Effect Size Database of Interventions to Prevent Children Involvement in Violence for the Youth Endowment Fund: Final Report

Hikari Umezawa[1], Hannah Gaffney[2,3], Howard White[1,4], Hugh Sharma Waddington[5]

1 The Campbell Collaboration
2 Institute of Criminology, University of Cambridge
3 Clare Hall college, University of Cambridge
4 Global Development Network and Lanzhou University
5 London School of Hygiene and Tropical Medicine and London International Development Centre

Final report submitted to the Youth Endowment Fund

16 February 2024

Table of Contents

Acknowledgements

The effect size database was produced by coders working in four teams led from Ghana, the Philippines, Uganda and the UK. The contributing researchers, listed alphabetically, are: Adam Abdul-Rahaman, Sheila Agyemang Oppong, Robert Apunyo, Kishore Basak, Diwakar Basnet, Alyssa Cyrielle Villanueva, Joseph Darko, Nina Dela Cruz, Favour Ezeh, Carrel Fokou, Hannah Gaffney (HG), Neha Gupta, Sangyoung Jung, Desmond Kaledzi, Solomon Kamurari, Maria Kizza, Pastan Lusiba, Saranya Mohan Das, Regina Ndagire, Francis Nkunzimaana, Philip Orishaba, Hugh Sharma Waddington (HSW), Lovely Tolin, Rachel Wangi, Howard White (HW), Hikari Umezawa (HU). Double-coding was provided by HG, SJ, HSW and HU. Ashima Mohan managed the EGM database and assigned coders to studies. Training materials were developed by HG and HSW. The coder handbook was produced by HG with contributions from HSW, SJ and HU. The study protocol was produced by HSW, HG and HW. This report and the analyses conducted were produced by HU, HSW and HG.

Abbreviations

ASBO	anti-social behaviour order
CS	controlled studies
ES	effect size
EGM	Evidence and Gap Map
ESDB	Effect Size Database
IE	impact evaluation
OR	odds ratio
QED	quasi-experimental design
RCT	randomised controlled trial
SD	standard deviation
SDG	Sustainable Development Goal
SMD	standardised mean difference
UCS	uncontrolled studies
YEF	Youth Endowment Fund
YJB	Youth Justice Board

Summary

Programmes to reduce the involvement of children and young people in crime and violence are an important area of social policy in the United Kingdom (UK) and worldwide. This Effect Size Database (ESDB) is a free online repository that reports effect sizes (ES) associated with each study included in the Youth Endowment Fund (YEF) Evidence and Gap map (EGM). It enables users to determine the direction and magnitude of impact reported in a particular study for a specific outcome, intervention, population group and context, reported as the standardised mean difference. The ESDB will be made freely available by the YEF.

The database aims to help researchers efficiently design impact evaluations by presenting expected effect sizes associated with particular interventions and contexts. By providing a variety information associated with each ES, the database also aims to help researchers conduct future systematic reviews and meta-analyses.

The data were collected by a global team of 23 coders based in Bangladesh, Cameroon, Ghana, India, the Philippines, UAE, Uganda and the UK. Coders were required to attend training sessions, which we conducted in the UK (2 events), India, Ghana and Uganda. We adopted a rigorous quality control process, whereby a random sample of 10 percent of all studies, stratified by coder, were double-coded by a quality control team, with resolution meetings happening with each coder afterwards.

This report was prepared to inform YEF of the process taken to create the ESDB, as well as present findings from analyses of the distributions of ES for relevant outcomes, participant groups, interventions and study designs. The ESDB includes 14,834 ES on child-centred and offending and crime outcomes from 1,217 studies[1], of which more than 90 percent are from controlled studies (e.g., RCTs). Most of the ES (84%) were measured for child-centred outcomes (such as externalising behaviour, mental health and social cognition); 16 percent of the ES were measured for offending and crime outcomes (e.g., delinquency, contact with police, etc). The most frequently reported age group was 10-14 years. More ES were reported for males than females, and for majority ethnicity children rather than minority groups.

[1] The database includes 15,206 ES from 1,218 studies in total. Among them, this report focuses on 14,834 ES on child-centred and offending and crime outcomes, from 1,217 studies (one study containing one ES was excluded from the EGM after the start of this ESDB project; 368 ES are on other outcome domains; three ES are reported only as Cohen's *d*, not as Hedges' *g*).

The distribution analysis, focusing on average values of ES, suggested that in controlled studies the mean ES was higher for child-centred outcomes (mean g=0.207; interquartile range, IQR= -0.034, 0.368; 11,713 estimates) than for offending and crime outcomes (mean g=0.068; IQR= -0.100, 0.227; 2,286 estimates). The analysis also suggested that mean ES for targeted interventions (mean g=0.252; IQR= -0.049, 0.461; 5,760 estimates for child-centred outcomes, and mean g=0.070; IQR= -0.133, 0.269; 1,742 estimates for offending and crime outcomes) were larger ES than those for interventions with universal (non-targeted) access (mean g=0.1578; IQR= -0.025, 0.277; 5,830 estimates for child-centred outcomes, and mean g=0.065; IQR= -0.050, 0.1458; 517 estimates).

For both child-centred and offending and crime outcomes, ES were on average larger for female participants (mean g=0.300; IQR= 0.006, 0.424; 923 estimates for child-centred outcomes and mean g=0.202; IQR= 0.016, 0.213; 102 estimates for offending and crime outcomes) than for male participants (mean g=0.230; IQR= -0.034, 0.422; 1,259 estimates for child-centred outcomes and mean g=0.063; IQR= -0.076, 0.2380; 366 estimates for offending). ES also tended to be greater on average for males (mean g=0.294; IQR= -0.060, 0.575; 683 estimates for child-centred outcomes and mean g=0.064; IQR= -0.010, 0.302; 280 estimates for offending and crime outcomes) and females (mean g=0.551; IQR= 0.028, 0.785; 352 estimates for child-centred outcomes and mean g=0.876; IQR= 0.278, 0.723; 12 estimates for offending and crime outcomes) in same-sex targeted programmes, and for ethnic groups in ethnicity-targeted programmes (mean g=0.287; IQR= -0.005, 0.409; 1,092 estimates for child-centred outcomes and mean g=0.028; IQR= -0.119, 0.240; 240 estimates for offending and crime outcomes). For child-centred outcomes, the average ES was greatest among 0 to 3-year-old children (mean g=0.359; IQR= 0.029, 0.618; 847 estimates) and youths aged 15 years and older (mean g=0.280; IQR= -0.037, -0.409; 879 estimates), while offending and crime outcomes interventions tended to have larger ES among children aged 4-9 years (mean g=0.197; IQR= -0.070, 0.425; 212 estimates).

While we present these means and IQRs to illustrate the range of ES contained in the ESDB for particular groups of participants and programmes, they do not provide policy meaningful effects since they are not calculated using appropriate methods of statistical synthesis. We would expect any syntheses based on the ESDB to use inverse-variance weighted random effects multi-level meta-analysis models.

1. Background

As of 2021, about one-third of the UK population (19 million) is aged 24 or below.[2] Programmes to reduce children and young people's involvement in crime and violence are an important part of social policy in the United Kingdom (UK) and worldwide. The involvement of children and young people in crime and violence can also be considered a public health issue. Research evidence has shown that individuals who have been convicted are at increased risk for physical and mental illness (e.g., Jolliffe et al., 2019; Skinner & Farrington, 2021).

In the UK, involvement in crime and violence has a significant impact on children's behaviour, safety, and overall wellbeing. Recent youth justice data suggest that there is an overall downward trend in the involvement of children and young people in crime and violence, but there are still an estimated 8,000 children and young people having contact with the youth justice system (Youth Justice Board, YJB, 2023).

Children are also too often the victims of violence and crime. A survey of over 2,000 children identified that 14 percent of respondents reported being a victim of violence in the previous 12 months (YEF, 2022). This same survey found that children are fearful of violence and are modifying their behaviours in ways that could have long lasting impacts such as missing school or inability to concentrate at school.

At the international level, almost half (40 percent) of the world's population are children and youth. There are 2 billion children aged up to 14 years and 1.2 billion young people aged 15 to 24 years.[3] Interventions that can work to address the challenges facing children and youths are important for achieving United Nations Sustainable Development Goals (SDGs) 3 (health and wellbeing), 4 (education), 5 (gender equality), 8 (decent work and economic growth), 10 (reduced inequalities), 11 (sustainable cities and communities) and 16 (peace and justice)[4].

There exists a wealth of research on the different intervention programmes and prevention strategies that can be implemented with children and young people involved in crime and violence. For the past 30 years, the 'what works' approach has been dominant in research

[2] https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland
[3] https://www.un.org/en/global-issues/youth#:~:text=Today%2C%20there%20are%201.2%20billion,cent%2C%20to%20nearly%201.3%20billion
[4] https://sdgs.un.org/goals

related to children and young people's involvement in and violence (for a good example, see Lipsey & Wilson, 1998). Several organisations in the UK, and further afield, favour the 'what works' approach to understanding the best possible ways in which policymakers can prevent and reduce crime. For example, the College of Policing have a What Works Centre for Crime Reduction and a crime reduction toolkit to summarise the best available evidence on various different forms of crime reduction and prevention[5]. In relation to children and young people, the Youth Endowment Fund have also created a Toolkit that focuses on providing an accessible 'best bets' menu of intervention and prevention programmes[6] (Farrington et al., 2022).

A wide range of interventions could be implemented to reduce, or prevent, children and young people becoming involved in crime and violence. Drawing from research in public health, scholars have suggested that these types of interventions can be classified as primary, secondary, and tertiary forms of prevention (e.g., Brantingham & Faust, 1976; Sutton et al., 2021). In this classification, primary prevention programmes are those that are implemented before offending has occurred and/or to prevent the development of risk factors for offending, whereas secondary prevention programmes are implemented with children identified at-risk of offending, normally through the presence of several risk factors or initial contact with the criminal justice system. Finally, tertiary intervention programmes are implemented with individuals who have already offended and engaged in criminal behaviours (Sutton et al., 2022). Another classification model categorises interventions as either developmental approaches, to prevent the development of risk factors associated with crime and violence, community prevention to change neighbourhood and social factors that may relate to criminal behaviour, and situational prevention approaches that modify the physical environment to reduce the opportunities for crime to occur and increase the risk associated with offending (Tonry & Farrington, 1995).

Interventions for children and young people sometimes have large, and desirable, effects on target populations. For example, meta-analyses have shown that child skills training can reduce delinquent behaviours by approximately 24-32 percent relative to a comparison group that received no intervention (Beelman & Lösel, 2021; Gaffney et al., 2021a). In addition, reviews have shown that strategies such as diverting children and young people from traditional or formal criminal justice system processing can have desirable impacts on the incidence of reoffending. One review demonstrated that this approach is associated with an approximate relative reduction of 20 percent (Petrosino et al., 2019; Gaffney et al., 2021b).

However, interventions do not always have a desirable impact, and may even have unintended or undesired harmful effects. For example, 'bootcamp' interventions where children and young people are exposed to military-style living conditions, exercise regimes and punitive discipline have been associated with an increase in offending behaviours. Results reported by Wilson et al. (2008) on the impact of bootcamps are associated with a relative increase in reoffending of approximately 3-5 percent (Gaffney et al., 2021c). Although a small increase, harmful effects

[5] https://www.college.police.uk/research/what-works-centre-crime-reduction
[6] https://youthendowmentfund.org.uk/toolkit/

of crime prevention programmes are always undesirable, and even small harms can be policy meaningful if they are experienced by a large number of people, including those on the receiving end of crime.

Where interventions have no effects on desired outcomes, this poses the challenge as to whether public and/or private resources should be further channelled to seemingly ineffective strategies; at the very least, further analysis is needed to establish why there are no impacts on preventing youth involvement in crime and violence and to (re-) designing and testing more effective programmes. A role of rigorous evaluation is to help policymakers identify which interventions are likely to be more effective, that is, have desirable impacts, measured as the effect size (ES), within reasonable resource outlays.

High quality research on the effectiveness of interventions for children and youth has been systematically collected by the Campbell Collaboration and disseminated in YEF's Evidence and Gap Map (EGM). The EGM was produced by staff associated with Campbell including the Campbell Secretariat, but it has not been through the Campbell Collaboration peer review process. The YEF EGM includes 2,191 studies of the effectiveness of interventions, projects and interventions to reduce the involvement of children and young people in crime and offending (White et al., 2021). The included studies were coded by intervention and outcome, as well as some study design and population characteristics, together with an assessment of the level of confidence in each study. The EGM aims to be useful for policymakers and researchers by bringing together all the available global literature relevant to the effectiveness of interventions and outcomes; it includes interventions at all levels (primary, secondary, tertiary and any combination of these), and outcomes related to child, their family, carer, peer, adult, school, professionals and community, as well as offending and crime outcomes.

Whilst, the EGM includes primary evaluations of intervention programmes, the preferred approach to understanding 'what works' is through systematic reviews and meta-analyses of specific intervention programmes (Weisburd et al., 2017). Meta-analyses are undertaken by skilled research teams, but many aspects of the rigorous data collection process that is needed to do them are laborious and often undertaken by multiple researchers working in different projects. The aim of the current project is to facilitate more and better meta-analyses to further our understanding of what works to prevent children and young peoples' involvement in crime and violence.

We extracted effect sizes from primary evaluations included the EGM of interventions to reduce the involvement of children and young people in crime and offending. In doing so, we created an Effect Size Database (ESDB) for a range of outcomes related to the involvement of children and young peoples' involvement in crime and violence. One aim of this database is to help reduce resource duplication in meta-analyses, which can occur when multiple teams independently extract effect sizes from the same papers. By having a centralised and accessible database of all effect sizes reported in primary evaluations, those wishing to conduct meta-analyses of the effects of youth crime prevention interventions could use the database. This would involve, for example, using the database as a 'second coder' for effect size calculations

in future evidence syntheses, thereby reducing the time and resources needed for rigorous meta-analyses, which should use data that has been double-coded (see Lipsey and Wilson, 2001).

Another purpose of the database is to enable systematic comparisons of the distribution of effect sizes across multiple different factors, such as types of interventions and types of outcomes, which would not be feasible with reviews that focus on just one type of intervention or outcome. We anticipate users of the ESDB to draw on the codes presented in several ways. The codes can be used to provide expected effect sizes for particular outcomes, interventions and contexts for sample size calculations to be used in the design of future impact evaluations. For these purposes, the ESDB will be made publicly available by YEF and can be used as open access resource.

There are also potential policy uses of the ESDB. In a vein similar to collaboration between research and policy in the US, this database may serve to inform future policy and practice in youth justice in the UK. In the US, a database created by Mark Lipsey and colleagues included controlled evaluations of various programmes implemented with children involved in serious, and violent, crimes. The findings suggested that interventions reduced reoffending by approximately 12 percent, but the impact could be as large as 40 percent (Howell, 2013; Lipsey & Wilson, 1998; Lipsey, 1999). This project led to a surge in evidence-based juvenile justice policy in the US (Howell, 2013). This report presents the approach we used to design and undertake the coding, which was done by a global team of researchers based in Africa, Asia and the UK and guided by methods and topic experts (Chapter 2). We present information about the database contents, including the magnitudes of effect sizes by key outcomes and study population characteristics (Chapter 3). In a short conclusion (Chapter 4), we discuss how the ESDB might be used and the limitations of the database.

2. Approach

A study protocol was developed and approved by YEF (Appendix 3), drawing on data collected in the EGM, which itself was done in consultation with YEF and sector stakeholders. The project has not been registered with the Campbell Collaboration.

Types of interventions, outcomes, populations and studies

Studies for inclusion in the ESDB were those that were included in the EGM of interventions to reduce the involvement of children and young people in crime and violence. Information about the selection of eligible studies, such as definitions of interventions and outcomes, inclusion/exclusion criteria for population and study design, the search strategy and data extraction approach can be found in the EGM technical report (White et al, 2021)[7]. The EGM includes evaluation studies that report the impact of interventions on more than 20 outcome subdomains across five outcome domains. The five outcome domains are: (1) child-centred outcomes; (2) family and carer outcomes; (3) peer and adult; (4) school/professionals and community outcomes; and (5) offending and crime outcomes. However, the ESDB contains effect sizes from child-centred outcomes and offending and crime outcomes only. The outcomes were chosen in collaboration with YEF primarily because they are thought of greatest interest to policy and practice.

Both the child-centred and offending and crime outcome domains incorporated different sub-domains to capture further salient information about the outcomes reported by primary evaluations. Child-centred outcomes were defined as those that are common risk-factors for involvement in crime and violence (Farrington & Welsh, 2007). Coders also recorded the specific names of outcomes and measurement instruments reported primary evaluations. Often evaluations of interventions implemented with children and young people will not report direct effects on crime and violence outcomes. For example, parenting interventions are often implemented with very young children before offending would reasonably occur and longitudinal follow-ups to measure such behaviours are relatively rare (Piquero et al., 2009).

As such, it is important to establish the impact of interventions on intermediate outcomes, particularly those known to be risk factors for offending behaviours, as supported by

[7] https://youthendowmentfund.org.uk/wp-content/uploads/2021/02/YEF-Evidence-and-Gap-Map-Technical-Report-FINAL.pdf

prominent criminological theory such as the Integrated Cognitive Antisocial Potential (ICAP) theory (Farrington & McGee, 2019). ES for outcomes related to family and carers, peers and adult, school/professionals and community were not collected in this first iteration of the ESDB.

Table 1 summarises the 12 outcome subdomains within this categorisation (definitions of each outcome are given in Annex 2).

Table 1. Outcome domains and subdomains included in the ESDB

Outcome domain	*Outcome subdomain*
Child-centred	Attitudes and beliefs (e.g., beliefs about violence)
	Mental health, internalizing behaviour and self-regulation (e.g., mood ratings)
	Social cognition, skills and pro-social behaviour (e.g., helpful behaviour)
	Attainment and knowledge (e.g., school grades)
	Externalizing and risk-taking behaviour (e.g., fighting)
	Victimisation, abuse and injury (e.g., experiencing bullying)
	Service use, attendance and engagement (e.g., school attendance)
Offending and crime	Violent offences (e.g., charges of assault)
	Serious non-violent offences (e.g., drug offences)
	Other offences (e.g., unspecified offences)
	Antisocial and 'delinquent' behaviour (e.g., obtaining an Anti-Social Behaviour Order)
	Contact with custody services or justice system (e.g., number of arrests)

The EGM includes studies of interventions targeting children and young people aged 0-17 from any country or date (White, 2021). We included all intervention categories in the EGM. For example, 'parents/main care giver(s) focused' interventions (e.g., parent training programmes) and 'mental health and therapeutic interventions' (e.g., counselling) are among the most common intervention categories captured in the EGM (White et al., 2021). The full list of interventions is presented in Annex 1 along with the definitions.

The interventions may target all children and youth, or particular population groups (such as specific ages, sexes/genders and/or or ethnicities). They may be implemented universally, or through targeted programmes and projects. The studies may also report information for all children and youth, or for particular groups, at different points in time following the intervention. The ESDB aims to provide some granularity on effect sizes from these different contexts by presenting information for different population groups, intervention characteristics and study periods.

The EGM contains impact evaluations (IEs) – that is, studies aiming to estimate the direction and magnitude of the effect of the intervention, project or programme on specified outcomes – and process evaluations. The ESDB contains information collected from IEs. We grouped the effect sizes into two main types: those from controlled studies (CS), which are studies where measurement was done contemporaneously (or near-contemporaneously) in an intervention and a control group at the same time; those from uncontrolled studies (UCS), where measurement was only made in the group receiving the intervention or programme, before and afterwards (i.e., pre-test post-test only).

Approach to the data collection

a. Approach to calculating effect sizes

The effect size is the standard measure of the change in outcome associated with the intervention in impact evaluations. There are a number of different measures of effect size used in impact evaluations (e.g., standardised mean differences or odds ratios). Whatever measure is used, effect sizes should be comparable across studies that use different measurement instruments, and primarily represent the magnitude of the change in the outcome. The effect size may also be influenced by other characteristics of the study such as the sample size, which is incorporated in the confidence interval associated with the effect size (Higgins et al., 2021; Lipsey & Wilson, 2001; Waddington et al., 2012).

However, the calculation of effect sizes is not only a statistical matter. There are additional decisions involved in selecting data used in their calculation particularly where an evaluation reports multiple measures of multiple outcomes, possibly in multiple subgroups at multiple time periods of post-intervention follow-up. In the final stage, different types of effect sizes may also need to be transformed into common metrics that are interpretable by decision makers.

We used two types of effect size measurement, which are the most common forms of effect size that we were able to compute: the standardised mean difference (SMD), or Hedges' g, calculated for continuous outcomes (e.g., number of days on probation); and the odds ratio (OR), calculated for dichotomous outcome variables (e.g., arrested versus not arrested). This section discusses effect size calculations for simple and more complex study designs. By simple study designs, we are referring to "straightforward" and clear designs such as a two-group randomised controlled trials (RCTs), where one group is exposed to a new intervention and another group is exposed to something else (e.g., existing services or nothing at all). Assignment of participants to these conditions is determined randomly and outcomes of interest can be measured before and/or after the implementation of an intervention. By more complex designs, we mean studies with multiple levels of sampling (e.g., cluster-RCTs) or quasi-experimental designs (QEDs), where exposures are not randomly allocated and statistical methods such as difference-in-differences or propensity score matching are used. As

there are multiple ways of estimating effect sizes from studies, the calculations here are presented in order of preference. So, calculations based on data fully reported in studies were preferred to calculations based on partial reporting or approximations from inferential statistics (i.e., the more complex option).

i) Continuous outcomes: standardised mean differences

The standardised mean difference (SMD) measures the size of the intervention effect in each study in units of standard deviation observed in that study. The simplest measure of SMD is referred to as Cohen's *d*, which is independent of units of measurement and can be compared between studies that may use different measurement instruments. For simple study designs such as randomised controlled trials (RCTs), the *d* statistic is calculated as the ratio of the mean difference in outcomes to the standard deviation of the outcome, *S(y)*:

$$d = \frac{y_t - y_c}{S(y)} \qquad (1)$$

where y_t is the outcome mean in the treatment group and y_c the outcome mean in the control group. For the denominator, *S(y)*, the pooled standard deviation S_p is preferred:

$$S_p = \sqrt{\frac{(n_t - 1)s_t^2 + (n_c - 1)s_c^2}{n_t + n_c - 2}} \quad (2)$$

where s_t and s_c are the standard deviations in treatment and control groups respectively, and n_t and n_c their respective sample sizes.

The 95 percent confidence intervals (95% Cis) are estimated using the standard error of *d*, se(*d*), given by:

$$se(d) = \sqrt{\frac{n_c + n_t}{n_c\ n_t} + \frac{d^2}{2\,(n_c + n_t)}} \qquad (3)$$

95% CIs are calculated as:

$$d \pm 1.96 * se(d) \qquad (4)$$

The calculation assumes a comparison at one point in time, such as endline (the point at which implementation of an intervention is completed) or a six-month follow up. This is a valid estimate of impact if the average value of the outcome was the same in the treatment and comparison group at baseline. For example, Minor (1988) reported job retention beliefs for probationers. The standardised mean difference and its standard error are calculated using equations (1) and (3), respectively, where the pooled standard deviation uses equation (2) (Table 2).

