their characteristics: academic year 2024/25* GOV.UK: HM Government. 2025. [cited Jun 5]. Available from: <https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics/2024-25>
96. Becker GS. A theory of the allocation of time. *Econ J*. 1965;75(299):493. <https://doi.org/10.2307/2228949>

97. Johnson R, Allard C, Soan C, Beach D, Al-Janabi H. "Care as capital": Developing theory about school investment in mental health and wellbeing. *Soc Sci Med*. 2025;366:117665. <https://doi.org/10.1016/j.socscimed.2024.117665> PMID: 39837080
98. Breheny K, Frew E, Williams I, Passmore S, Coast J. Use of economic evidence when prioritising public health interventions in schools: a qualitative study with school staff. *Int J Environ Res Public Health*. 2020;17(23):9077. <https://doi.org/10.3390/ijerph17239077> PMID: 33291788
99. Barrett CA, Spear SE, Clinkscales A, Wood LL, Maki KE. What interventions are cost-effective? A systematic review of cost-effectiveness analyses of school-based programs from 2000 to 2020. *Sch Psychol*. 2024;39(6):658–71. <https://doi.org/10.1037/spq0000590> PMID: 37902703
100. Levin HM, McEwan PJ, Belfield C, Bowden AB. Cost concepts. *Economic evaluation in education: Cost-effectiveness and benefit-cost analysis*. Third ed. Thousand Oaks: SAGE Publications, Inc.; 2018. pp. 45–60.
101. Andronis L, Maredza M, Petrou S. Measuring, valuing and including forgone childhood education and leisure time costs in economic evaluation: Methods, challenges and the way forward. *Soc Sci Med*. 2019;237:112475. <https://doi.org/10.1016/j.socscimed.2019.112475> PMID: 31408769
102. Oort FJ, Visser MRM, Sprangers MAG. Formal definitions of measurement bias and explanation bias clarify measurement and conceptual perspectives on response shift. *J Clin Epidemiol*. 2009;62(11):1126–37. <https://doi.org/10.1016/j.jclinepi.2009.03.013> PMID: 19540722
103. Baranowski T, Allen DD, Mâsse LC, Wilson M. Does participation in an intervention affect responses on self-report questionnaires? *Health Educ Res*. 2006;21 Suppl 1:i98–109. <https://doi.org/10.1093/her/cyl087> PMID: 17060350
104. Murray AL, Booth T, Eisner M, Ribeaud D, McKenzie K, Murray G. An analysis of response shifts in teacher reports associated with the use of a universal school-based intervention to reduce externalising behaviour. *Prev Sci*. 2019;20(8):1265–73. <https://doi.org/10.1007/s11121-019-00999-2> PMID: 30847752
105. Hansen WB, Pankratz MM, Bishop DC. Differences in observers' and teachers' fidelity assessments. *J Prim Prev*. 2014;35(5):297–308. <https://doi.org/10.1007/s10935-014-0351-6> PMID: 24903491
106. Humphrey N, Lendrum A, Ashworth E, Frearson K, Buck R, Kerr K. Implementation and process evaluation (IPE) for interventions in education settings: An introductory handbook. Education Endowment Foundation; 2016. pp. 1.
107. Humphrey N, Lendrum A, Ashworth E, Frearson K, Buck R, Kerr K. Implementation and process evaluation (IPE) for interventions in educational settings: A synthesis of the literature. London: Education Endowment Foundation; 2016.
108. Catalán Molina D, Porter T, Oberle C, Haghghat M, Fredericks A, Budd K, et al. How to measure quality of delivery: focus on teaching practices that help students to develop proximal outcomes. *J Res Educ Effect*. 2022;15(4):898–923. <https://doi.org/10.1080/19345747.2022.2054481>
109. Devlin BL, Paes TM, Geer EA, Bryant LM, Zehner TM, Korucu I, et al. Moving beyond dosage and adherence: a protocol for capturing dimensions of active child engagement as a measure of fidelity for social-emotional learning interventions. *Front Psychol*. 2023;13:1014713. <https://doi.org/10.3389/fpsyg.2022.1014713> PMID: 36698587
110. Tong F, Tang S, Irby BJ, Lara-Alecio R, Guerrero C, Lopez T. A process for establishing and maintaining inter-rater reliability for two observation instruments as a fidelity of implementation measure: a large-scale randomized controlled trial perspective. *Stud Educ Eval*. 2019;62:18–29. <https://doi.org/10.1016/j.stueduc.2019.04.008>
111. Spacciapoli M, Viana M, Saunders Wilder O, Sullivan J, McCallum T, Wilder-Smith B. An equitable and scalable approach to track fidelity of implementation in partnership with teachers. *Front Educ*. 2022;7. <https://doi.org/10.3389/educ.2022.1020204>
112. Shiffman S, Stone AA, Hufford MR. Ecological Momentary Assessment. *Ann Rev Clin Psychol*. 2008;4(1):1–32. <https://doi.org/10.1146/annurev.clinpsy.3.022806.091415>
113. Barnes SP, Domitrovich CE, Jones SM. Editorial: Implementation of social and emotional learning interventions in applied settings: approaches to definition, measurement, and analysis. *Front Psychol*. 2023;14:1281083. <https://doi.org/10.3389/fpsyg.2023.1281083> PMID: 37744606
114. Zetterlund J, von Thiele Schwarz U, Hasson H, Neher M. A slippery slope when using an evidence-based intervention out of context. how professionals perceive and navigate the Fidelity-Adaptation Dilemma—a qualitative study. *Front Health Serv*. 2022;2:883072. <https://doi.org/10.3389/frhs.2022.883072> PMID: 36925897
115. Wigelsworth M, Mason C, Verity L, Qualter P, Humphrey N. Making a case for core components: new frontiers in SEL theory, research, and practice. *School Psychol Rev*. 2021;53(5):1–14. <https://doi.org/10.1080/2372966X.2021.2004863>
116. Lendrum A, Askell-Williams H. Types of knowledge teachers use when solving educational problems: a case study of the implementation of the Promoting Alternative Thinking Strategies (PATHS) Program. *Problem solving for teaching and learning*. Routledge; 2019. pp. 140–56.