Table 2. Example ES calculation for SMD effect size

	Mean T	Mean C	SD Treat	SD Control	nt	nc	S(p)	d	se(d)
Retention beliefs	19.50	19.21	2.54	2.62	22	23	2.581	0.110	0.298

Source: data collected from Minor (1988).

SMD should be corrected for small sample bias in evaluations by applying the standard Hedges' g correction factor:

$$g = d\left[1 - \frac{3}{4(n_t + n_c - 2) - 1}\right] \qquad (5)$$

The standard error of g uses the standard formula (equation (3)), incorporating g rather than d, thus:

$$se(g) = \sqrt{\frac{n_c + n_t}{n_c\ n_t} + \frac{g^2}{2\ (n_c + n_t)}} \qquad (6)$$

For example, Bursal and Buel (1980) reported school absences for girls in a small group design. Hedges' g is calculated by applying equation (5), while its standard error uses equation (6) (Table 3).

Table 3. Example of transformation from d to Hedges' g

	n t	n c	d	se(d)	g	se(g)
School absence	17	12	0.603	0.385	0.586	0.385

Source: data collected from Bursal and Buel (1980).

The same approach can be applied to more complex designs and those based on non-randomised assignment of interventions, called quasi-experimental designs (QEDs). Examples of QEDs include difference-in-differences statistical adjustment and matched comparisons. Further information and formulae for these different designs and ES is provided in the study protocol (Annex 3).

ii) Dichotomous outcomes: odds ratios

In the case of dichotomous outcomes, the odds ratio is calculated from the two-by-two frequency table:

$$OR = \frac{p_t/(1 - p_t)}{p_c/(1 - p_c)} \qquad (7)$$

where p_t is the proportion of successful cases (e.g., those who have not been arrested) in the treatment group, $1 - p_t$ the proportion of unsuccessful outcomes (e.g., those who have been arrested). The standard error of the logarithm of OR (log-odds ratio) is given as:

$$se(lnOR) = \sqrt{\frac{1}{n_t p_t} + \frac{1}{n_t(1 - p_t)} + \frac{1}{n_c p_c} + \frac{1}{n_c(1 - p_c)}} \qquad (8)$$

For example, Gillis et al. (2008) reported numbers of youths re-arrested at 6-months and one-year after participation in Behaviour Management through Adventure training (BmtA), compared to those attending Youth Development Centers (YDC). The odds ratios and standard errors are calculated using equations (7) and (8), respectively (Table 4).

Table 4. Example ES calculation for OR effect size

	Treatment		Control		OR	se(lnOR)
Outcome	Not arrested	Arrested	Not arrested	Arrested		
No re-arrest at 6 months	280	67	232	115	2.072	0.177
No re-arrest at 1 year	234	113	179	168	1.944	0.157

Source: data collected from Gillis et al. (2008).

In instances where a 2x2 table used to estimate an odds ratio included a zero (e.g., no participants were arrested), an odds ratio continuity principle was applied. This approach adds a value of 0.5 to all cells in the table, therefore overcoming the inclusion of a '0' in the calculation. Further information and formulae for transformations to OR are given in the study protocol (Annex 3).

iii) Transformations to ensure comparability of effects

Transformations are needed to ensure comparability of effect sizes due to the outcome measure, sampling, and the effect size itself. We converted OR into SMD effect sizes using the Cox transformation:

$$d = \ln(OR)\,\frac{\sqrt{3}}{\pi} \qquad (9)$$

with standard error of Cox-transformed d given as:

$$se(d) = \frac{\sqrt{3}}{\pi}\sqrt{\frac{1}{n_t p_t} + \frac{1}{n_t(1-p_t)} + \frac{1}{n_c p_c} + \frac{1}{n_c(1-p_c)}} \qquad (10)$$

All effect sizes in the ESDB were calculated so that an increase in the effect size measures an improvement or a desirable impact of the intervention (i.e., a positive value of g or an odds ratio > 1). This was consistent across outcomes included in the ESDB, including those where an increase in the outcome represents a desirable intervention effect (i.e., self-esteem, academic achievement, or prosocial behaviour) and outcomes where a reduction represents a desirable intervention effect (e.g., offending, arrests, delinquency, or mental health problems).

In instances where a reduction in the outcome represented a desirable intervention effect (e.g., exclusion from school), d was multiplied by -1, or in the case of OR, the natural log of the odds ratio (lnOR) was multiplied by –1 to change the direction of the effect size. Therefore, a positive value for g or an odds ratio greater than 1 represent a desirable intervention effect for all outcomes. It follows that negative values for g (odds ratios less than 1) represent an undesirable

intervention effect (e.g., an increase in offending, arrests or mental health problems and a decrease in self-esteem, academic achievement, or prosocial behaviour).

Transformations and formulae for variances that address clustering of observations are given in the coder manual (Annex 4).

b. Data selection

Calculating the effect size requires extracting statistical data from the study such as sample size for treatment and control, standard deviation of the outcome for continuous variables, and so on. In practice, authors reported their findings in a wide variety of ways, so a coding manual was developed to account for both differences in study design and ways in which results are reported (Annex 4). A flow chart indicating the decisions that needed to be taken by coders and the types of information they should be collecting is given in Figure 1.

Addressing dependency

As the left-hand side of the decision-making flow chart in Figure 1 shows, effect size calculation is not only a statistical matter. There are additional decisions involved in selecting data used in their calculation, particularly where an evaluation may report multiple measures of multiple outcomes, possibly in multiple subgroups and multiple follow-up periods. Thus, it will usually be possible to calculate multiple effect sizes from each study. For example, in piloting work using included studies in the YEF EGM, we were able to calculate 182 estimates from just three studies (Bursal & Buel, 1980; Gillis et al., 2008; Minor, 1988).

The EGM includes studies that report information about the effectiveness of intervention and prevention programmes. Included evaluations may therefore report the impact of programmes on a number of different outcomes within our categorisation of child-centred and offending outcome domains. However, to add complexity, there are often:

- multiple publications on each study,
- multiple interventions in each study,
- multiple participant subgroups in each study,
- multiple outcome measures in each study (e.g., self-report, parent-report, and official records),
- multiple follow-ups in each study, and/or
- multiple specifications in each study (e.g., adjusted versus unadjusted treatment effects).

The issue of having multiple effect sizes from a study is known as dependency. Dependency is not necessarily a bad thing if it is accounted for properly. For example, where studies report comparative effects of multiple interventions, or multiple outcomes of interest along a causal pathway (e.g., attitudes, behaviour, and re-arrest rates), these can provide useful information for decision making. But in other cases, dependency can be problematic, particularly if a user

of the EGM were to incorrectly assume effect sizes were independent and include them in a meta-analysis without the necessary statistical adjustments for dependency.

We aimed to extract as many relevant effect sizes as possible per evaluation, within the child-centred and offending outcome domains. We therefore reported effect sizes for each relevant intervention, comparison, outcome construct, measurement instrument, or subgroup of participants reported. Where there were multiple publications on the same study, we consulted each publication to gain the most complete record of the evaluation. For example, a research team may publish the findings from a large evaluation separately for different outcomes or for different groups. In this case we may have had to code effect sizes from the same evaluation from separate publications. Similarly, if an evaluation was published as a dissertation and then later a journal article, we favoured the publication with the most complete record of the data. For example, due to journal restrictions on the length of an article, a journal article may not report all information needed to compute an effect size. Where there were multiple follow-up periods, we collected data from the immediate and last follow-up. We included a study identification number for each outcome, so users would know where effects come from the same study, even if paper authorship varies.[8]

In addition, there may also be multiple specifications, such as when results from single difference analysis are presented alongside difference-in-differences. For example, a study may report mean values and standard deviations for a particular outcome after implementation of an intervention but also report the results from regression models. Where studies presented multiple specifications, we used the decision rules indicated in the decision flow diagram (Figure 1) and p. 36 of the coder handbook (Annex 2). Guidance on calculating all-population effects from sub-groups[9] is also given in the handbook.

c. Coding framework, data extraction tool and code book

To help coders efficiently and correctly extract data in accordance with all the rules and principles described earlier, we developed a spreadsheet in which they would code necessary data and compute ES (Annex 5), along with a complete coding manual (Annex 4). The formulae to compute ES are provided in the spreadsheet so that once coders have filled in all the cells with relevant data, the ES and associated statistics are automatically calculated by the coders, who need to apply the appropriate effect size calculation row from a menu of choices.

For each ES we collected the following information:

[8] There was an issue as to what constitutes the same study. It is usually taken to mean studies using the same study sample. But it can also be argued to be studies of the same intervention with the same study population, even if the studies are undertaken by different research groups and the data were collected some years apart. This issue was not resolved, so the ESDB also reports these studies using separate identification numbers.

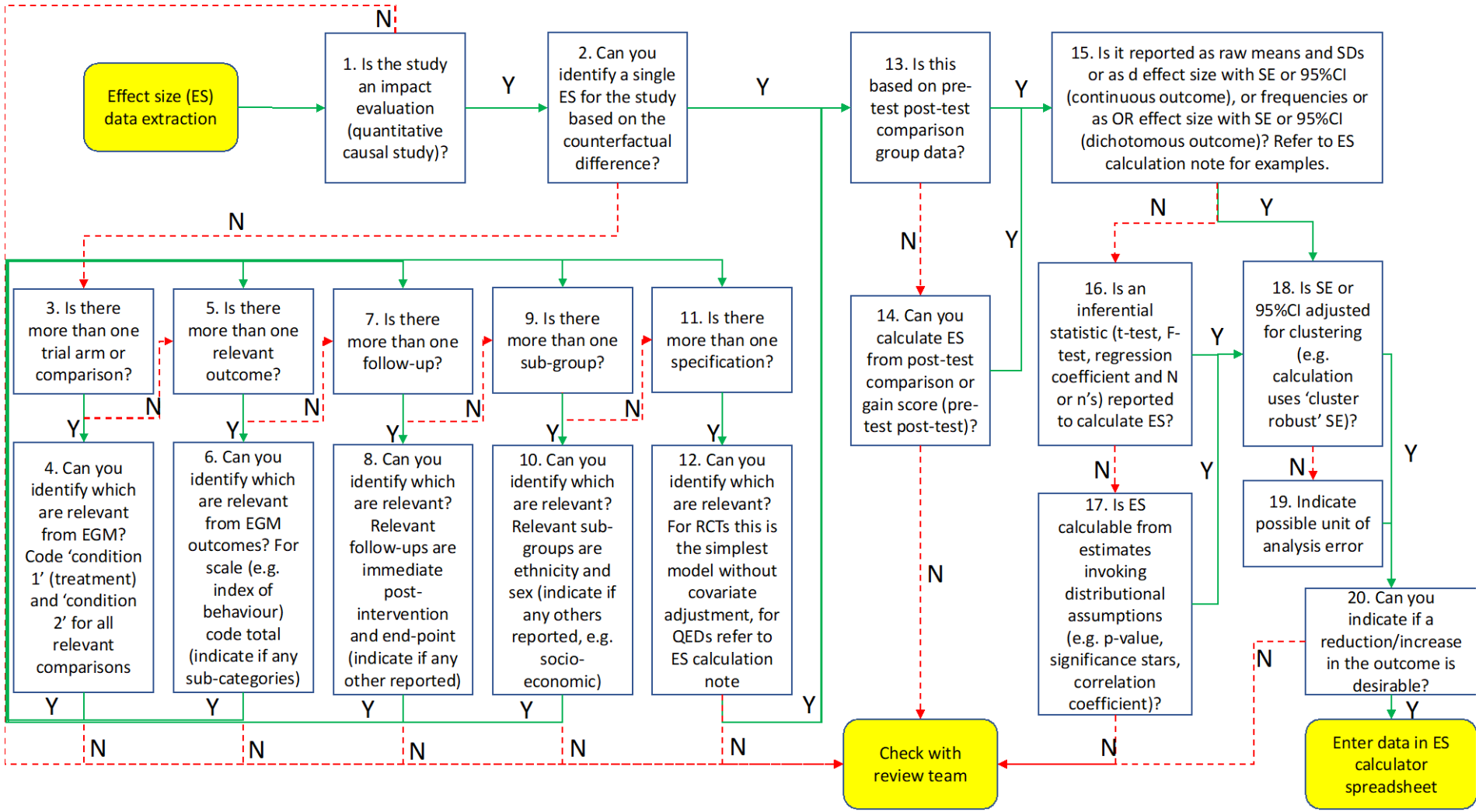
[9] We combined subgroups using "synthetic averages". If the outcome was reported only for subgroups, we calculated effects for all participants by combining these subgroups using the sample-weighted average.

- Coder initials, Study ID, reference for data (page number).
- Intervention name(s) and type(s) of comparison.
- Outcome name(s) and domain(s)/subdomain(s), additional information related to outcome measurement, such as name of the measurement instrument, reporter, name of subgroup, months of follow-up.
- Basic information to compute ES (e.g. Means/SD, regression coefficients, SE/*t*-statistics/*p*-value, sample sizes), number of clusters (if applicable).
- Additional information such as whether increase in outcome is desirable (i.e., the direction of the effect size), whether it is the intent-to-treat estimate, notes on assumptions.

The outline of each variable coded was given in the coder handbook (Annex 4, Table 1). Once the information was collected, the coders would then copy and paste the most relevant formula from example coding at the top of the spreadsheet, which automatically calculated values of *g*.

We also merged the ESDB dataset with the EGM dataset, which includes information about intervention category and subcategory, and characteristics of participants (e.g., age), interventions (e.g., universal or targeted) and studies (e.g., design, confidence assessments).

Figure 1. Coder decision flow chart



Timeline and engagement process

This ESDB was developed following a co-production process where the teams from the YEF and the Campbell Collaboration worked together on defining the coding framework, piloting and coding of studies. The main steps involved in this project are outlined below in Table 5.

There was extended stakeholder engagement in the preparation of the EGM and subsequently the framework for the ESDB, including process management. Initial meetings were held with YEF in May 2022 and subsequently between October 2023 and June 2023, when the research team presented the preliminary findings. Irregular meetings were also held with consultants with YEF on related projects. Coder team meetings were held weekly or fortnightly for within-team meetings, and monthly for all-coder meetings (screenshot from one such meeting in Annex 6).

Table 5. Project timeline

Major step	*Start date*	*End date*
Scoping, developing coding framework, pilot coding, developing coder handbook	April 2022	June 2022
Coder recruitment (round 1)	June 2022	June 2022
Initial coding	July 2022	November 2022
Coder recruitment (round 2)	December 2022	February 2023
Coding of controlled studies	February 2023	May 2023
Coding of uncontrolled studies	May 2023	June 2023
Presentation of initial findings	June 2023	June 2023
Additional analysis	August 2023	August 2023
Generating final database	August 2023	August 2023
Writing final report	August 2023	September 2023
Revisions to database and report following peer review	December 2023	February 2024

Data collection and coder management

a. Training and recruitment

We initially expected to extract effect sizes from over 2,200 effectiveness studies[10] included in the EGM, with 10 percent of studies double-coded as part of quality control. To do this, 23 research assistants were recruited, in this report referred to as coders. The coders were

[10] We included 1,218 studies in the ESDB, since the EGM includes many process evaluations, and in some of the impact evaluations, from which effect size information are taken, ES could not be included for various reasons (e.g., the study did not present sufficient data, the study was a protocol, the study only reported outcomes that were irrelevant to the ESDB project).

recruited and trained during two recruitment drives: the first in June 2022 and second in December-February 2023. We recruited six coders initially,[11] who participated in a three-day in-person workshop, held in Cambridge, UK in June 2022. The first day of training presented coders with the teaching materials, with days two and three reserved for working through examples and problem-solving tasks. The coders initially conducted data extraction only, but at the later stage of the project, they also worked in quality control and as team leads of coders who were recruited in the second drive. For the second recruitment drive, training was conducted with applicants across different continents in January and February 2023: two sessions in-person in India and Ghana, and one online session for candidates in Uganda, United Kingdom, UAE and Cameroon. In this second round, 21 coders were recruited to the project.

b. Team management and quality control

Data extraction work was carried out remotely and took place between July 2022 and June 2023. The 23 coders were grouped into five teams based on the geographic location of the team lead: Ghana team (4), India team (2), Philippines team (3), Uganda team (8), UK team (5 coders), and quality control (2). Within each team, one more experienced coder was appointed as the team lead. The team leads were the first point of contact for queries from their coders. The team leads also managed the work of their team members: each coder was assigned batches of studies each containing 20 to 50 papers depending on their experience. When the coders completed a batch, they first sent their coding sheets to the team lead, who would check for the quality of their work (i.e., minor mistakes such as blank cells and error codes). Once the checking was done, they then forwarded the coding sheet to the quality control team (QC) who undertook double-coding. Double-coding was done independently, for a random sample of studies, following which reconciliation meetings were held between each coder and the QC (and in some cases their team lead). Each coder was responsible for finalising their coding after the discussion and reconciliation with the QC team. Once the coding sheet was finalised, the coders would send it to the team lead for a final check, who then forwarded it to the global team for database compilation. The team leads also coded their own batches of studies. The list of project team members and the organogram are presented in Annex 6

c. Double-coding

Ten percent of studies were double-coded to ensure the quality of the work done by this extensive group of coders. These studies were selected randomly by the QC team. Therefore, each coder did not know which studies would be double coded during their initial coding. The random sample was stratified by each batch of studies, to ensure that each coder would equally have opportunities to discuss their coding with the QC team. Once a coding sheet (with initial coding) was sent to the QC team, they coded the randomly selected 10 percent of the studies independently, and then compared their findings with the initial codes. If there was any disagreement, they would get in touch with the initial coder for discussion and reconciliation. In some cases where coders showed greater concerns about their work, the coders worked in

[11] Four applicants initially participated in the training workshop, but they became unavailable at a later stage.

duplicate to double-code all the studies included in a batch. This was done particularly at the beginning of their contracted period when they received the first set of studies to be coded.

Presentation of the ESDB

The database is presented in the form of a spreadsheet. The columns in the left present information defining each ES, such as outcome name, outcome categories, type of comparison, participant characteristics, months of follow up. To the right of the spreadsheet the computed effect size and variance are given. A column named "notes/assumptions" provides information that needs attention when interpreting the estimated ES. Not all studies reported the required information clearly, and therefore coders had to make some assumptions where reasonable to do so. For example, when a study reported the total sample size, but not the sample sizes of each treatment group, and all the other information such as means and SDs were provided, the coder would have assumed the equal distribution of participants between treatment and control/comparison groups. This column gives explanations on assumptions coders needed to make to compute ES, such as estimating SD in cases where a study only reported the confidence intervals or a value for the standard error.

In the next chapter, we present descriptive analyses using simple means, medians, and other moments of the distribution of effect sizes. Users of the ESDB need to use appropriate statistical methods – invariably, inverse-variance weighted random effects multi-level meta-analysis – to draw generalisable conclusions for decision making about what works.

3. Findings

This chapter presents summary information about the numbers of effect sizes contained in the ESDB, together with an overview of the means, medians and additional distributional moments.

What the ESDB contains

a. Overview

The database currently contains in total 14,834 effect sizes on child-centred and offending and crime outcomes from 1,217 studies. Most of the ES (95%) are from controlled studies (CS): 11,754 ES measure impacts on child-centred outcomes from CS, 2,287 measure offending and crime outcomes from CS, 723 measure child-centred outcomes in uncontrolled studies (UCS), and 70 measure offending and crime outcomes from UCS.

The means of ES (noted as *g*) vary across study type and outcome domain (Table 6). The mean across all ES is 0.23 and standard deviation (SD) is 1.80. Mean and SD for child-centred outcomes in CS are similar to those of the overall average, while offending and crime outcomes in CS have smaller mean and SD. For UCS, the means for both outcome domains are higher with smaller SD.

Table 6. Summary of effect sizes (ES) measured as Hedges' g

Variable	*Observations*	*Mean*	*SD*	*Min*	*Max*
All ES	14,834	0.23	1.80	-18.7	119.9
Child-centred outcome, CS	11,754 (79.2%)	0.26	2.00	-18.7	119.9
Offending and crime outcome, CS	2,287 (15.4%)	0.07	0.48	-2.7	5.2
Child-centred outcome, UCS	723 (4.9%)	0.35	0.81	-8.8	6.2
Offending and crime outcome, UCS	70 (0.5%)	0.73	1.59	-4.4	5.4

Notes: CS controlled studies; UCS uncontrolled studies; SD standard deviation.

b. Summary of ES by outcome domain and subdomain

Table 7 provides the breakdown of outcome sub-domain by study type. Within the child-centred outcomes, the most frequently reported outcome sub domains are 'externalizing and risk-taking behaviour' (n = 3,704 from CS and 196 from UCS), 'mental health, internalizing

behaviour and self-regulation' (n = 2,435 from CS and 256 from UCS), and 'social cognition, skills and prosocial behaviour' (n = 3,032 from CS and 150 from UCS),

Among offending and crime outcomes, for both CS and UCS, the most common outcome subdomain is 'contact with justice system/any custody service' (n = 961 for CS and n = 45 for UCS). While 'antisocial and delinquent behaviour' is the second most common outcomes reported for CS (n = 466), there are only two ES in this subdomain among UCS (n = 2). The 'violent offences' outcome is also commonly reported for CS with 390 ES, but only 9 for UCS.

Table 7. Number of effect sizes by outcome subdomain

Outcome (sub) domain	*Controlled studies*	*Uncontrolled studies*	*Total*
Child-centred	11,754	723	12,477
Externalizing and risk-taking behaviours	3,704	196	3,900
Mental health, internalizing behaviour and self-regulation	2,435	256	2,691
Social cognition, skills and pro social behaviour	2,033	150	2,183
Attitudes and beliefs	1,171	24	1,195
Attainment and knowledge	1,127	8	1,135
Victimisation, abuse and injury	704	58	762
Service use, attendance and engagement	580	31	611
Offending and crime	2,287	70	2,357
Contact with justice system/any custody service	961	45	1,006
Antisocial and delinquent behaviour	466	2	468
Violent offences	390	9	399
Serious non-violent offences	233	4	237
Other offences	237	10	247
Total	14,041	793	14,834

c. Number of ES related to SDQ

Table 8 shows the number of some specific outcomes, namely Strengths and Difficulties Questionnaire (SDQ) (Goodman, 1997) and self-reported delinquency. The database includes 375 effect sizes using the SDQ. Although studies do not always report the 'total difficulties score' along with its subscales, the database includes 163 ES measured by the externalising score (143 from CS and 20 from UCS), and 118 by the internalising score (96 from CS and 22 from UCS). We have also counted the number of self-reported delinquency (regardless of measurement), of which there are 342 ES, mainly from CS.

Table 8. Number of ES related to Strengths and Difficulties Questionnaires (SDQ)

Outcome	*Controlled studies*	*Uncontrolled studies*	*Total*
SDQ (total)	316	59	375
Total difficulties score	21	8	29
Externalising score	143	20	163
Internalising score	96	22	118
Pro-social behaviour	50	8	58

Impact scores	6	1	7
Self-reported delinquency	341	1	342

d. Number of ES by level of intervention targeting

Table 9 shows the proportion of ES associated with targeted interventions, against that from universal interventions. Targeted interventions refer to those that are only available to children or youth at risk of problem behaviour, and universal interventions are those available to all children and youth. Across outcome domains and study types, in general there are more ES from studies of targeted interventions, although the share of ES from universal interventions is slightly higher for child-centred outcomes from controlled studies.

Table 9. Number of ES by level of intervention targeting

	Controlled studies		*Uncontrolled studies*	
Level of targeting	*Child-centred*	*Offending and crime*	*Child-centred*	*Offending and crime*
Targeted intervention	5,779	1,743	479	58
Universal intervention	5,852	517	244	12

e. Number of ES by study design

Table 10 shows the number of ES by study design for each outcome domain (CS only). For both outcome domains, 98 percent of the effect sizes are from either randomised controlled trials (RCTs) or quasi-experimental design (QED), and most ES are from RCTs. The share of RCTs is higher for child-centred outcomes than for offending and crime outcomes, likely reflecting the greater opportunities to design prospective studies like RCTs to measure impacts on child-centred outcomes, than offending and crime. These outcomes are further along the causal pathway and therefore rarer, requiring larger sample sizes for statistical precision, and therefore more likely to be measured in QEDs that use existing data (e.g., household surveys, administrative data).

Table 10. Number of ES by study design (controlled studies only)

	Child-centred	*Offending and crime*
Randomised controlled trial	7,588	1,322
Quasi-experimental design	4,034	938

f. Number of ES by study level confidence

Table 11 shows the share of ES by study-level confidence, which uses the critical appraisals provided by the EGM. For both outcome domains, more than 75 percent of effect sizes are from studies rated as at 'low confidence', which is not surprising given that the ESDB includes both controlled and uncontrolled studies, the latter always being coded as at low confidence. The share of ES based on high-confidence studies is greater for child-centred outcomes (8%) than offending and crime outcomes (4%).

Table 11. Number of ES by study level confidence (controlled studies only)

	Child-centred	*Offending and crime*
High: High confidence on all items	1,033 (RCT=888, QED=145)	110 (RCT=97, QED=13)
Medium: No lower than medium confidence on any item	1,229 (RCT=813, QED=416)	380 (RCT=230, QED=150)
Low: At least one low confidence	9,230 (RCT=5,786, QED=3,444)	1,732 (RCT=979, QED=753)

g. Number of ES by age group

Table 12 shows the share of ES by age groups, using study level characteristics collected for the EGM. Across study types and outcome domains, the most reported age group is 10 to 14-year-olds. ES for 0 to 3-year-old children were rarely reported for offending and crime outcomes (both CS and UCS), as we would expect. On the other hand, the share of ES on adolescents of 15+ years is higher for offending and crime outcomes than child-centred outcomes, which we would expect as it is usually this group that commits more serious offences. ES for 4 to 9-year-old children are mostly reported in CS for child-centred outcomes.

Table 12. Number of ES by age group

	Controlled studies		*Uncontrolled studies*	
	Child-centred	Offending and crime	Child-centred	Offending and crime
0-3 years	847	32	115	-
4-9 years	3,231	213	98	4
10-14 years	5,445	1,411	343	50
15+ years	879	347	66	6

h. Number of ES by sex of participant

Table 13 shows the number of ES by participant sex (not all ES are associated with a sex subgroup). Across study type and outcomes domain, ES for males outnumber those for females, particularly for offending and crime.

Table 13. Number of ES by sex

	Controlled studies		*Uncontrolled studies*	
	Child-centred	*Offending and crime*	*Child-centred*	*Offending and crime*
Female	924	103	13	2
Male	1,263	366	97	11

i. Number of ES by participant ethnicity

Table 14 shows the share of ES by ethnicity and/or the ethnic identity of included children and young people. Similar to participant sex subgroup, not all ES are associated with ethnicity subgroups. Across study type and outcomes domain, ES for majority or white children outnumber those for minority or Black, Asian and Minority Ethnic (BAME) children, which is not surprising as the EGM includes more studies from Western countries, such as Australia, Canada, UK and USA (White, 2021). The EGM does not include an equal proportion of studies conducted in each global region and there is an over-representation of evaluations conducted

in Australia, Canada, UK, and the USA. As a result, we label children and young people identified as White as the majority ethnic group as this is typically the case in such contexts. Moreover, this means that children and young people identified as Black, Asian, mixed ethnicity, or non-White, are often collectively labelled as minority ethnicities/BAME. Participant ethnicity is not generally well reported in evaluation studies, and as such, we adopted a simplified approach to recording participant ethnicity: use the information reported by primary studies. The exception was in any instance where the study was deemed to have used derogatory, prejudiced, or racist language.

Increasingly, children and young people identified as minority ethnicities are over-represented in the criminal justice system, but this disproportionality also does not impact all minority ethnicities uniformly (e.g., YEF, 2023). As such, it is highly concerning, that in our database of effect sizes, the majority of effects for White children and young people outnumber those for minority ethnicities.

Table 14. Number of ES by ethnicity

	Controlled studies		*Uncontrolled studies*	
	Child-centred	*Offending and crime*	*Child-centred*	*Offending and crime*
Minority ethnic	1,180	254	117	8
Majority ethnic	2,485	299	149	22

Analysis of effect size distributions

This section presents the distributions of effect sizes (the standardised mean difference or *g*) contained in the ESDB. Each effect size has been transformed such that a positive value indicates a desired change in outcome. The analysis was performed using Stata. Here we limit ES to those having absolute values of *g* less than 5.[12] Each distribution is presented graphically, using histograms (showing the distribution of effects), usually overlaid with box-plots (showing the mean (in red), median, inter-quartile range and full range). Below each box plot is a table presenting summary information about the distribution including the mean and median, interquartile range, the standard deviation and the sample size (the number of values of *g* for the relevant indicator). We present the analysis for controlled studies (figures associated with each analysis are usually presented in Annex 7; analysis for uncontrolled studies and additional figures are presented in Annex 8). We interpreted each ES magnitude as high, moderate, low, null or harmful based on the bands shown in Table 15.

Table 15. Thresholds for ES magnitude categories

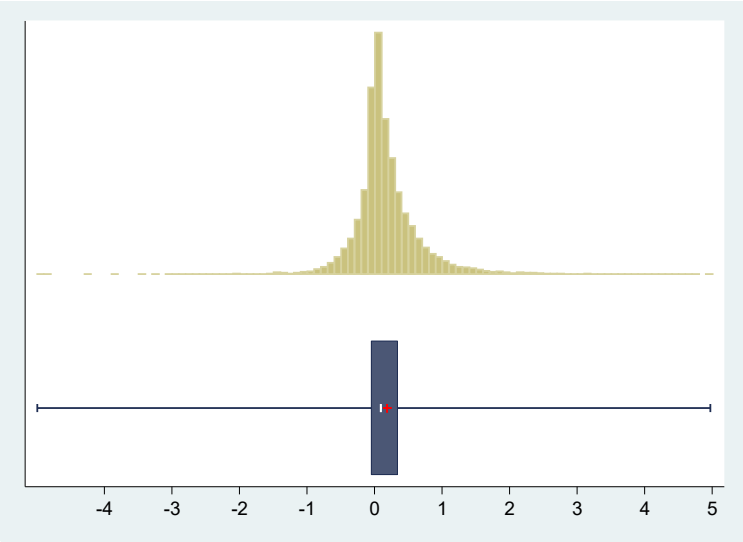
Effect size	*Category of magnitude*
g < -0.04	Harmful
-0.04 <= g <= 0.04	No impact

[12] As a result, 47 ES were excluded from the analysis (44 ES on child-centred outcome and 3 ES on offending and crime outcomes).

Effect size	*Category of magnitude*
0.04 < g < 0.1	Low
0.1 <= g < 0.25	Moderate
g => 0.25	High

Figure 2 presents the overall distribution of ES from controlled studies. The overall mean value of effect size is moderate, with g = 0.18 and SD = 0.56, and the median value at 0.09 is low. The value of ES ranges from -4.99 to 4.97. Skewness is higher than 1 (1.33), indicating the distribution is highly skewed to the right. Kurtosis is also high (15.79) above 3, meaning the distribution has heavy tails and therefore greater tendency for outliers.

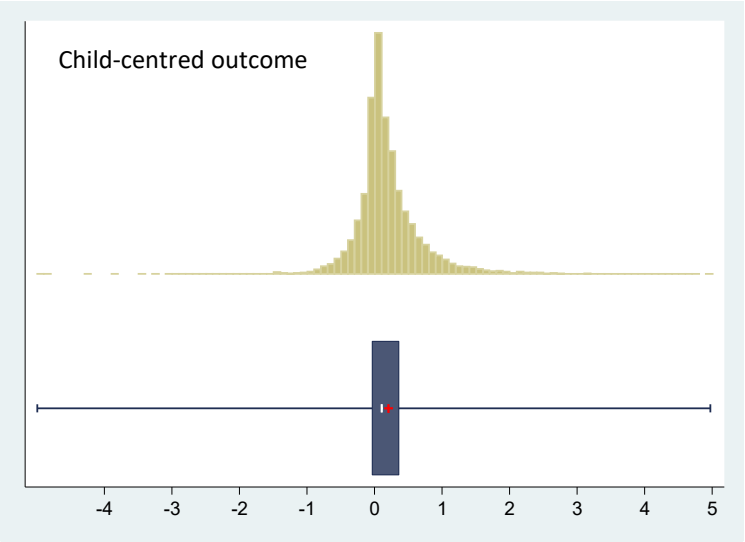
Figure 2. The overall distribution of effect sizes (CS)



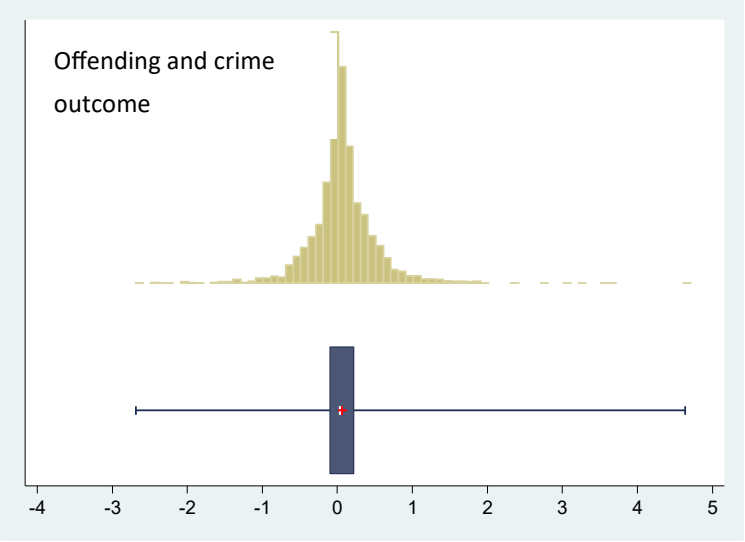
Observation	13999
Mean	0.1839
Std. Dev.	0.5586
Variance	0.3121
Skewness	1.3259
Kurtosis	15.7861
Smallest	-4.9964
Quartile 1	-0.0462
Quartile 2 (Median)	0.0946
Quartile 3	0.3481
Largest	4.9760

Figure 3 shows the distribution of effect sizes by outcome domain. Child-centred outcomes have a higher mean value (M) of 0.21 (but still considered as moderate), higher standard deviation (SD) of 0.57 and higher median of 0.11, compared to offending and crime outcomes (M = 0.07 (low impact); SD = 0.47; median = 0.03). The range is wider for child-centred outcomes at 9.97, while the range for offending and crime outcomes is 7.33. Both outcome domains have right skewness, but child-centred outcomes are more skewed than offending and crime outcomes. Kurtosis is high for both outcome domains thus greater tendency for outliers.

Figure 3. Distribution of effect sizes by outcome domain



Observation	11,713
Mean	0.2066
Std. Dev.	0.5714
Variance	0.3265
Skewness	1.3493
Kurtosis	15.6822
Smallest	-4.9964
Quartile 1	-0.0340
Quartile 2 (Median)	0.1077
Quartile 3	0.3678
Largest	4.9760



Observation	2,286
Mean	0.0675
Std. Dev.	0.4711
Variance	0.2220
Skewness	0.8668
Kurtosis	15.2298
Smallest	-2.6868
Quartile 1	-0.1002
Quartile 2 (Median)	0.0312
Quartile 3	0.2274
Largest	4.6393

The remaining figures relating to these distributions are presented in Annex 7. Annex 7 Figure 4 shows the distribution of effect sizes by outcome subdomain for offending and crime outcome. 'Contact with custody services or justice system' subdomain has a higher mean value of 0.09 (SD = 0.46), which is a low impact, and median of 0.06, compared to other outcome subdomain: for violent offences mean is 0.04 (SD = 0.53) and median is 0.01, and serious nonviolent offences has a mean of 0.01 (SD=0.46) and median of 0.007. For both outcomes the ES are very close to zero and considered to be no impact (for violent offences, the mean falls within the range of low, but very close to no impact). The range is narrower for serious nonviolent offences (2.98) compared to contact with 'contact with custody services or the justice system' (6.38) and violent offences (5.46). Unlike the other outcomes, 'contact with custody services or justice system' is moderately skewed to the right; violent offences and serious nonviolent offences are approximately symmetrically distributed (skewness value ranges between -0.5 to 0.5). For all outcome domains, kurtosis is greater than 3, suggesting potential for noticeable outliers.

Annex 7 Figure 5 shows the distribution of effect sizes by outcome subdomain for offending and crime using administrative data sources only. Like the previous analysis, 'contact with custody services or the justice system' has a higher mean value of 0.08 (SD = 0.47) and median of 0.06, compared to serious nonviolent offences (M = 0.02, SD = 0.48, median = 0.008), but both impacts are of small magnitude or null (having no effects). The mean for violent offences is negative (M = -0.02, SD = 0.57) with the median fairly close to zero (0.005), which suggests on average the effect of the interventions on this outcome is null. The distributions' range, skewness and kurtosis are similar to the previous analysis: the range is narrower for serious non-violent offences (2.98) compared to contact with justice system/any custody (6.38) and violent offences (5.46). 'Contact with justice system/any custody' is moderately skewed to the right, while violent offences and serious nonviolent offences are approximately symmetric. Based on the values for kurtosis greater than 3, there are potential noticeable outliers for all outcome subdomains.

Annex 7 Figure 6 shows the distribution of effect sizes for outcomes measured using SDQ. The total difficulties score has a higher mean value of 0.24 (SD = 0.27) and median of 0.14, which indicates moderate (close to high) impact on the outcomes measured with this score. For the subscales, externalizing behaviour score, internalising behaviour score and prosocial behaviour score, the mean ranges between 0.12 and 0.15 which is moderate effect, and SD between 0.31 to 0.40. The range is 1.02 for total difficulties score, 4.03 for externalising behaviour score, 2.69 for internalising behaviour and 2.08 for prosocial behaviour score. Apart from total difficulties score, the distributions are heavily skewed to right, and the values for kurtosis are considerably high, suggesting great instances of noticeable outliers. The distribution of ES for externalising behaviours, internalising behaviours and prosocial behaviours from non-SDQ indexes are presented in Annex 8 as additional figures.

Annex 7 Figure 7 shows the distribution of effect sizes for self-reported delinquency. ES is usually moderate with mean of 0.11 (SD = 0.37) and median of 0.06, with the range of 2.58. The distribution has a positive (right) skewness and heavy tails (outliers).

Annex 7 Figure 12 reveals the distribution of effect sizes for child-centred outcome by age group. High level of effect sizes can be observed for 0-3 year (M=0.36, SD=0.67, median =0.31) and +15 year (M = 0.28, SD = 0.68, median = 0.09). The mean ES for 4-9 years is 0.21 (SD = 0.53, median = 0.14) and for 10-14 year, the most frequently reported ES, mean is the lowest (M = 0.18, SD = 0.55, median = 0.08) but still considered to be moderate. The range of the ES is 7.59 for 0-3 year, 8.91 for 4-9 year, 9.97 for 10-14 year and 7.26 for 15+ year. Apart from for the 0-3 years group, the distributions are skewed to the right. All age group distributions have considerable tendency for outliers.

Annex 7 Figure 13 shows the distribution of effect sizes for crime and offending outcome by age group. The effect sizes are smaller than for child-centred outcomes, but relatively higher effect sizes can be observed for 4-9 year (M = 0.20, SD = 0.54, median = 0.16), which is moderate. The other age groups have mean ES of close to zero (M= 0.08, SD = 0.55, median =

0.13 for 0-3 year; M = 0.05, SD = 0.50, median = 0.019 for 10-14 year; M = 0.08, SD = 0.34, median = 0.04 for 15+year) indicating the impact is low (or close to null) on these age group. The range of the ES is 2.80 for 0 to 3-year-olds, 5.89 for 4 to 9-year-olds, 6.38 for 10 to 14-year-olds and 2.71 for those aged 15+. Apart from 0-3 years, the distributions are skewed to the right. All age group distributions have a great tendency for outliers.

Table 16 provides an overview of all distributional analyses described above and presented in Annex 7, giving mean and median ES with colour coding to indicate the likely magnitude of the estimated effects.

Table 16. All distributional analyses

	Controlled studies				*Uncontrolled studies*			
	Child-centred		*Offending & crime*		*Child-centred*		*Offending & crime*	
	mean	*median*	*mean*	*median*	*mean*	*median*	*mean*	*median*
Overall	0.2066	0.1077	0.0675	0.0312	0.3623	0.2808	0.5952	0.0745
Contact with justice system/any custody service	-	-	0.0874	0.0563	-	-	0.5159	0.0603
Serious non-violent offences	-	-	0.0143	0.0074	-	-	0.8030	0.6988
Violent offences	-	-	0.0406	0.0126	-	-	1.4373	1.6076
Contact with justice system/any custody service (admin. data)	-	-	0.0785	0.0563	-	-	0.5790	0.0707
Serious non-violent offences (admin. data)	-	-	0.0229	0.0081	-	-	1.2052	1.2052
Violent offences (admin. data)	-	-	-0.0208	0.0050	-	-	2.0942	2.2219
SDQ (total difficulties score)	0.2427	0.1391	-	-	0.2614	0.4192	-	-
SDQ (externalising score)	0.1504	0.1299	-	-	0.2377	0.2023	-	-
SDQ (internalising score)	0.1224	0.0959	-	-	0.2072	0.2224	-	-
SDQ (pro-social behaviour)	0.1527	0.0527	-	-	0.0026	-0.0564	-	-
Self-reported delinquency	-	-	0.1102	0.0587	-	-	0.5696	0.5696
RCT	0.2087	0.1094	0.0636	0.0359	-	-	-	-
QED	0.1944	0.1025	0.0753	0.0240	-	-	-	-
Aged 0-3 years	0.3593	0.3078	0.0805	0.1330	0.4674	0.4626	-	
Aged 4-9 years	0.2124	0.1393	0.1974	0.1622	0.3041	0.1940	0.0265	0.0806
Aged 10-14 years	0.1751	0.0805	0.0458	0.0190	0.3431	0.2618	0.6600	0.0745
Aged 15+ years	0.2803	0.0892	0.0797	0.0375	0.4631	0.4708	1.4005	1.4064
Female	0.3001	0.1169	0.2016	0.0765	0.3888	0.3451	0.0353	0.0353
Male	0.2298	0.1184	0.0631	0.0493	0.2579	0.1775	0.8985	0.6397
Minority ethnic	0.2829	0.1366	0.0301	0.0257	0.0931	0.0540	-0.0318	-0.0419
Majority ethnic	0.2329	0.1385	0.0536	0.0300	0.5240	0.4294	0.7958	0.3519
Targeted intervention	0.2519	0.1471	0.0695	0.0336	0.3531	0.2935	0.7173	0.1090
Universal intervention	0.1578	0.0829	0.0649	0.0280	0.3807	0.2607	0.0254	0.0434
Female, gender-targeted intervention	0.5512	0.2857	0.8758	0.4901	-		-	
Female, non-gender-targeted intervention	0.1416	0.0769	0.1117	0.0661	0.3888	0.3451	0.0353	0.0353

	Controlled studies				*Uncontrolled studies*			
	Child-centred		*Offending & crime*		*Child-centred*		*Offending & crime*	
	mean	*median*	*mean*	*median*	*mean*	*median*	*mean*	*median*
Male, gender-targeted intervention	0.2940	0.1823	0.0638	0.0748	0.2557	0.1596	1.4698	1.2052
Male, non-gender-targeted intervention	0.1537	0.0892	0.0608	0.0170	0.2700	0.3391	-0.1012	-0.1623
Minority ethnic, targeted intervention	0.2874	0.1334	0.0280	0.0198	0.0946	0.0540	0.0840	0.0976
Minority ethnic, non-targeted intervention	0.1555	0.2380	-0.1521	-0.1521	0.0325	0.0692	-0.1476	-0.1122

Note: in most cases of a harmful effect, the sample sizes are small (e.g., n=8, n=4, n=2).

Legend	
High (g => 0.25)	No impact (-0.04 <= g <= 0.04)
Moderate (0.1 <= g < 0.25)	Harmful (g < -0.04)
Low (0.04 < g < 0.1)	- Not applicable/ no observations

Conclusions

This final chapter concludes with a discussion of the findings, the limitations of the ESDB and presents information about how to use it.

Discussion

We used a rigorous process to collect data on over 15,000 effect sizes on child-centred and offending and crime outcomes from over 1,200 studies, of which more than 90 percent were from controlled studies (RCTs and quasi-experiments). Most of the ES (84%) were measured for child-centred outcomes (such as externalising behaviour, mental health and social cognition), while 16 percent were measured for offending and crime outcomes (e.g., delinquency, contact with police, etc).

The data were collected by 23 coders who were trained by us for the purposes of this exercise. We conducted five separate training events in the UK (2 trainings), India, Ghana and Uganda. We also adopted a rigorous quality control process, whereby a random sample of 10 percent of all studies, stratified by coder, were double-coded by the quality control team.

The discussion of ES presented in this report uses descriptive analysis that is based on simple means, medians, and other distributional moments. Our aim here was to describe observed patterns and tendencies in ES distributions. Thus, we have not taken into account ES variance when calculating the means, nor dependencies in the data – for example, due to the ESDB containing multiple ES from each study. We would expect any users of the ESDB to use sophisticated methods of statistical meta-analysis, taking into account effect size variances, dependency and heterogeneity of interventions and other contextual factors. Attempts to draw more generalisable, policy-relevant conclusions from the data were beyond the scope of our project.