117. Lovett JM, Schonert-Reichl KA, Zinsser KM, Lawlor MS. Beyond fidelity: unveiling the landscape of teacher adaptation in social and emotional learning programs. *Front Educ*. 2024;9. <https://doi.org/10.3389/educ.2024.1444588>
118. Harder JA. The multiverse of methods: extending the multiverse analysis to address data-collection decisions. *Perspect Psychol Sci*. 2020;15(5):1158–77. <https://doi.org/10.1177/1745691620917678> PMID: 32598854
119. Ahmed S, Airlie J, Clegg A, Copsey B, Cundill B, Forster A, et al. A new opportunity for enhancing trial efficiency: Can we investigate intervention implementation processes within trials using SWAT (study within a trial) methodology? *Res Methods Med Health Sci*. 2022;3(3):66–73. <https://doi.org/10.1177/26320843221080734>
120. van Poortvliet M, Clarke A, Gross J. Improving social and emotional learning in primary schools: guidance report. London, UK: Education Endowment Foundation and Early Intervention Foundation; 2019.

121. Domitrovich CE, Li Y, Mathis ET, Greenberg MT. Individual and organizational factors associated with teacher self-reported implementation of the PATHS curriculum. *J Sch Psychol*. 2019;76:168–85. <https://doi.org/10.1016/j.jsp.2019.07.015> PMID: [31759464](#)
122. Wigelsworth M, Verity L, Mason C, Qualter P, Humphrey N. Social and emotional learning in primary schools: a review of the current state of evidence. *Br J Educ Psychol*. 2022;92(3):898–924. <https://doi.org/10.1111/bjep.12480> PMID: [34921555](#)
123. Blanken TF, Van Der Zweerde T, Van Straten A, Van Someren EJW, Borsboom D, Lancee J. Introducing network intervention analysis to investigate sequential, symptom-specific treatment effects: a demonstration in co-occurring insomnia and depression. *Psychother Psychosom*. 2019;88(1):52–4. <https://doi.org/10.1159/000495045> PMID: [30625483](#)
124. Axén I, Björk Brämberg E, Galaasen Bakken A, Kwak L. Recruiting in intervention studies: challenges and solutions. *BMJ Open*. 2021;11(1):e044702. <https://doi.org/10.1136/bmjopen-2020-044702> PMID: [33495262](#)
125. Girio-Herrera E, Ehrlich CJ, Danzi BA, La Greca AM. Lessons learned about barriers to implementing school-based interventions for adolescents: ideas for enhancing future research and clinical projects. *Cogn Behav Pract*. 2019;26(3):466–77. <https://doi.org/10.1016/j.cbpra.2018.11.004>
126. Laas Sigurðardóttir LB, Melendez-Torres GJ, Backhaus S, Gardner F, Scott S, European Parenting Program Research Consortium, et al. Individual participant data meta-analysis: individual differences in mediators of parenting program effects on disruptive behavior. *J Am Acad Child Adolesc Psychiatry*. 2025;64(5):564–76. <https://doi.org/10.1016/j.jaac.2024.10.003> PMID: [39395649](#)
127. Domitrovich CE, Bradshaw CP, Greenberg MT, Embry D, Poduska JM, Ialongo NS. Integrated models of school-based prevention: logic and theory. *Psychol Sch*. 2010;47(1):71–88. <https://doi.org/10.1002/pits.20452> PMID: [27182089](#)
128. Styles B, Torgerson C. Randomised controlled trials (RCTs) in education research –methodological debates, questions, challenges. *Educ Res*. 2018;60(3):255–64. <https://doi.org/10.1080/00131881.2018.1500194>
129. Wigelsworth M, Eccles A, Santos J. Social and emotional learning: a survey of English primary school's priorities, perceptions, and practices. *Int J Emot Educ*. 2021;13(2):23–39.