From the analysis of distributions, we observed several broad trends in average values of ES in the ESDB. In general, means of ES were larger in the uncontrolled studies than controlled studies. We expected this to be the case, since UCS provide 'low confidence' findings that do not usually represent the difference that the intervention, project or programme has made to participants, over and above other factors that might cause outcomes to change (such as

general improvements in conditions for children and young people). In order to measure effects reliably, one usually needs information from controlled studies like RCTs and QEDs, and some types of uncontrolled quasi-experiments (e.g., interrupted time-series study designs made possible by administrative data collection).

In controlled studies, the mean ES was larger for child-centred outcomes than for offending and crime outcomes. This is not surprising, since programmes have potentially more influence on child-centred outcomes than offending and crime, which are further along the causal pathway and hence subject to a much greater set of external factors (e.g., the extent of crime and violence in society, the extent and capability of policing). This observation is consistent with the funnel of attrition (White, 2014). However, this was not the case for uncontrolled studies, suggesting UCS may be misrepresenting the true magnitude of changes in outcomes.

Studies in the ESDB using RCT methods tended to yield ES of greater magnitude than those using QEDs, whereas ES from studies with high confidence were also larger than those from medium or low confidence studies. Within the same confidence level, RCTs tended to have larger effect sizes than QEDs. We calculated ES using the intention-to-treat (ITT) where possible, which was usually possible in RCTs but not always for QEDs. The ITT is usually smaller than other approaches (such as the effect of adhering to treatment protocols, also called the average treatment on the treated effect, ATET) because it includes all study participants who were offered the intervention, regardless of their eventual participation – hence ITT is sometimes also considered the more policy-relevant quantity. The finding that RCTs of higher confidence have larger effects than other studies is consistent with 'site selection effects'. That is, where RCTs are well designed and conducted, they tend to generate bigger effects due to having been designed more appropriately to the study context and participants and/or have better implementation (or perhaps there may be a motivational effect of the trial on the staff delivering the intervention) (Allcott, 2015). This finding potentially makes the case for more prospective studies, and RCTs in particular, of 'real world' interventions to improve programming for children and youth.

Studies with targeted interventions tended to yield larger ES than those with universal interventions. In addition, interventions that were gender-targeted tended to have greater effects on participants of the same gender, particularly female participants in female-targeted programmes, for whom the effects were very large. Effects for males in male-targeted programmes were also large for child-centred outcomes, but not for violence and offending.

On average, for both child-centred and offending and crime interventions, ES were larger for female subgroups than male subgroups on average. As expected, for both subgroups, the average ES was greater when interventions took a tailored approach for the targeted sex subgroup. While we did observe differences in average ES for child-centred outcomes between ethnic minority and other groups (where minority groups tended to have on average bigger ES), differences in ES were smaller for offending and crime outcomes (which also tended to have smaller ES on average than for child-centred outcomes). However, these differences are only suggestive. More research is also needed of the ESDB, by, for example, analysing

distributions of ES by gender and ethnicity subgroups and by intervention type (and other relevant information). This analysis should use appropriate methods of statistical analysis which addresses issues of dependency in the ES contained in the database, accounts for between-study heterogeneity, by using random effects meta-analysis models, and uses a more appropriate method of weighting studies, such as by the inverse of the variance.

The subgroup analysis by age suggested that the average impact was higher among 0 to 3-year-olds and youths aged 15 years and older for child-centred outcomes, while for offending and crime outcomes, programmes tended to have larger ES on average among children aged 4-9 years. This might suggest that targeted interventions, implemented before the onset of involvement in crime and violence and/or development of certain risk factors, may be particularly effective, and could be explored in future studies. ES were most frequently reported for children aged 10 to 14 years old, but the magnitude of the average effect tended to be lower than for other age groups, for both child-centred and violence and offending outcome domains. Further statistical analysis of the ESDB using appropriate meta-analysis models is needed to confirm this finding, and therefore whether future programmes need to be redesigned for this age group.

Limitations of the ESDB

We independently double-coded a random sample of 10 percent of studies, stratified by coding batch. The final codes were reached through reconciliation. We indicate in the database which ES were independently double-coded. We were not able to independently double-code and verify the remaining 90 percent of studies, hence our suggestion to use the codes cautiously in research, for example, by using the ESDB as a 'second coder' in evidence synthesis work.

We could not capture all the information given in each study due to resource constraints. For example, we could not compute ES for outcomes that were classified outside of the 'child-centred' or 'offending and crime' domains, or from non-priority sub-groups (e.g., particularly 'at-risk' participants), or some data collection points (e.g., intervention mid-point). Future iterations of the ESDB could attempt to incorporate these additional outcomes, subgroups and time points as they could lead to interesting and impactful research questions.

We could not compute ES from all the studies included in the EGM, due to lack of necessary data from impact evaluations (e.g., studies reporting means only, without standard deviations), and also because the EGM included process evaluations which do not aim to provide effect sizes. Similarly, not all studies gave clear information needed, and therefore we had to make some reasonable assumptions to make the full use of the evidence base: for example, when a study reported only total sample size (without sample sizes for each treatment group) and all the other information was provided (e.g., means and SD), we would assume the sample size of the intervention group was the same as that of the control group. This was done to produce the most approximate possible estimation of ES, but we cannot be certainly sure that these assumptions are valid or not.

The assessments of study confidence were conducted at the study level (as part of the EGM process). More rigorous systematic analysis requires these assessments to also be undertaken at the outcome level (Higgins et al., 2021). This should be addressed in any systematic reviews and meta-analyses that use the ESDB.

While the outcome categorisation and definitions were developed through engagement with external stakeholders when producing the EGM, we acknowledge that there might be disagreement among experts on these definitions. Since the ESDB is a follow-on project from the YEF EGM, we followed the definitions developed by the EGM project to ensure consistency. Due to the structure of the ESDB, it will be a straightforward process to move particular outcomes into different domains and sub-domains.
There are also a few additional technical limitations of the ESDB:

- Regarding double-difference estimates, in the small number of cases where raw means or frequencies were used to compute ES, the correlations between pre-test and post-test data were not addressed in the effect size formulae.
- Where clustering was present, but the study did not report cluster-adjusted inferential statistics, we did not calculate a variance with cluster adjustment; in such cases the cluster robust variance is reported as NA.
- The issue of dependency arising from multiple papers on the same intervention and same participants was not resolved. Potentially linked studies can be identified by filtering the spreadsheet by intervention name (column E "Condition 1 (name intervention)"), which will help users to decide how to use the information.

How to use the ESDB

This ESDB contains 14,834 ES from 1,217 studies, or 12 ES on child-centred or violence and offending outcomes per study on average. It is a unique publicly available and open access resource that will potentially inform researchers and research commissioners of the possible impacts reported in studies on youth violence prevention. In this section, we present guidance on how to use the database for researchers, and the limitations of what we have done.

We envisage two main groups of users of the ESDB: 1) those involved in designing impact evaluations; and 2) those conducting systematic reviews and meta-analyses. Ethical design of primary studies, particularly large and expensive studies like RCTs, requires knowledge about the likely effect(s) associated with the intervention, if possible, in the study context, for the targeted populations and outcome constructs. For clustered designs, information is also needed on the likely correlations of observations within clusters for particular outcomes. This information is contained in, or can be extracted from, the ESDB, helping to ensure that future impact evaluations are designed based on existing empirical findings.

For those conducting systematic reviews and meta-analyses, the ESDB contains more information than presented in our analysis in Chapter 3, which is based on only some aspects

of the data; in other words, outcome domains, sex, age group and ethnicity. These include intervention name, study setting, country, specific outcome measures (e.g., 'Child Behaviour Checklist'), the outcome reporter, and the number of months from intervention endpoint until follow up. We envisage systematic review and meta-analysis authors using the ESDB as the 'second coder' for the effect sizes and relevant data extraction variables, from which rigorous meta-analyses can be conducted.

The analysis of ES distributions presented in this report are for descriptive purposes only. Appropriate analysis of the ESDB to draw conclusions for decision making, would necessarily use more advanced statistical methods including inverse-variance weighted random effects meta-analysis with multi-level modelling to account for dependencies in the data contained in the ESDB.

The method we have developed can be applied to engage more fully with evidence on youth offending, such as by collecting outcomes measured among family, teachers, peers and other non-priority outcomes which were excluded from the ESDB. The data extraction framework and approach can also be applied to other policy areas, to produce ESDBs on other topics.

Annexes

Annex 1: Intervention subcategories definitions

Category	*Sub-category*	*Definitions*
Supporting better behaviours	Mental health and therapeutic interventions	Any recognised talking therapy, or intervention aimed specially at improving or treating mental health concerns.
		Includes both individual and group. (Castillo, 2019)
	Mentoring and supportive relationships	Interventions [that] connect people who have specific skills and knowledge (mentors) with individuals (protégés) who need or want the same skills and advantages to move up in work, skill level, or school performance.' (Community tool box, n.d.)
		This broadly includes building supportive relationships with key adults.
	Educational and vocational interventions	Interventions that focus on gaining specific knowledge or that lead to educational or career progressions (Lestrud, 2013; Mau, 2008)
	Sports, recreation and community activities	Interventions that promote the pursuit of positive activities such as sport or creative endeavours. (Khasnabis, 2010)
	Social and emotional interventions	Interventions which aim to improve children's interaction with others and self-management of emotions (2) (Education Endowment foundation, n.d.)
	Practical life skills	Activities that focus on developing skills of daily living and/or planning for adult life. (Prajapati, 2017)
Addressing problem behaviours	Gang and criminal network interventions	A gang is defined by the social relationships of its members with each other and with those outside the group. This category, therefore, includes any intervention aiming to reduce gang related outcomes such as gang membership and activities or involvement in organised crime (Michael Sierra-Arevalo, 2017)
	Child exploitation and contextual safeguarding	Practices and procedures to reduce harm to children outside of the family home

Category	*Sub-category*	*Definitions*
		(University of Bedfordshire, 2020) , including those specifically related to child exploitation
	Alcohol and Drug interventions	Interventions addressing alcohol and/or drug related outcomes, including but not limited to direct use.
	Anti-bullying interventions	Any interventions that identify as 'anti-bullying' or is aimed at reducing persistent aggressive behaviour that is intended to cause another child harm or discomfort (American Psychological Association, n.d.).
	Direct violence prevention	Any intervention specifically aiming to reduce or eliminate violence. Examples would include dating violence programmes or programmes to reduce reactive aggression.
Family and carer interventions	Parent/main care giver(s) focused	Interventions that focus on addressing behaviours/attitudes/outcomes for parental figures. Parenting skills would be categorised here.
	Family members focused	Interventions that target, or address, whole families and/or family systems, or include familial relationships outside the main carers. This is equivalent in care settings.
System approaches	Schools and service coordination and improvements	Interventions pertaining any changes in the way services are delivered including developing service personnel or systems or procedures. Co-ordination between services in the same sector are captured here (e.g., transition between schools). Excluding justice system or contextual safeguarding specific activities.
	Public health and multi-agency working approaches	Pertaining to changes in whole systems or multi-agency working to promote maximal health for all (Public Health England, 2019). Co-ordination between services across sectors are captured here.
Justice and opportunity-based crime prevention	Justice system interventions	Changes or adjustments to justice processes or interventions targeted at justice professionals and/or are conducted in justice settings such as prisons or police facilities.
	Opportunity based crime prevention	Interventions that increase risk/difficulty of committing a crime (Clarke, 1995). For our purposes, this would include behaviour restrictions (e.g., curfews and anti-social behaviour orders, ABSOs) as well as environmental factors (e.g., lighting and closed-circuit television, CCTV).

Source: White et al. (2021)

Annex 2: Outcome subdomain definitions

Outcome Domain	*Outcome subdomain*	*Definition*
Child-centred	Attitudes and beliefs	Any attitude or belief relating directly to crime/aggression or identified risk and protective factors (e.g., violence ideation, offending attitudes, moral beliefs, and attitudes to school). For our purposes this includes goals and future aspirations.
	Mental health, internalizing behaviour and self regulation	Outcomes relating to managing emotions, impulses such anger management, ability to manage impulsivity and distractedness and other mental health components, favourable or encouraging estimate or opinion/belief and attitude among oneself (e.g., self-esteem and self-worth). Mental health status and diagnoses are also included in this category (Bogee, 1998).
		Internalizing behaviour problems are described as inward occurrences, displaying as an inhibited style described as withdrawn, lonely, depressed, and anxious (McGrath, 2015).
		Self-regulation refers to skills described above, outside of the mental health context, for example general anger management. Resilience is also included here.
	Social cognition, skills and pro social behaviour	Pertaining to understanding and relating to others. Including: empathy, attribution style, conflict resolution style (Frith, 2006).
		Outcomes related to improved interactive and communication skills with others in the society and community measures of an individual's social network (Kugler, 2015) and sense of connectedness.
		Pro social behaviours are positive behaviours that children can engage in for example assisting with household or classroom tasks.
	Attainment and knowledge	Outcomes relating to achievements (academic or extra-circular), or measures of specific knowledge gained. For example, educational attainment, sports achievements or knowledge about knife crime. This includes cognitive outcomes such as memory and task switching, as well as age-dependent developmental measures.
	Externalizing and risk-taking behaviours	Any measure of externalizing behaviours including aggression and rule breaking behaviour or risk-taking behaviour such as gambling, running away, truancy and drug and alcohol use (Movallali, 2017).
	Victimisation, abuse and injury	Any measure of individual victimisation including victim of crime, abuse/neglect, victim of bullying or harassment, an imminent risk of serious harm and/or relevant physical health outcomes such as wound severity or diagnoses (Barajas, 2017).
	Service use, attendance and engagement	Any measures of participation in activities/services/community, including measures of involvement with activities/services; e.g., service utilisation, involvement with family/peer activities, use of community activities, employment and classroom behaviour.
Offending and crime	Violent offences	Any measure or record of recognised violent crimes such as assault, murder/manslaughter, use of weapons, robbery at an individual and community level (NIJ, n.d.).

Outcome Domain	*Outcome subdomain*	*Definition*
	Serious non-violent offences	Any measure or record of recognised serious crime that is not violent which includes drug supply offences and possession of an offensive weapon as serious non-violent offences, and burglary of a dwelling (not aggravated burglary). At an individual or community level.
	Other offences	Any measure or record of undifferentiated offences, total offences including for individuals and communities, and offences not included above.
	Antisocial and delinquent behaviour	Any measure or record of acting/behaviour that is likely to cause alarm or distress over a period of time (Shelter Scotland, n.d.).
	Contact with justice system/any custody service	Any measure or record of contact with any teams of services within the criminal justice or custody service.

Source: White et al. (2021)

Annex 3: Study protocol is provided as a separate file

Annex 4: Coder handbook is provided as a separate file

Annex 5: Data extraction tool is provided as a separate file

Annex 6: List of project members and organogram

Name	*Role*
Ashima Mohan	Global team
Hannah Gaffney	Global team and quality control team lead
Howard White	Global team director
Hugh Sharma Waddington	Global team lead and quality control team
Carrel Fokou	UK team
Diwakar Basnet	UK team
Favour Ezeh	UK team
Saranya Mohan Das	UK team
Hikari Umezawa	UK team lead and quality control team
Adam Abdul-Rahaman	Ghana team
Desmond Kaledzi	Ghana team
Joseph Darko	Ghana team
Sheila Agyemang Oppong	Ghana team lead
Kishore Basak	India team
Neha Gupta	India team
Alyssa Cyrielle Villanueva	Philippines team
Lovely Tolin	Philippines team
Nina Dela Cruz	Philippines team lead
Sangyoung Jung	Quality control team
Francis Nkunzimaana	Uganda team
Maria Kizza	Uganda team
Pastan Lusiba	Uganda team
Philip Orishaba	Uganda team
Rachel Wangi	Uganda team
Regina Ndagire	Uganda team
Solomon Kamurari	Uganda team
Robert Apunyo	Uganda team lead

Figure A6.1. Organogram

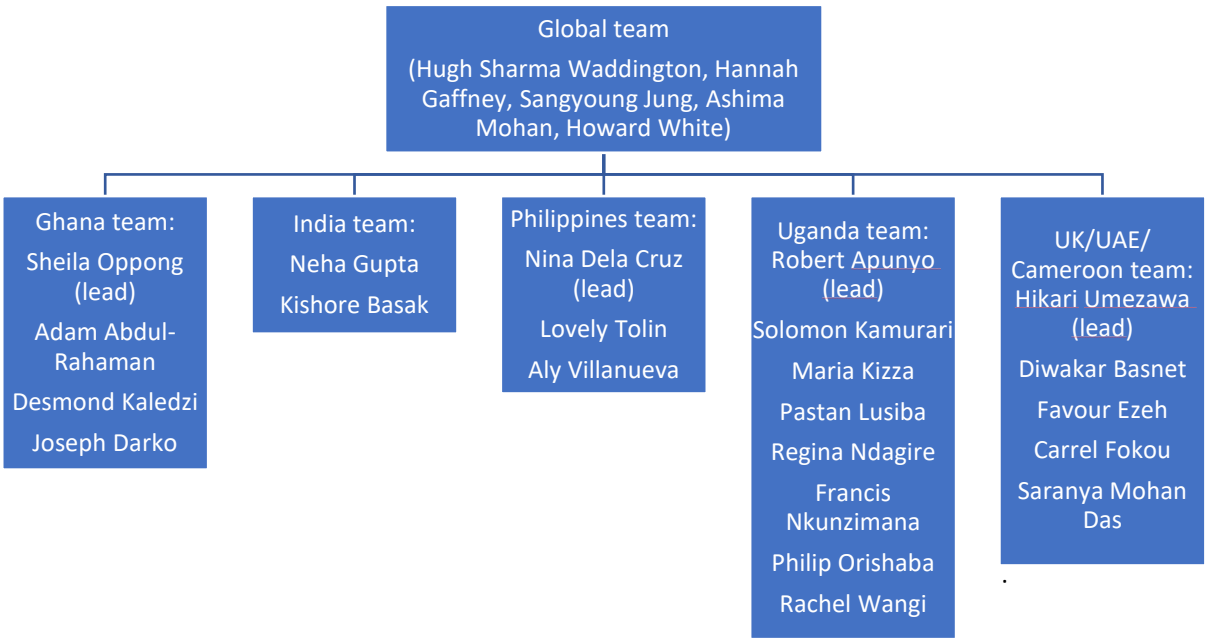
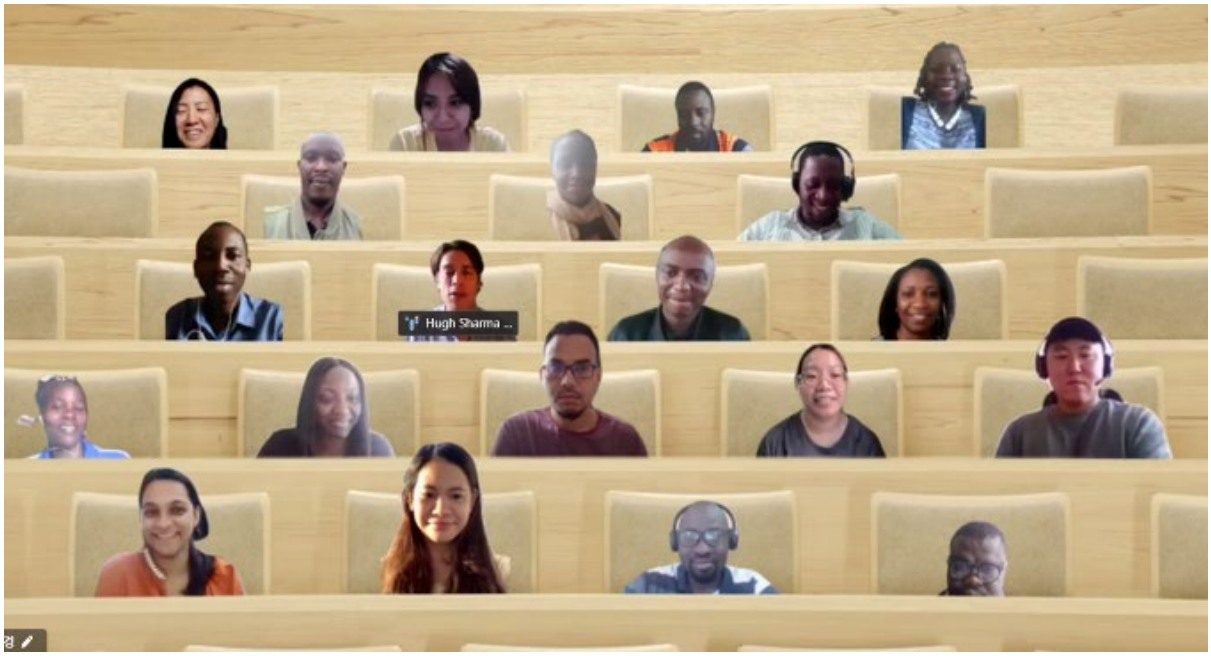
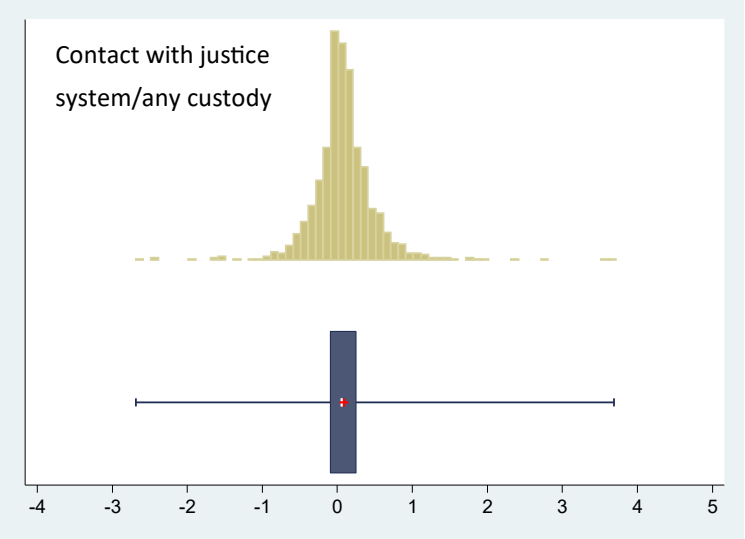


Figure A6.2. All-coder meeting screenshot

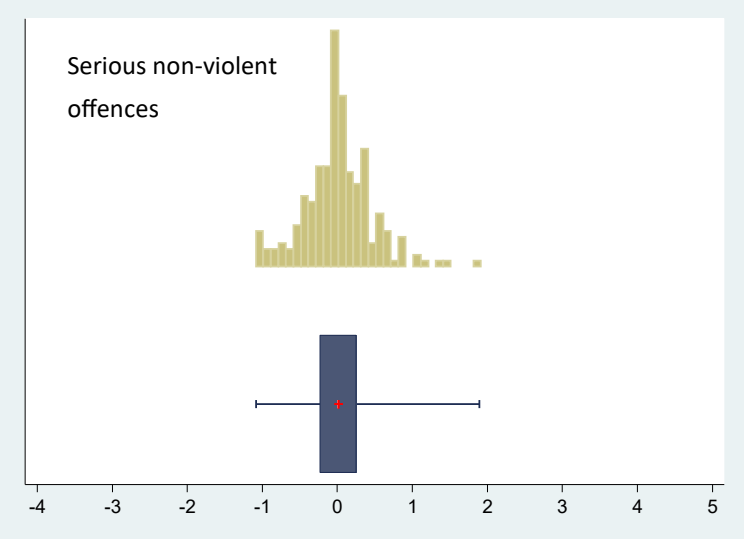


Annex 7: Figures used in the distribution analysis

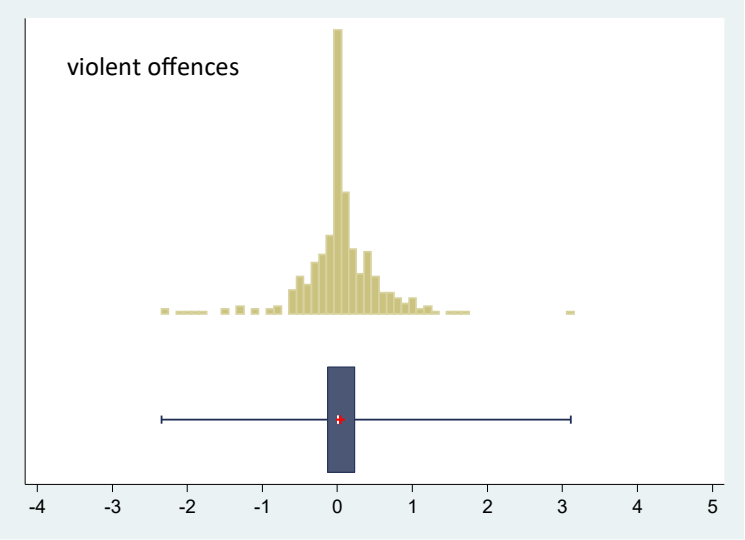
Figure 4. Distribution of effect sizes by outcome subdomain (offending and crime)



Observation	961
Mean	0.0874
Std. Dev.	0.4631
Variance	0.2145
Quartile 1	-0.0932
Quartile 2 (Median)	0.0563
Quartile 3	0.2598
Range	6.3757
Skewness	0.7517
Kurtosis	15.5629

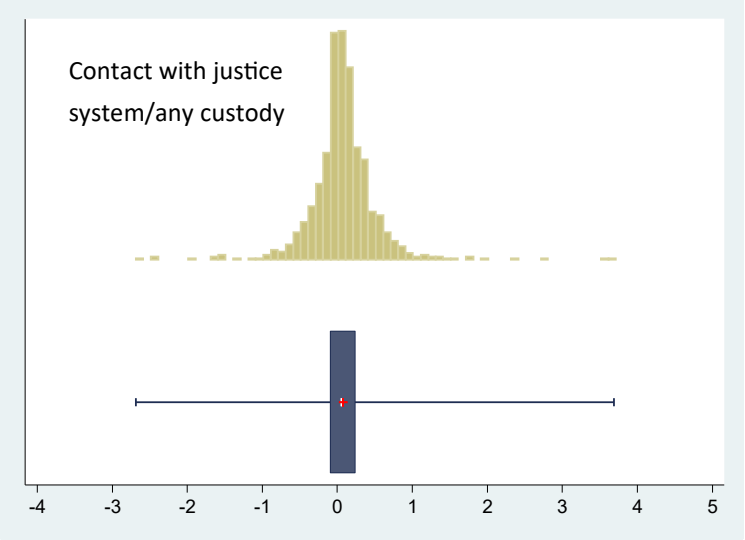


Observation	233
Mean	0.0143
Std. Dev.	0.4555
Variance	0.2074
Quartile 1	-0.2292
Quartile 2 (Median)	0.0074
Quartile 3	0.2580
Range	2.9780
Skewness	0.3150
Kurtosis	4.4762

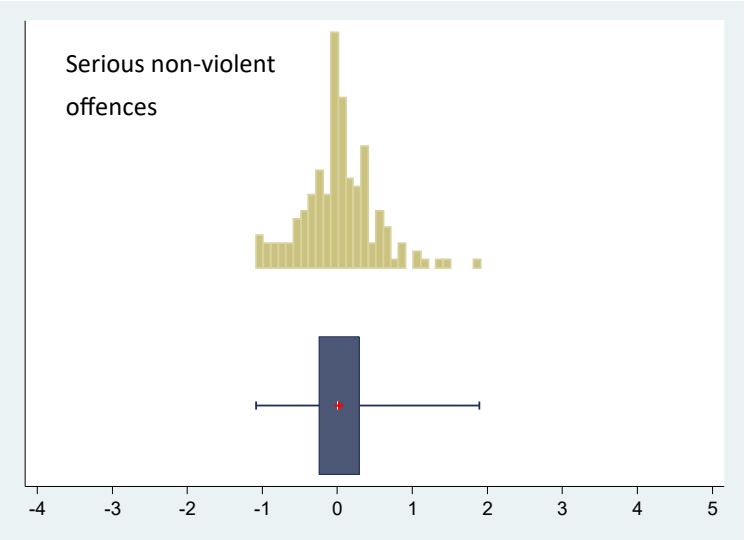


Observation	390
Mean	0.0406
Std. Dev.	0.5258
Variance	0.2764
Quartile 1	-0.1269
Quartile 2 (Median)	0.0126
Quartile 3	0.2354
Range	5.4553
Skewness	-0.3288
Kurtosis	9.5785

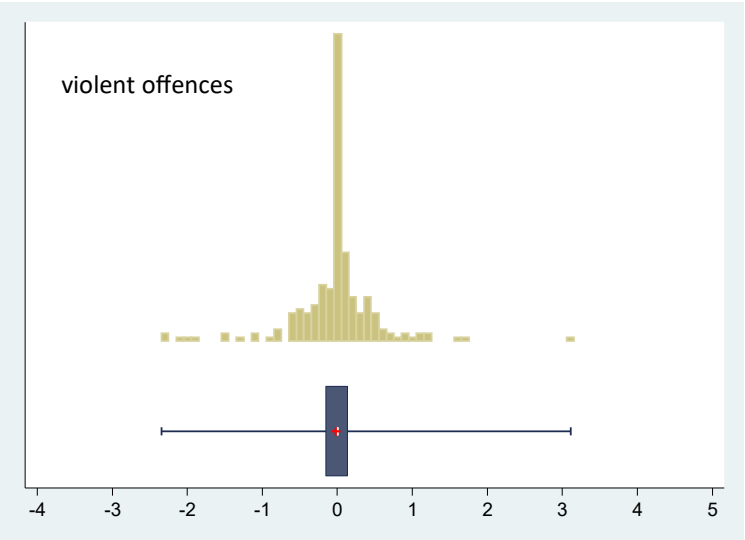
Figure 5. Distribution of effect sizes by outcome subdomain (offending and crime, administrative data sources)



Observation	827
Mean	0.0785
Std. Dev.	0.4722
Variance	0.2230
Quartile 1	-0.0930
Quartile 2 (Median)	0.0563
Quartile 3	0.2534
Range	6.3757
Skewness	0.7189
Kurtosis	16.2705

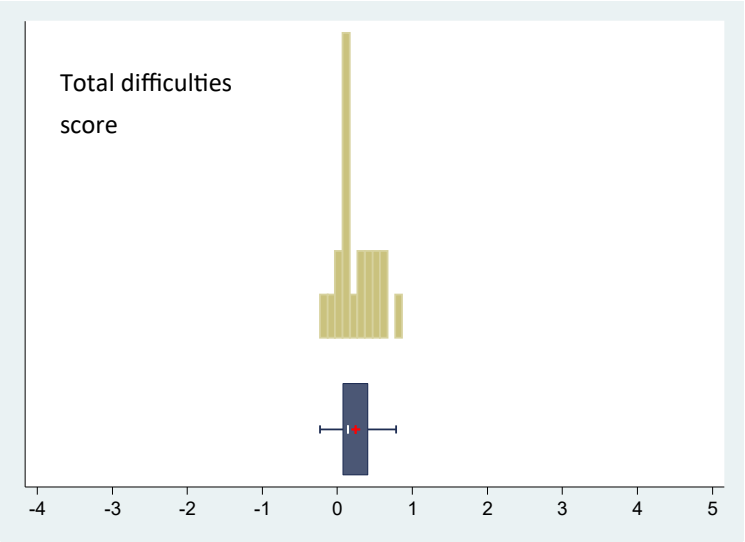


Observation	170
Mean	0.0229
Std. Dev.	0.4810
Variance	0.2314
Quartile 1	-0.2425
Quartile 2 (Median)	0.0081
Quartile 3	0.3005
Range	2.9780
Skewness	0.3950
Kurtosis	4.3288

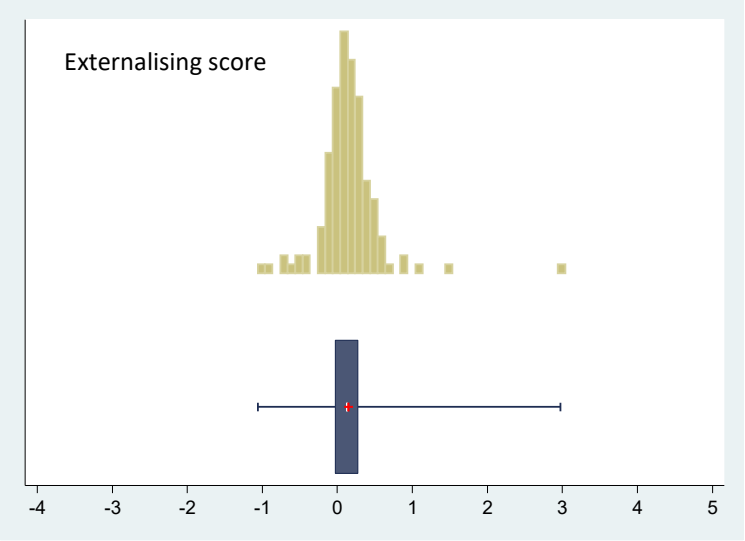


Observation	222
Mean	-0.0208
Std. Dev.	0.5693
Variance	0.3241
Quartile 1	-0.1518
Quartile 2 (Median)	0.0050
Quartile 3	0.1385
Range	5.4553
Skewness	-0.2306
Kurtosis	10.4929

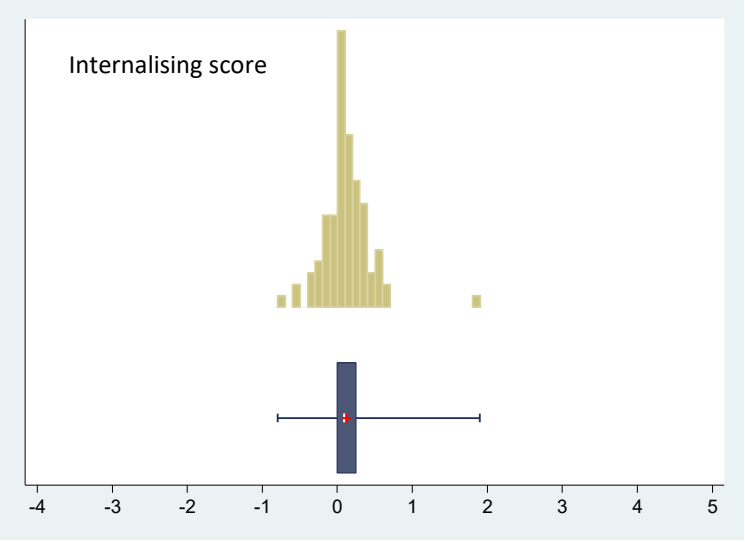
Figure 6. Distribution of effect sizes for SDQ



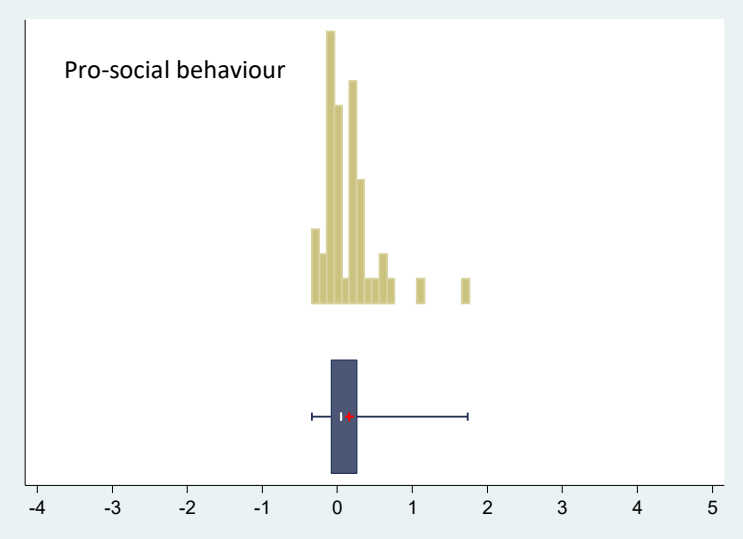
Observation	21
Mean	0.2427
Std. Dev.	0.2659
Variance	0.0707
Quartile 1	0.0768
Quartile 2 (Median)	0.1391
Quartile 3	0.4091
Range	1.0156
Skewness	0.4043
Kurtosis	2.3056



Observation	143
Mean	0.1504
Std. Dev.	0.4003
Variance	0.1602
Quartile 1	-0.0241
Quartile 2 (Median)	0.1299
Quartile 3	0.2797
Range	4.0341
Skewness	2.3068
Kurtosis	20.3703

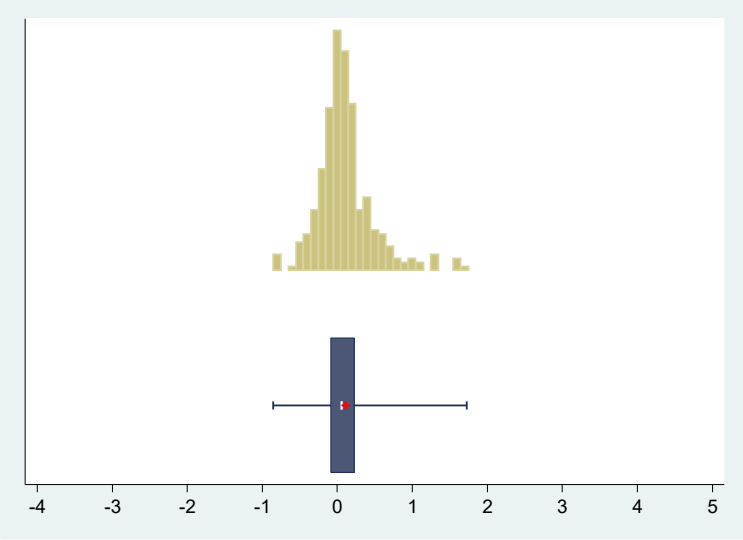


Observation	96
Mean	0.1224
Std. Dev.	0.3128
Variance	0.0978
Quartile 1	-0.0005
Quartile 2 (Median)	0.0959
Quartile 3	0.2559
Range	2.6927
Skewness	1.5652
Kurtosis	12.9458



Observation	50
Mean	0.1527
Std. Dev.	0.3572
Variance	0.1276
Quartile 1	-0.0643
Quartile 2 (Median)	0.0527
Quartile 3	0.2649
Range	2.0782
Skewness	2.1928
Kurtosis	9.7698

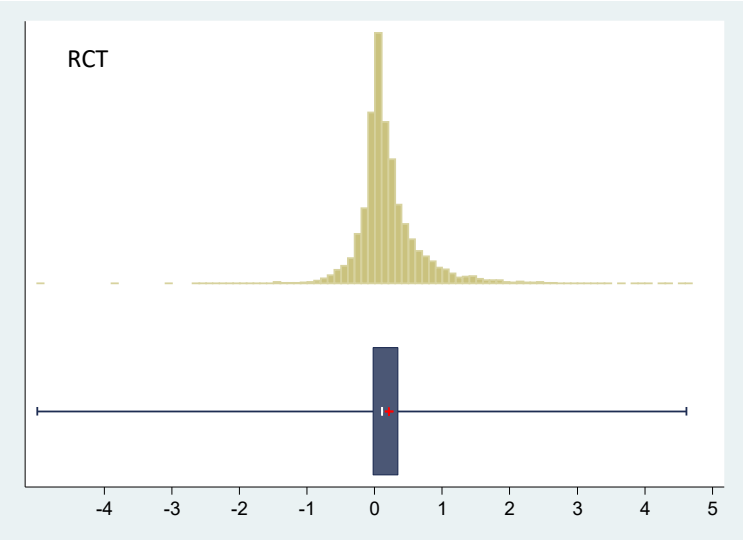
Figure 7. Distribution of effect sizes for self-reported delinquency



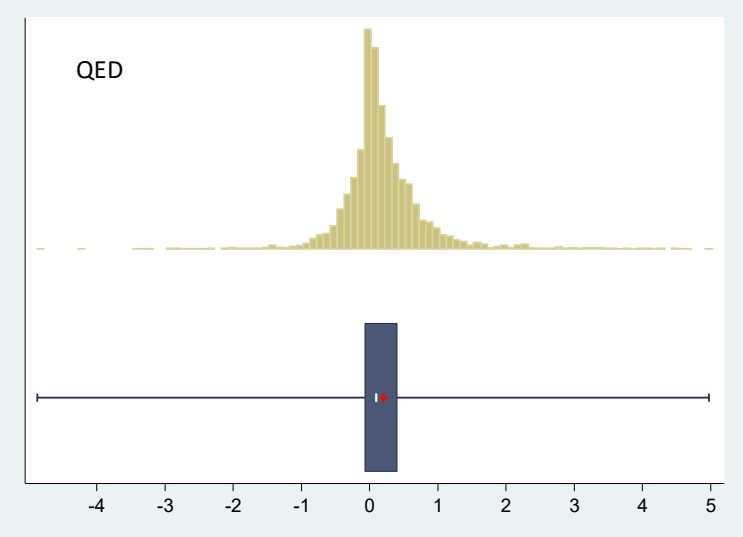
Observation	341
Mean	0.1102
Std. Dev.	0.3726
Variance	0.1388
Quartile 1	-0.0865
Quartile 2 (Median)	0.0587
Quartile 3	0.2310
Range	2.5775
Skewness	1.2826
Kurtosis	6.6316

Figure 8 shows the distribution of effect sizes for child-centred outcomes by study design (RCT and QED). The effect sizes are moderate: in RCT, mean is 0.21 (SD=0.51) and median is 0.11, and in QED mean is 0.19 (SD=0.67) and median is 0.10. The ranges are similar for both types of study: 9.62 for RCTs and 9.85 for QED. ES from RCTs have a more positively (to the right) skewed distribution than from QEDs. Both distributions have great tendency for outliers.

Figure 8. Distribution of effect sizes for child-centred outcomes by study design



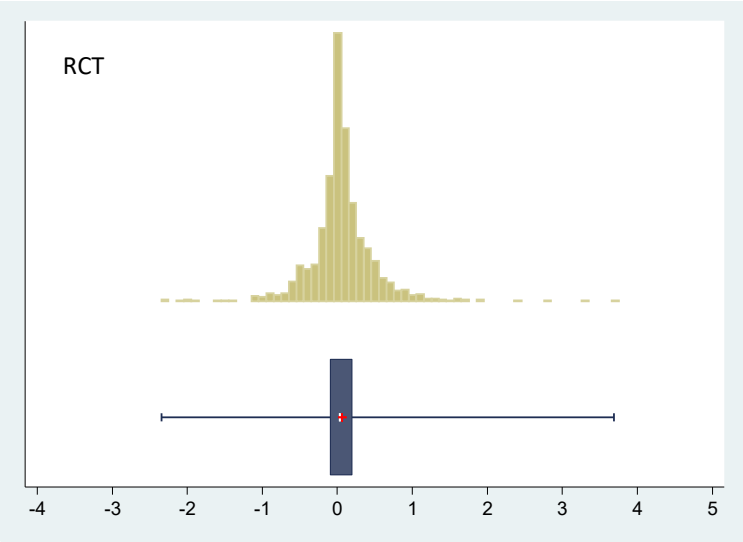
Observation	7,572
Mean	0.2087
Std. Dev.	0.5108
Variance	0.2609
Quartile 1	-0.0214
Quartile 2 (Median)	0.1094
Quartile 3	0.3519
Range	9.6158
Skewness	1.5829
Kurtosis	15.2397



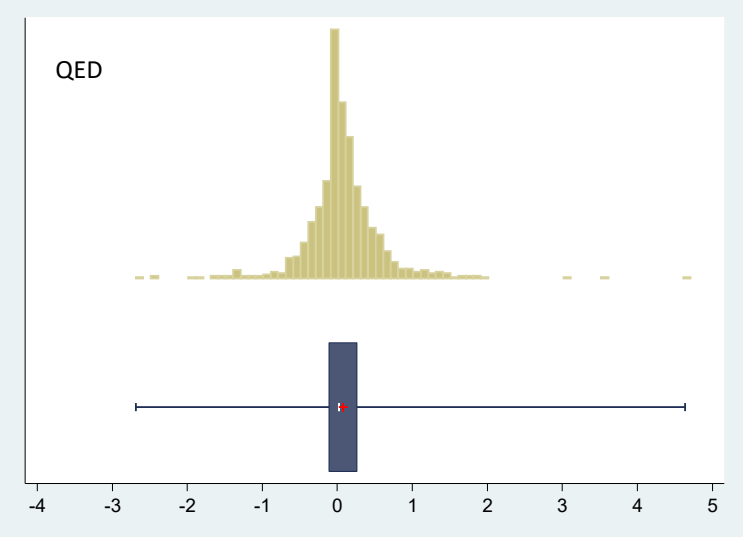
Observation	4,009
Mean	0.1944
Std. Dev.	0.6652
Variance	0.4424
Quartile 1	-0.0677
Quartile 2 (Median)	0.1025
Quartile 3	0.4003
Range	9.8517
Skewness	1.0913
Kurtosis	14.7127

Figure 9 shows the distribution of effect sizes for offending and crime outcomes by study design (RCT and QED). The effect sizes are small: in RCTs the mean is 0.06 (SD=0.44) and median 0.04; in QEDs the mean is 0.08 (SD=0.52) and median 0.02. The range is wider for QEDs (7.33) compared to RCTs (6.03). ES from QEDs have a more positively (to the right) skewed distribution than from RCT. Both distributions have great tendency for outliers.

Figure 9. Distribution of effect sizes for offending and crime outcomes by study design



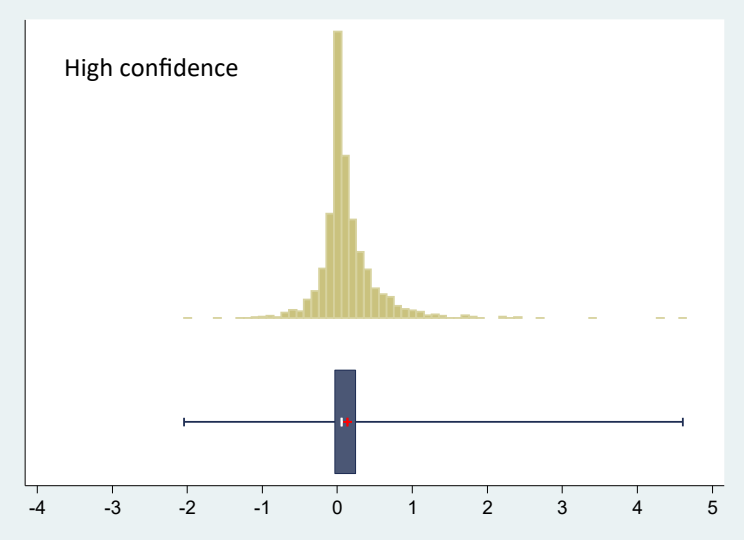
Observation	1,322
Mean	0.0636
Std. Dev.	0.4380
Variance	0.1919
Quartile 1	-0.0895
Quartile 2 (Median)	0.0359
Quartile 3	0.2113
Range	6.0328
Skewness	0.7515
Kurtosis	13.7459



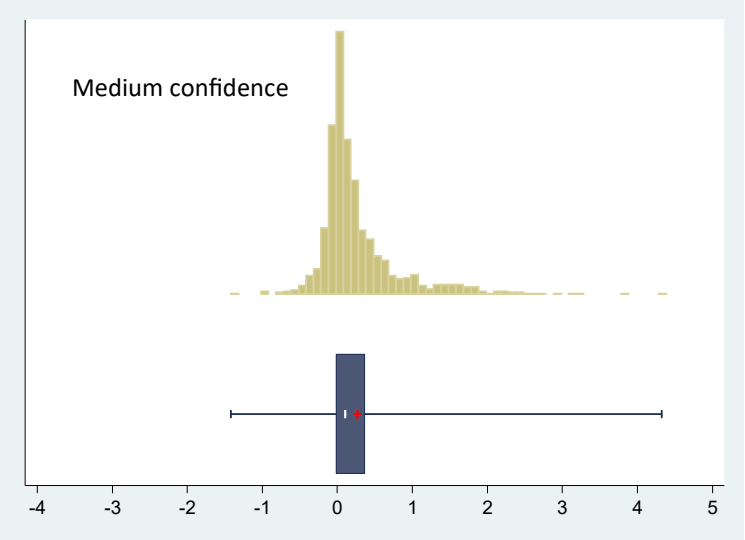
Observation	937
Mean	0.0753
Std. Dev.	0.5177
Variance	0.2680
Quartile 1	-0.1088
Quartile 2 (Median)	0.0240
Quartile 3	0.2660
Range	7.3261
Skewness	0.9419
Kurtosis	15.5047

Figure 10 describes the distribution of effect sizes (outcome domain combined) by study confidence level. Only the mean of the high confidence studies is high (M=0.26, SD=0.56) and median of 0.10, while that of lower confidence studies is moderate. For the medium confidence studies, the mean is 0.14 (SD=0.42) and median is 0.06. For low confidence studies, mean is 0.19 (SD=0.58) and median is 0.10. The range for low confidence is wider (9,97) and for high confidence it is lower (5.74). Distributions of studies of all confidence levels are right-skewed and contain considerable numbers of outliers.

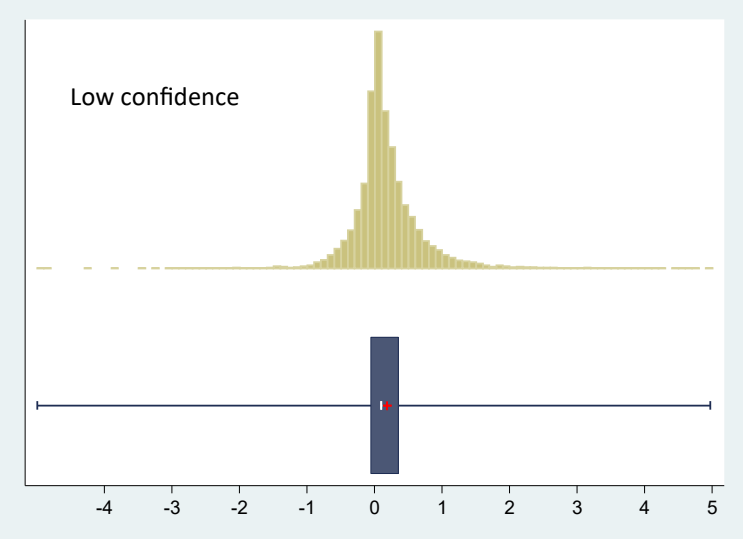
Figure 10. Distribution of effect sizes by study confidence



Observation	1,143
Mean	0.2584
Std. Dev.	0.5556
Variance	0.3087
Quartile 1	-0.0230
Quartile 2 (Median)	0.0970
Quartile 3	0.3642
Range	5.7421
Skewness	2.1955
Kurtosis	10.9833



Observation	1,607
Mean	0.1361
Std. Dev.	0.4205
Variance	0.1768
Quartile 1	-0.0304
Quartile 2 (Median)	0.0555
Quartile 3	0.2387
Range	6.6504
Skewness	2.7179
Kurtosis	23.7487

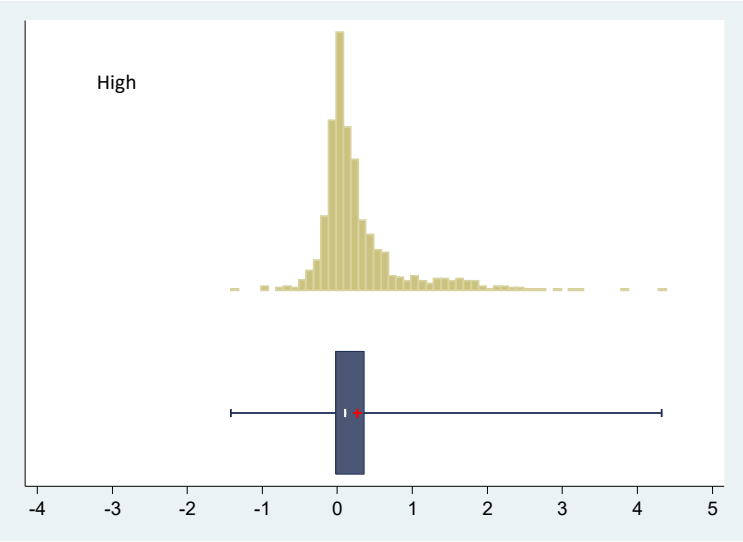


Observation	11,007
Mean	0.1850
Std. Dev.	0.5779
Variance	0.3340
Quartile 1	-0.0513
Quartile 2 (Median)	0.1020
Quartile 3	0.3629
Range	9.9725
Skewness	1.1522
Kurtosis	15.4779

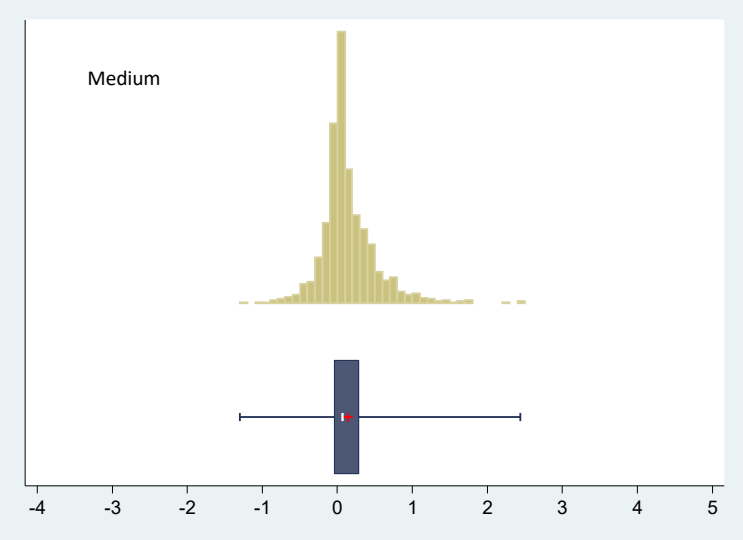
We further analysed distributions for study confidence by study design (Figure 11). The analysis shows that, regardless of study design, studies rated at high confidence tended to present larger ES (M=0.27 for RCT and 0.20 for QED) than medium or low confidence studies, while low confidence studies tended to have higher mean ES (M=0.18 for both RCT and QED) than medium confidence (M=0.14 for RCT and M=0.12 for QED). Within the same confidence level, RCTs tended to have higher effect sizes than QEDs.

Figure 11 Distribution of effect sizes by study confidence and study design

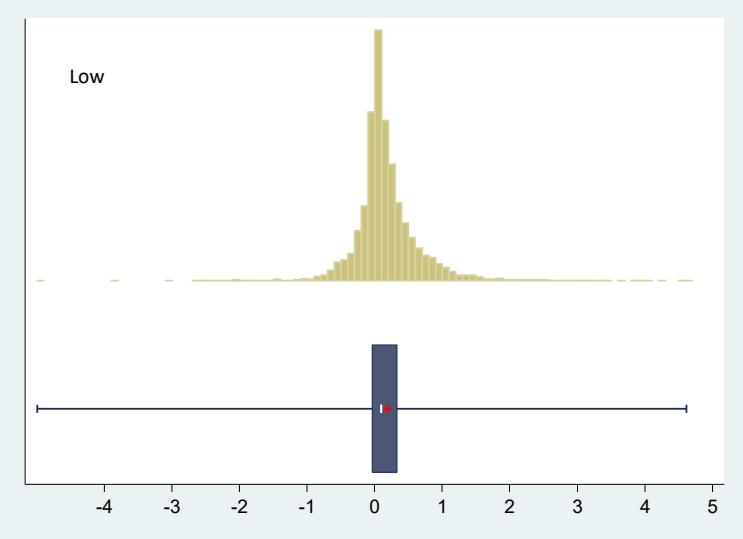
a) RCTs by confidence level



Observation	985
Mean	0.2671
Std. Dev.	0.5621
Variance	0.3160
Quartile 1	-0.0174
Quartile 2 (Median)	0.1078
Quartile 3	0.3540
Range	5.7421
Skewness	2.4071
Kurtosis	11.6252

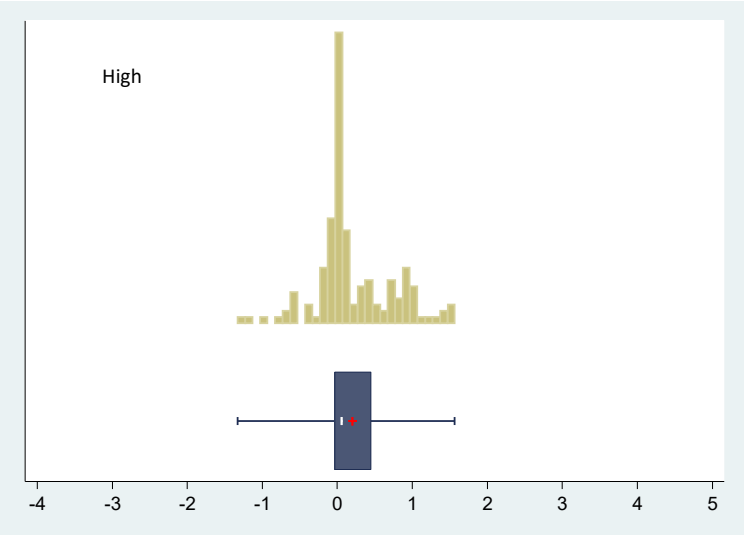


Observation	1,043
Mean	0.1435
Std. Dev.	0.3691
Variance	0.1362
Quartile 1	-0.0368
Quartile 2 (Median)	0.0701
Quartile 3	0.2895
Range	3.7377
Skewness	1.4184
Kurtosis	8.8829

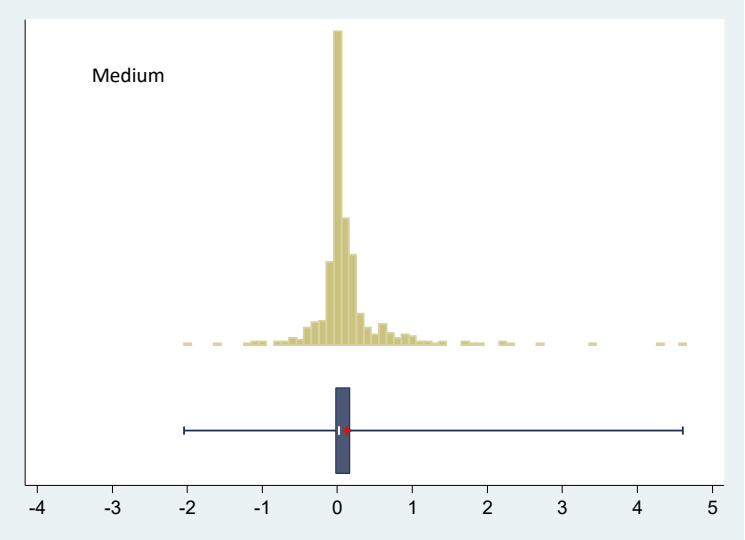


Observation	6,749
Mean	0.1840
Std. Dev.	0.5131
Variance	0.2633
Quartile 1	-0.0328
Quartile 2 (Median)	0.1009
Quartile 3	0.3387
Range	9.6158
Skewness	1.2926
Kurtosis	15.4160

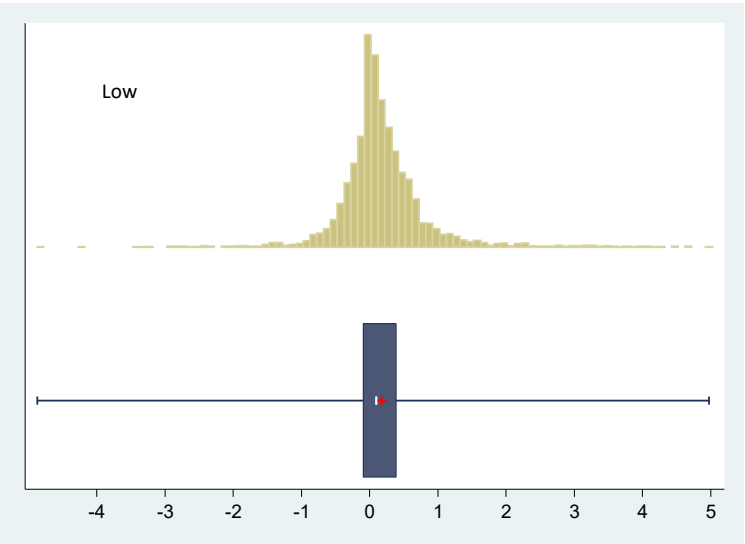
b) QEDs by confidence level



Observation	158
Mean	0.2041
Std. Dev.	0.5117
Variance	0.2618
Quartile 1	-0.0313
Quartile 2 (Median)	0.0554
Quartile 3	0.4548
Range	2.8906
Skewness	0.3822
Kurtosis	3.7152

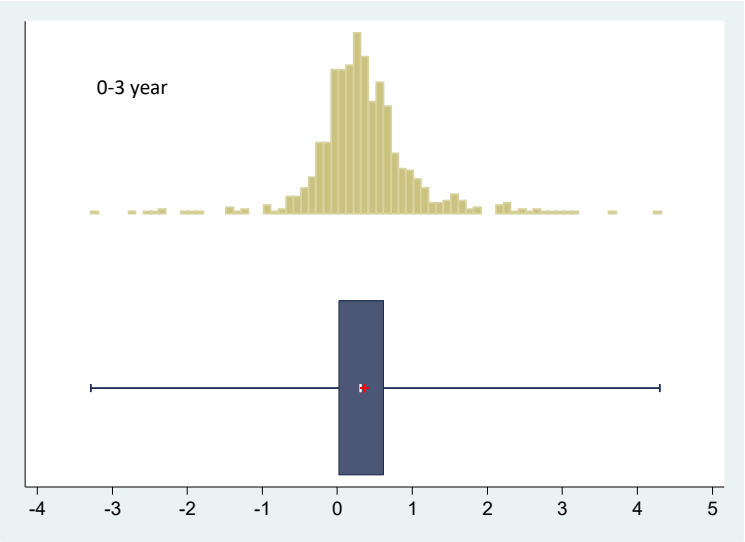


Observation	564
Mean	0.1224
Std. Dev.	0.5020
Variance	0.2520
Quartile 1	-0.0228
Quartile 2 (Median)	0.0297
Quartile 3	0.1759
Range	6.6504
Skewness	3.5539
Kurtosis	28.9118

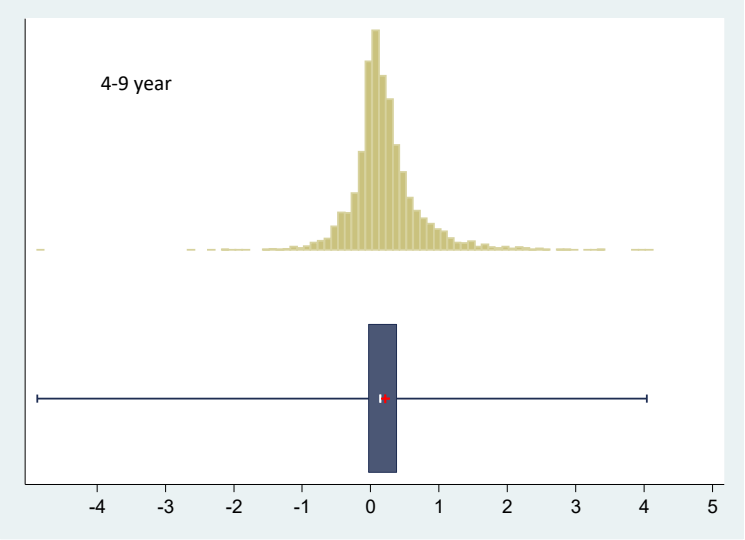


Observation	4,173
Mean	0.1782
Std. Dev.	0.6619
Variance	0.4381
Quartile 1	-0.0894
Quartile 2 (Median)	0.1033
Quartile 3	0.3954
Range	9.8517
Skewness	0.9825
Kurtosis	14.4761

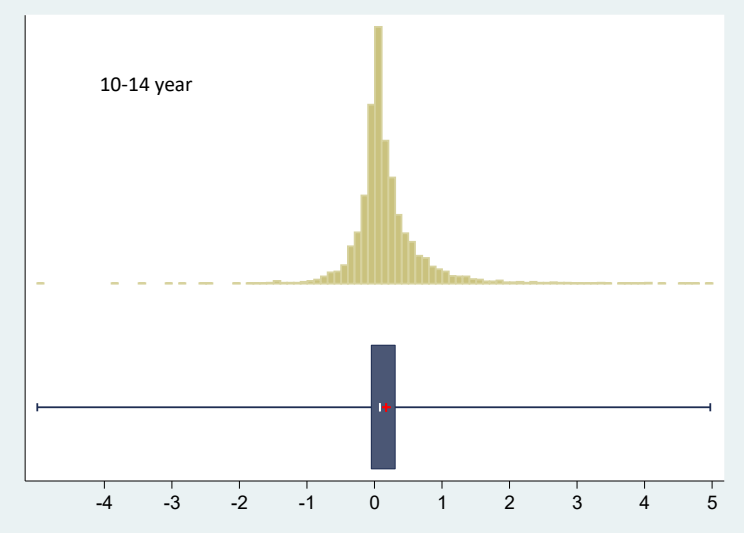
Figure 12. Distribution of effect sizes for child-centre outcomes by age group



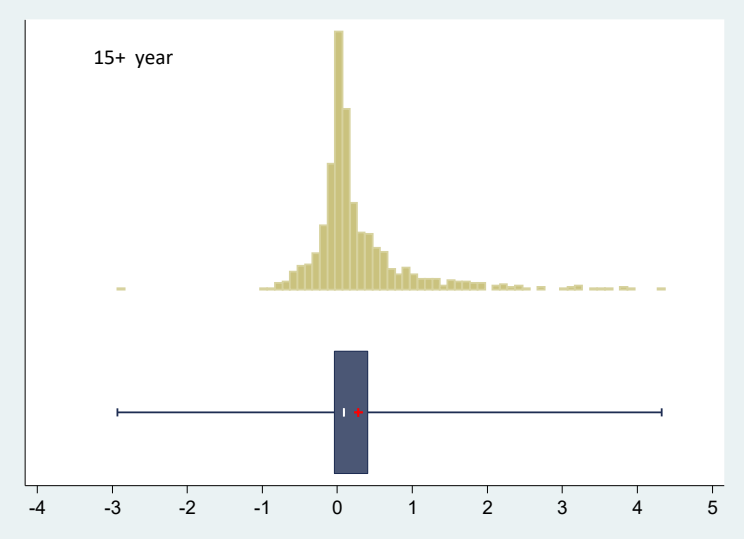
Observation	847
Mean	0.3593
Std. Dev.	0.6710
Variance	0.4503
Quartile 1	0.0293
Quartile 2 (Median)	0.3078
Quartile 3	0.6182
Range	7.5855
Skewness	0.4319
Kurtosis	9.4280



Observation	3,221
Mean	0.2124
Std. Dev.	0.5342
Variance	0.2853
Quartile 1	-0.0286
Quartile 2 (Median)	0.1393
Quartile 3	0.3879
Range	8.9147
Skewness	1.0112
Kurtosis	12.3032

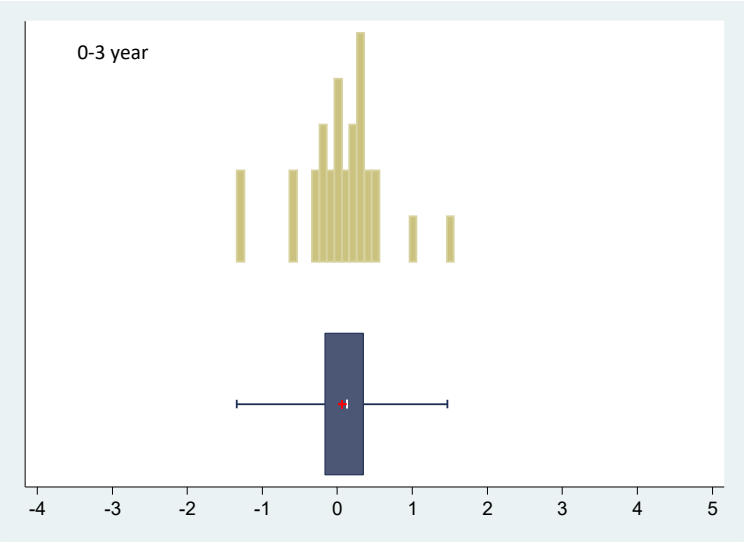


Observation	5,416
Mean	0.1751
Std. Dev.	0.5512
Variance	0.3038
Quartile 1	-0.0470
Quartile 2 (Median)	0.0805
Quartile 3	0.3130
Range	9.9725
Skewness	1.7004
Kurtosis	20.5460

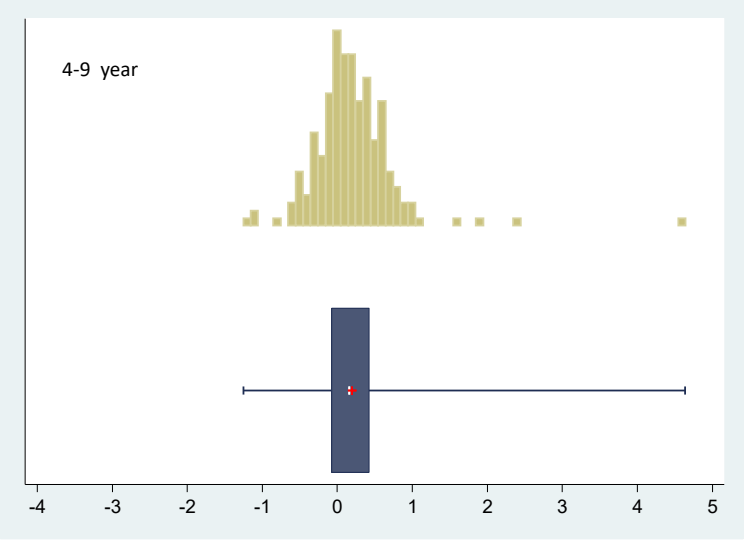


Observation	879
Mean	0.2803
Std. Dev.	0.6794
Variance	0.4616
Quartile 1	-0.0369
Quartile 2 (Median)	0.0892
Quartile 3	0.4087
Range	7.2577
Skewness	2.2508
Kurtosis	11.1735

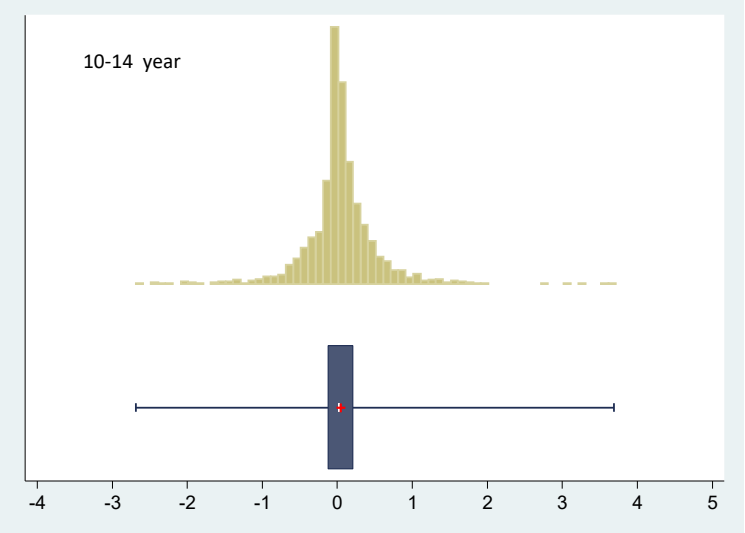
Figure 13. Distribution of effect sizes for offending and crime outcomes by age group



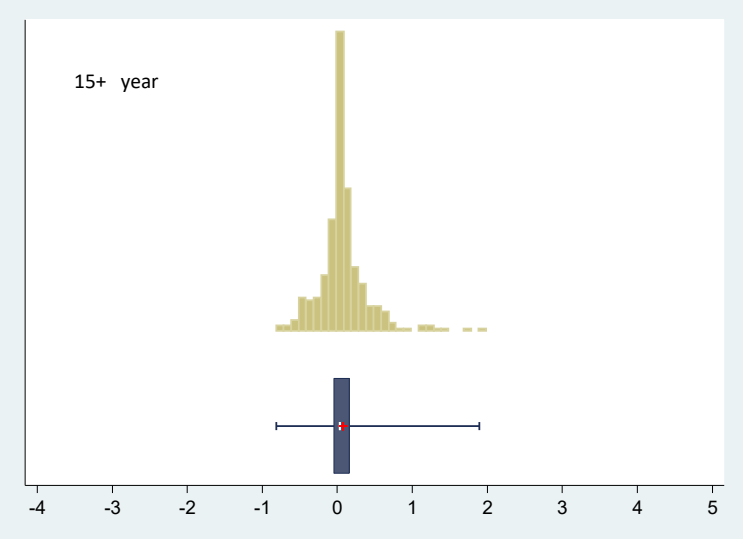
Observation	32
Mean	0.0805
Std. Dev.	0.5507
Variance	0.3033
Quartile 1	-0.1571
Quartile 2 (Median)	0.1330
Quartile 3	0.3570
Range	2.8046
Skewness	-0.4428
Kurtosis	4.6329



Observation	212
Mean	0.1974
Std. Dev.	0.5394
Variance	0.2909
Quartile 1	-0.0702
Quartile 2 (Median)	0.1622
Quartile 3	0.4252
Range	5.8930
Skewness	2.9149
Kurtosis	24.9215



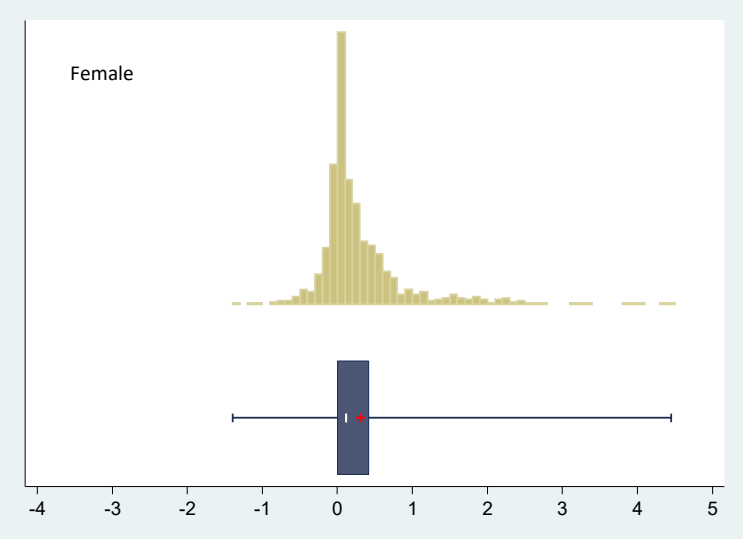
Observation	1,411
Mean	0.0458
Std. Dev.	0.5027
Variance	0.2527
Quartile 1	-0.1203
Quartile 2 (Median)	0.0190
Quartile 3	0.2160
Range	6.3757
Skewness	0.4544
Kurtosis	12.6385



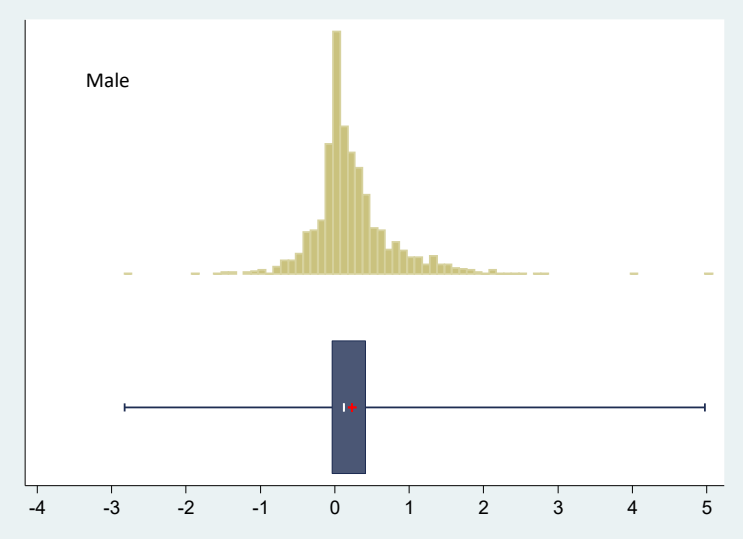
Observation	347
Mean	0.0797
Std. Dev.	0.3354
Variance	0.1125
Quartile 1	-0.0451
Quartile 2 (Median)	0.0375
Quartile 3	0.1704
Range	2.7072
Skewness	1.4027
Kurtosis	8.5388

Figure 14 shows the distribution of effect sizes for child-centred outcomes by study participant sex. In this analysis, we combined ES coded as gender subgroups from studies containing both types of participants with ES coded in studies that targeted female or male population groups only. The average ES is high for female subgroup (M=0.30, SD=0.62, median=0.12) while for male the impact is moderate (M=0.23, SD=0.55, median=0.12). The range of the ES is 5.85 for females, and 7.80 for males. The distribution for females is more skewed to the right than for males. Both distributions have a great tendency for outliers with high values of kurtosis.

Figure 14. Distribution of effect sizes for child-centred outcomes by gender



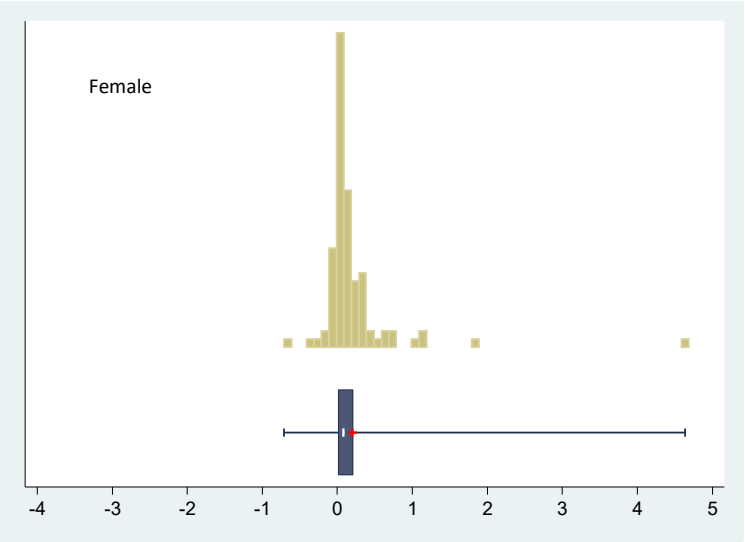
Observation	923
Mean	0.3001
Std. Dev.	0.6181
Variance	0.3821
Quartile 1	0.0055
Quartile 2 (Median)	0.1169
Quartile 3	0.4237
Range	5.8460
Skewness	2.6375
Kurtosis	13.5186



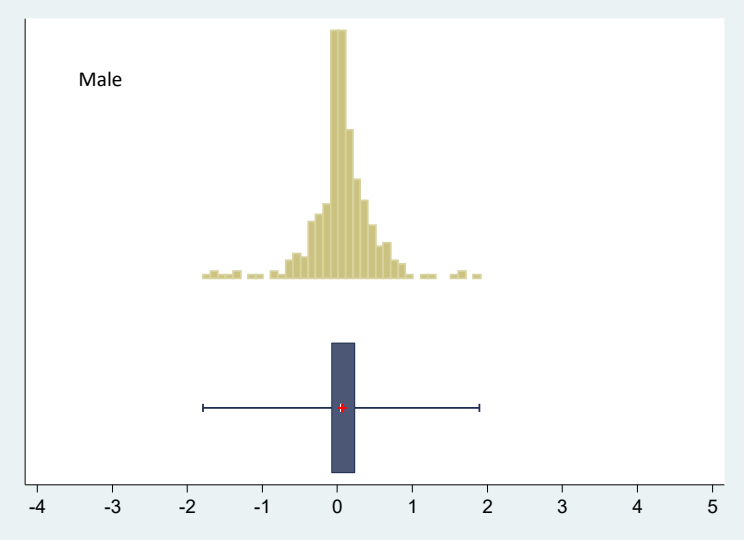
Observation	1,259
Mean	0.2298
Std. Dev.	0.5536
Variance	0.3065
Quartile 1	-0.0341
Quartile 2 (Median)	0.1184
Quartile 3	0.4222
Range	7.8017
Skewness	1.4261
Kurtosis	11.5648

Figure 15 shows the distribution of effect sizes for offending and crime outcome by gender. The effect sizes are smaller than for child-centred outcomes in general, but again ES is greater for females (M=0.20 (moderate), SD=0.54, median=0.08), than for males (M=0.06 (low), SD=0.42, median=0.05). The range of the ES is 5.35 for females, and 3.68 for males. The distribution for females is skewed to the right. Both distributions have a great tendency for outliers with high values of kurtosis.

Figure 15. Distribution of effect sizes for offending and crime outcomes by sex



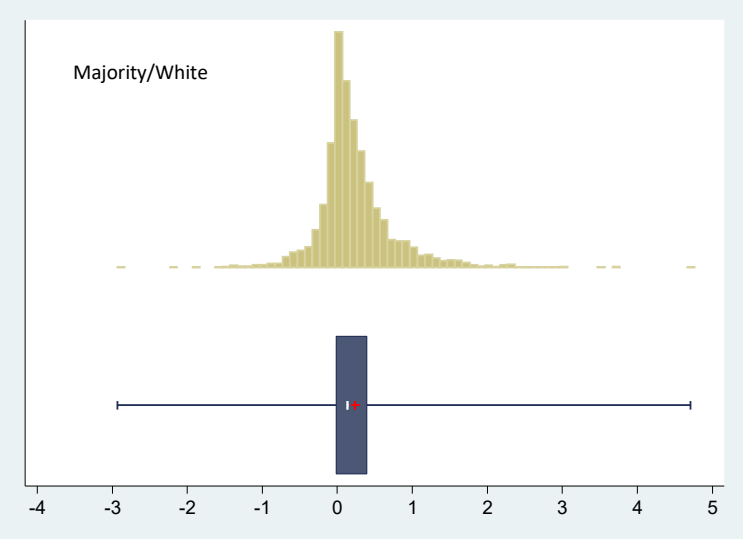
Observation	102
Mean	0.2016
Std. Dev.	0.5423
Variance	0.2941
Quartile 1	0.0164
Quartile 2 (Median)	0.0765
Quartile 3	0.2129
Range	5.3492
Skewness	5.8247
Kurtosis	46.1089



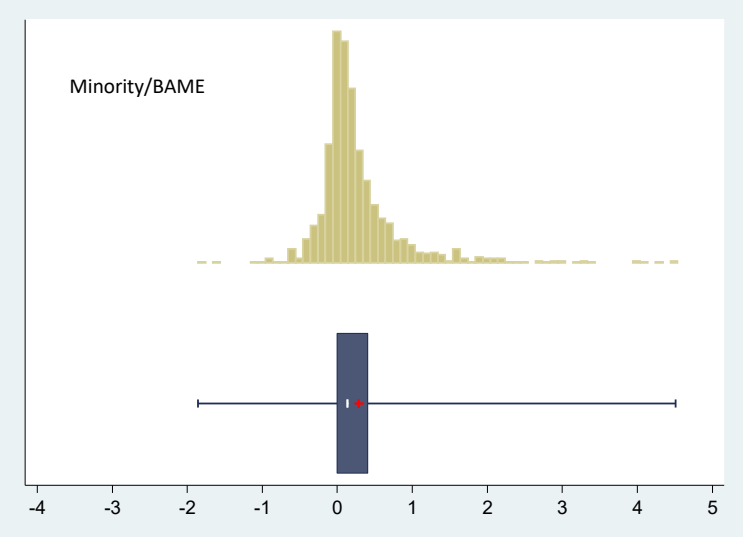
Observation	366
Mean	0.0631
Std. Dev.	0.4220
Variance	0.1781
Quartile 1	-0.0759
Quartile 2 (Median)	0.0493
Quartile 3	0.2380
Range	3.6802
Skewness	-0.3067
Kurtosis	7.9579

Figure 16 provides the distribution of effect sizes for child-centred outcome by ethnicity. In this analysis, we combined ES coded as ethnicity subgroups with studies targeted particular ethnic groups only. The mean ES is high for BAME children (M=0.28, SD=0.61, median=0.14), while the average impact is moderate for majority ethnic/White children and youths (M=0.23, SD=0.54, median=0.14). The range of the ES is 9.51 for majority ethic/White, and 6.37 for minority ethnic/BAME. The distribution for minority ethnic/BAME children and youth is more skewed to the right than for majority ethnic/White. Both distributions have a great tendency for outliers with high values of kurtosis.

Figure 16. Distribution of effect sizes for child-centre outcome by ethnicity



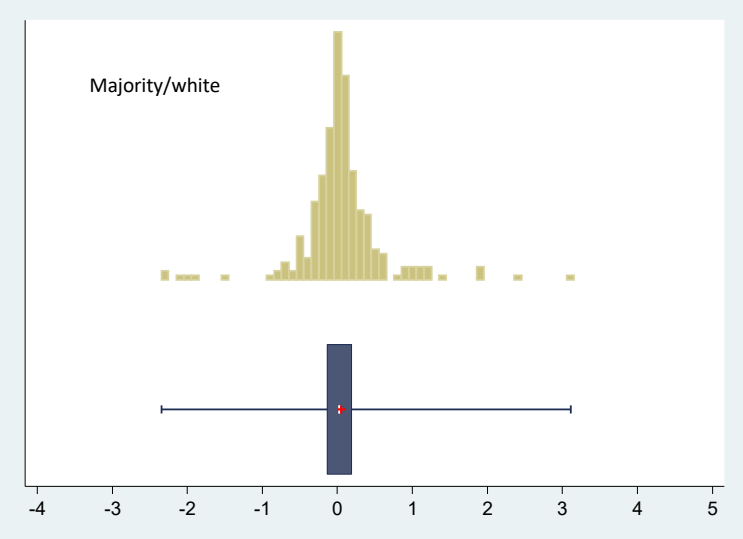
Observation	2,481
Mean	0.2329
Std. Dev.	0.5365
Variance	0.2878
Quartile 1	-0.0156
Quartile 2 (Median)	0.1385
Quartile 3	0.3977
Range	9.5060
Skewness	1.1385
Kurtosis	14.0763



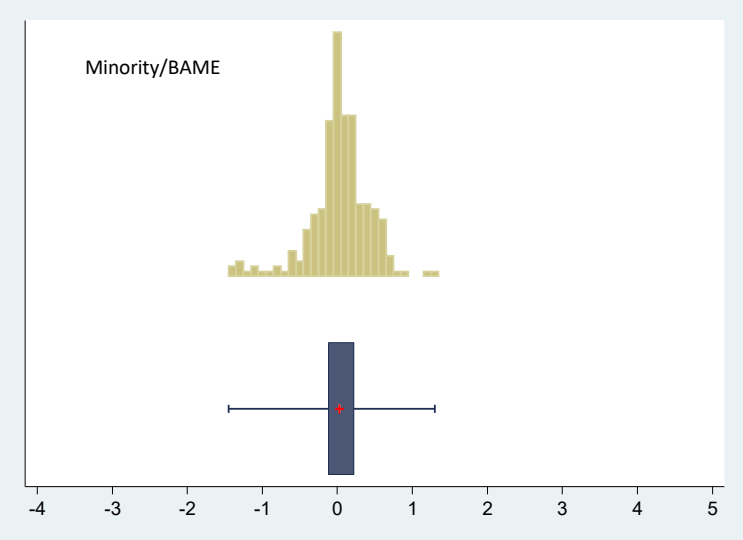
Observation	1,171
Mean	0.2829
Std. Dev.	0.6136
Variance	0.3764
Quartile 1	-0.0029
Quartile 2 (Median)	0.1366
Quartile 3	0.4049
Range	6.3699
Skewness	2.7589
Kurtosis	15.2130

Figure 17 shows the distribution of effect sizes for offending and crime outcome by ethnicity. The mean ES for majority ethnic/White is 0.05 (SD=0.54, median=0.03) and for minority ethnic/BAME the mean ES is 0.03 (SD=0.40, median=0.03), suggesting there are no impact on minority ethnic/BAME and low effect on majority ethnic/White for this outcome domain. The range of the ES is 5.46 for majority ethnic/White, and 2.75 for minority ethnic/BAME. The distribution for minority ethnic/BAME is moderately skewed to the left, while that of majority ethnic/White is approximately symmetrical. Both distributions have a great tendency for outliers with high values of kurtosis.

Figure 17. Distribution of effect sizes for offending and crime outcome by ethnicity



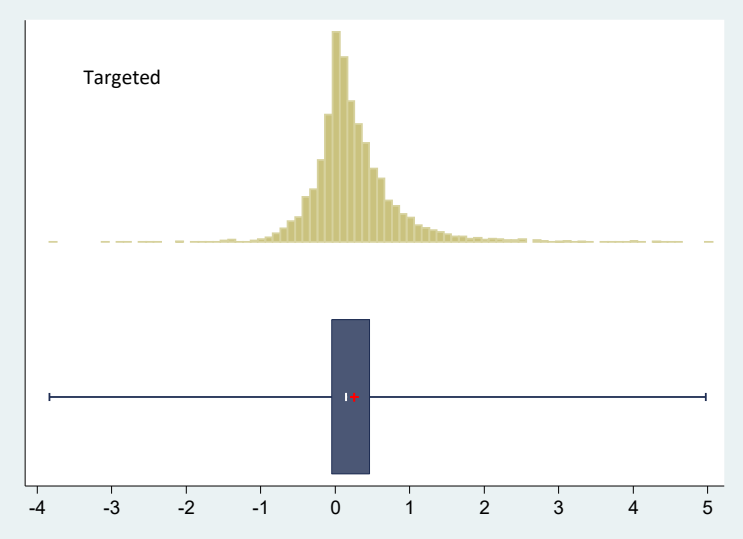
Observation	299
Mean	0.0536
Std. Dev.	0.5382
Variance	0.2897
Quartile 1	-0.1355
Quartile 2 (Median)	0.0300
Quartile 3	0.1980
Range	5.4553
Skewness	0.3086
Kurtosis	11.7874



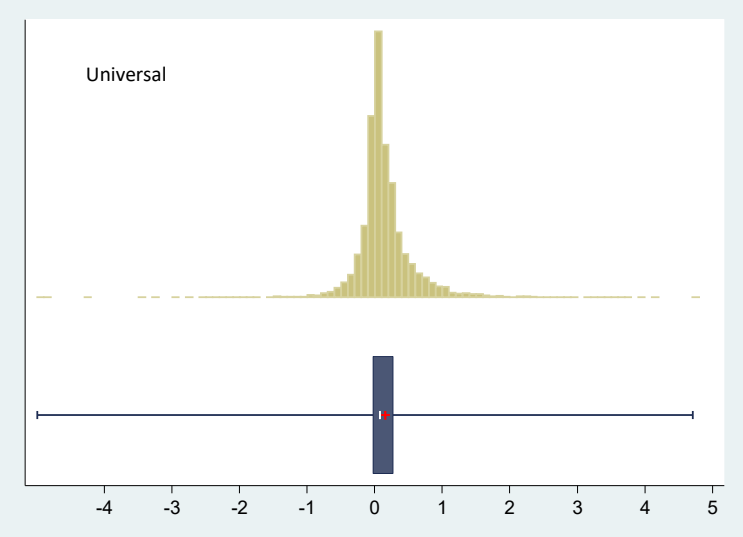
Observation	254
Mean	0.0301
Std. Dev.	0.4026
Variance	0.1621
Quartile 1	-0.1160
Quartile 2 (Median)	0.0257
Quartile 3	0.2274
Range	2.7474
Skewness	-0.8342
Kurtosis	5.5928

Figure 18 compares the distributions of effect sizes from targeted intervention against universal intervention, for child-centred outcomes. Targeted interventions have a high mean ES (M=0.25, SD=0.63, median 0.15), compared to universal interventions, which show moderate impact (M=0.16, SD=0.49, median= 0.08). The range of the ES is high, 8.81 for targeted interventions, and 9.71 for universal interventions. The distribution for targeted intervention is highly skewed to the right, while that of universal intervention has a moderate right skewness. Both distributions have a great tendency for outliers with high values of kurtosis.

Figure 18. Distribution of effect sizes for child-centred outcomes, targeted vs universal intervention



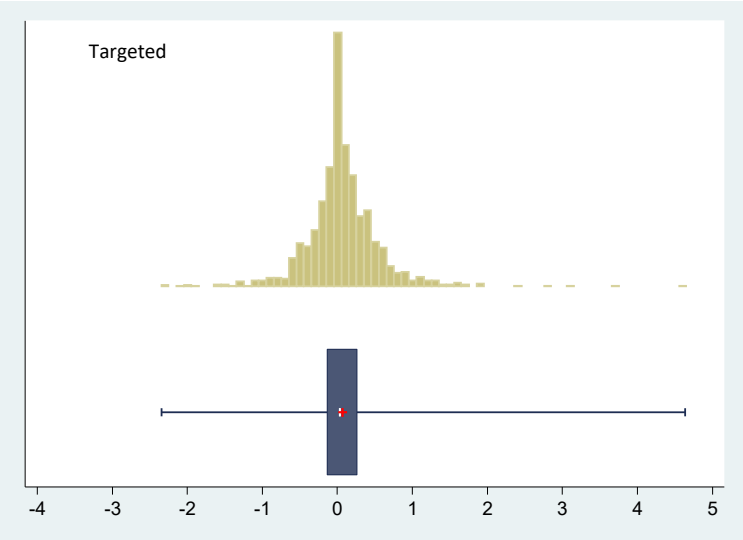
Observation	5,760
Mean	0.2519
Std. Dev.	0.6308
Variance	0.3979
Quartile 1	-0.0486
Quartile 2 (Median)	0.1471
Quartile 3	0.4605
Range	8.8135
Skewness	1.5486
Kurtosis	12.2986



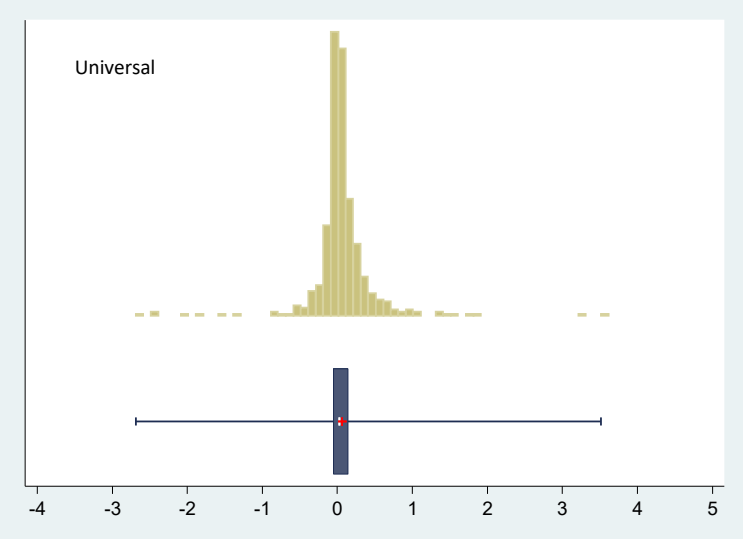
Observation	5,830
Mean	0.1578
Std. Dev.	0.4936
Variance	0.2437
Quartile 1	-0.0249
Quartile 2 (Median)	0.0829
Quartile 3	0.2767
Range	9.7074
Skewness	0.7118
Kurtosis	22.4447

Figure 19 compares distributions of effect sizes from targeted intervention against universal intervention, for offending and crime outcomes. The ES are small for both: the mean is 0.07 for targeted interventions (SD=0.48, median 0.03) and 0.06 for universal interventions (SD=0.43, median= 0.03). The range of the ES is 6.98 for targeted interventions, and 6.20 for universal interventions. Both distributions are moderately skewed to the right. Both distributions have a great tendency for outliers with high values of kurtosis.

Figure 19. Distribution of effect sizes for offending and crime outcomes, targeted vs universal interventions



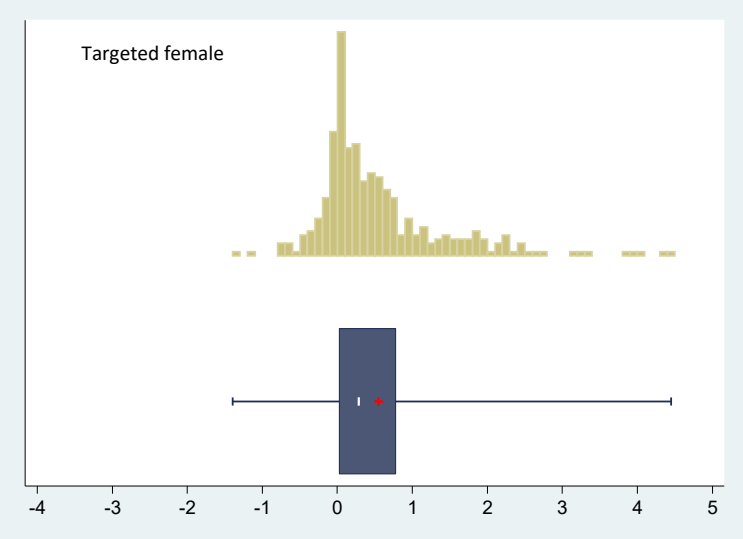
Observation	1,742
Mean	0.0695
Std. Dev.	0.4837
Variance	0.2339
Quartile 1	-0.1326
Quartile 2 (Median)	0.0336
Quartile 3	0.2690
Range	6.9832
Skewness	0.9127
Kurtosis	13.1160



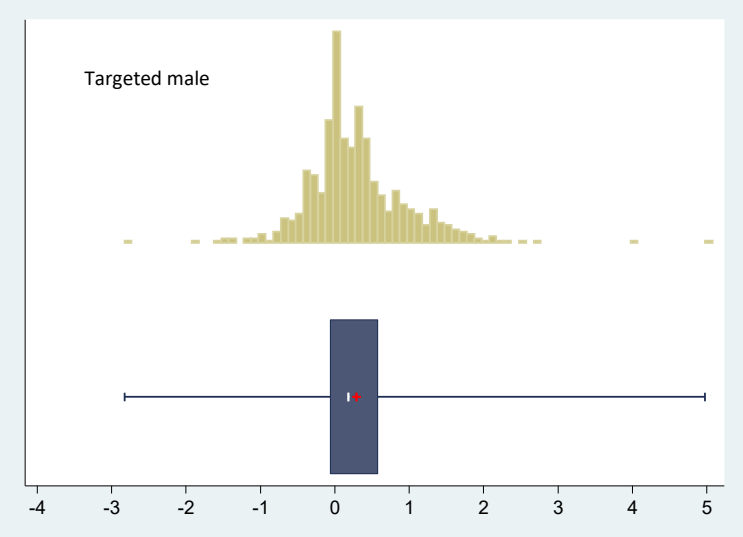
Observation	517
Mean	0.0649
Std. Dev.	0.4338
Variance	0.1882
Quartile 1	-0.0502
Quartile 2 (Median)	0.0280
Quartile 3	0.1458
Range	6.2001
Skewness	0.6447
Kurtosis	25.3198

Figure 20 compares distributions of effect sizes from gender targeted intervention against non-gender targeted intervention, for child-centre outcomes. Overall, the mean ES for females/males from female-/male-targeted interventions are higher than those from non-gender targeted interventions: the mean ES for female participants of female-targeted intervention is very large (M=0.55, SD=0.85, median=0.29); for their male equivalents, the mean is also large 0.29 (SD=0.68, median=0.18). The ES for female/male participants of non-gender targeted interventions are on average 0.14 (SD=0.33, median 0.08) and 0.15 (SD=0.34, median=0.09), respectively, which are both moderate impacts. The range of the ES is wider for gender-targeted interventions than for non-targeted ones: 5.85 for female-targeted, 7.80 for male- targeted, 3.36 for females participating in non-targeted interventions and 3.90 for non-targeted males. All the distributions are skewed to the right and have noticeable outliers.

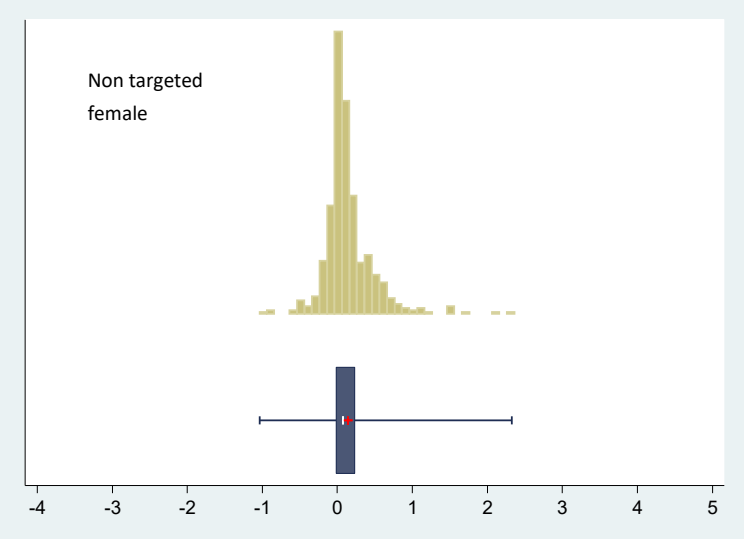
Figure 20. Distribution of effect sizes for child-centred outcomes, gender-targeted vs non-targeted interventions



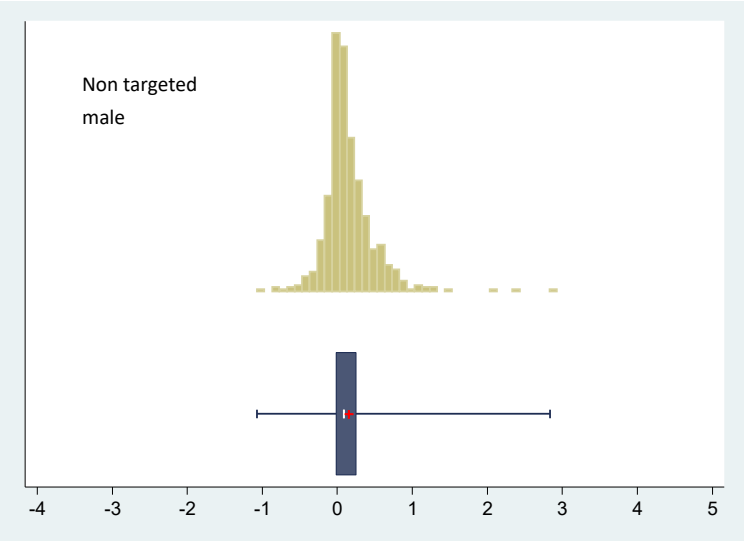
Observation	352
Mean	0.5512
Std. Dev.	0.8510
Variance	0.7242
Quartile 1	0.0279
Quartile 2 (Median)	0.2857
Quartile 3	0.7851
Range	5.8460
Skewness	1.7283
Kurtosis	6.9763



Observation	683
Mean	0.2940
Std. Dev.	0.6763
Variance	0.4574
Quartile 1	-0.0604
Quartile 2 (Median)	0.1823
Quartile 3	0.5751
Range	7.8017
Skewness	1.0107
Kurtosis	8.4148



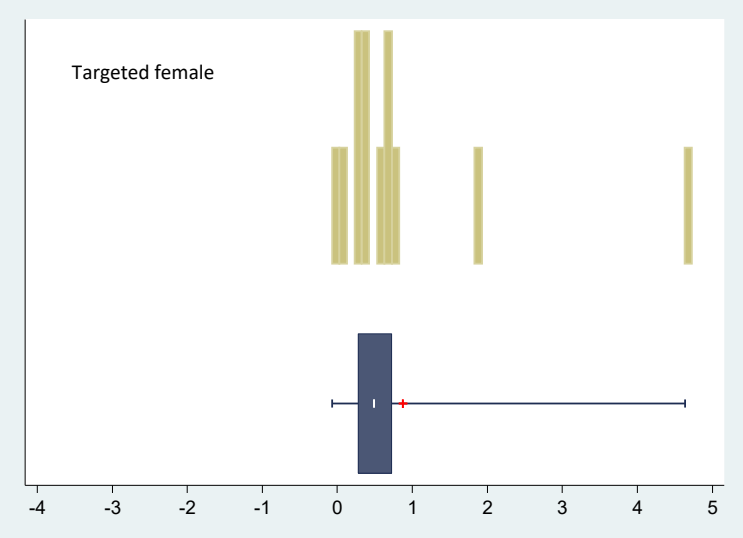
Observation	560
Mean	0.1416
Std. Dev.	0.3320
Variance	0.1102
Quartile 1	-0.0140
Quartile 2 (Median)	0.0769
Quartile 3	0.2455
Range	3.3649
Skewness	1.7202
Kurtosis	11.3570



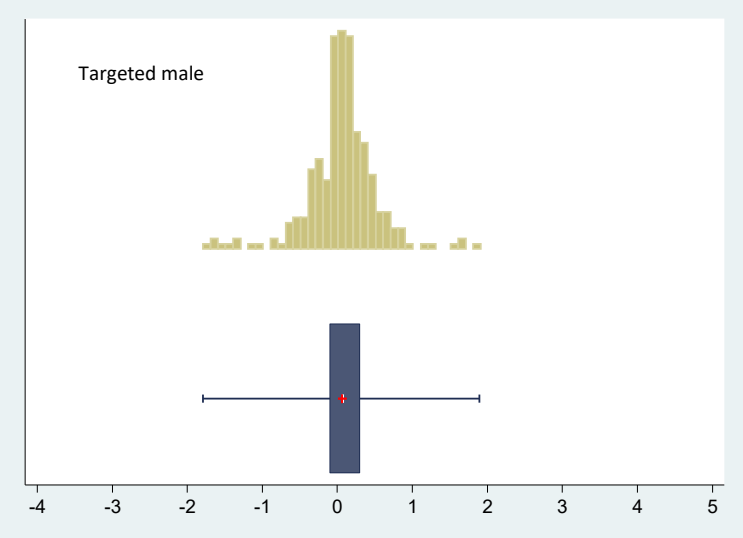
Observation	569
Mean	0.1537
Std. Dev.	0.3410
Variance	0.1163
Quartile 1	-0.0174
Quartile 2 (Median)	0.0892
Quartile 3	0.2630
Range	3.9039
Skewness	2.0987
Kurtosis	15.0938

Figure 21 compares distributions of effect sizes from gender-targeted interventions against non-gender targeted intervention, for offending and crime outcomes. Overall, the mean effect sizes for females/males from female-/male-targeted interventions are higher than those from non-gender targeted interventions: the mean ES for female participants of female-targeted interventions (N=12) is very high (M=0.88, SD=1.28, median=0.49), whereas for male equivalents, the mean is low with the value of 0.06 (SD=0.47, median=0.07). The mean ES for female and male participants of non-gender targeted intervention is moderate and low: 0.11 (SD=0.25, median 0.07) and 0.06 (SD=0.22, median=0.02), respectively. The range of the ES is wider for gender-targeted interventions than for non-targeted ones: 4.71 for female-targeted, 3.68 for male-targeted, 1.81 for females in non-targeted interventions and 1.26 for non-targeted males. The distributions are mostly skewed to the right (except for male- targeted) and have potential noticeable outliers.

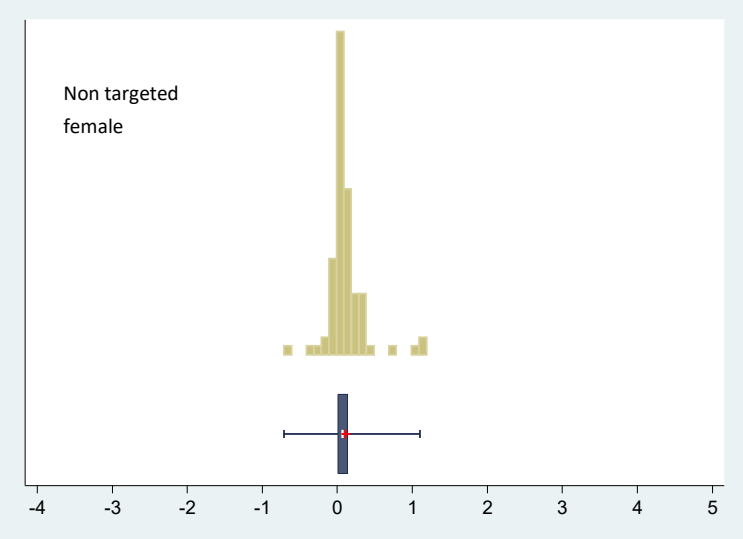
Figure 21. Distribution of effect sizes for offending and crime outcomes, gender-targeted vs non-targeted interventions



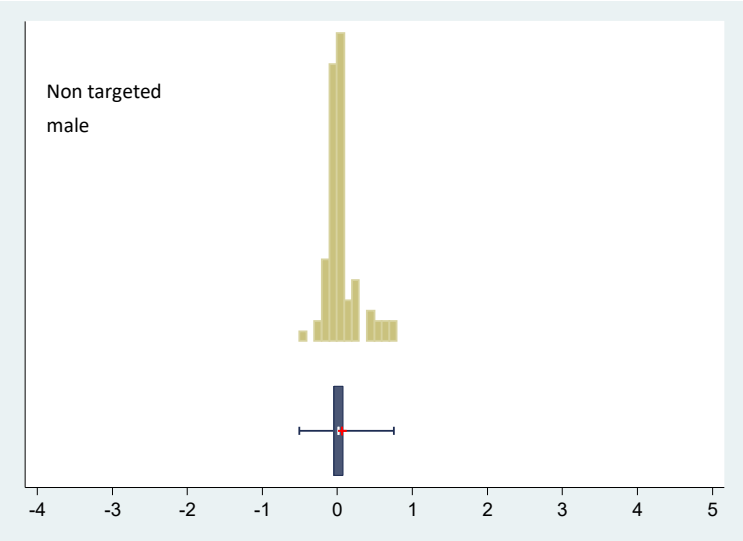
Observation	12
Mean	0.8758
Std. Dev.	1.2842
Variance	1.6492
Quartile 1	0.2783
Quartile 2 (Median)	0.4901
Quartile 3	0.7237
Range	4.7095
Skewness	2.3406
Kurtosis	7.4120



Observation	280
Mean	0.0638
Std. Dev.	0.4675
Variance	0.2186
Quartile 1	-0.0998
Quartile 2 (Median)	0.0748
Quartile 3	0.3015
Range	3.6802
Skewness	-0.3410
Kurtosis	6.8414



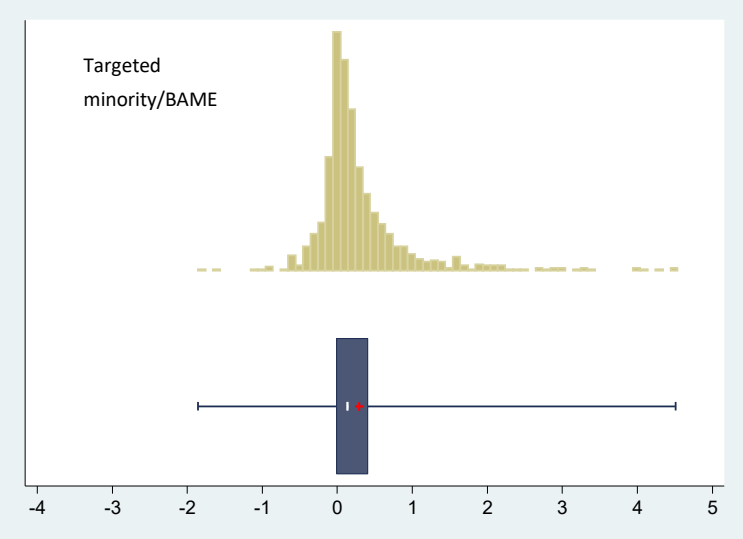
Observation	90
Mean	0.1117
Std. Dev.	0.2458
Variance	0.0604
Quartile 1	0.0085
Quartile 2 (Median)	0.0661
Quartile 3	0.1345
Range	1.8124
Skewness	1.7405
Kurtosis	10.2036



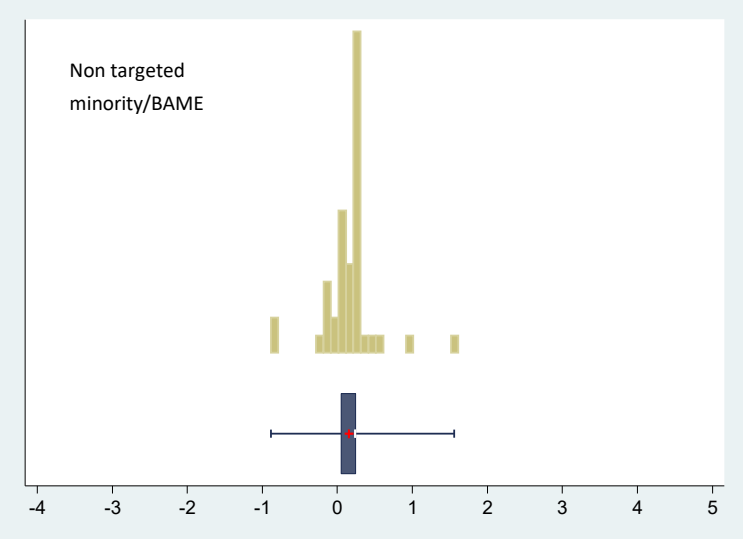
Observation	86
Mean	0.0608
Std. Dev.	0.2176
Variance	0.0473
Quartile 1	-0.0454
Quartile 2 (Median)	0.0170
Quartile 3	0.0750
Range	1.2621
Skewness	1.3820
Kurtosis	5.5947

Figure 22 compares distributions of effect sizes from BAME-targeted interventions against non-BAME targeted interventions, for child-centred outcomes. The mean ES from targeted interventions is large (M=0.29, SD=0.63, median =0.13), and greater than that from non-BAME targeted interventions (M=0.16 (moderate), SD=0.36, median = 0.24). However, the median is higher for ES on minority/BAME participants of non-BAME targeted interventions. The range of the ES is wider for targeted interventions (6.37) than for non-targeted ones (2.44). The distribution for BAME-targeted interventions is more skewed to the right and both distributions have noticeable outliers.

Figure 22. Distribution of effect sizes for child-centred outcomes, minority-targeted vs non-targeted interventions



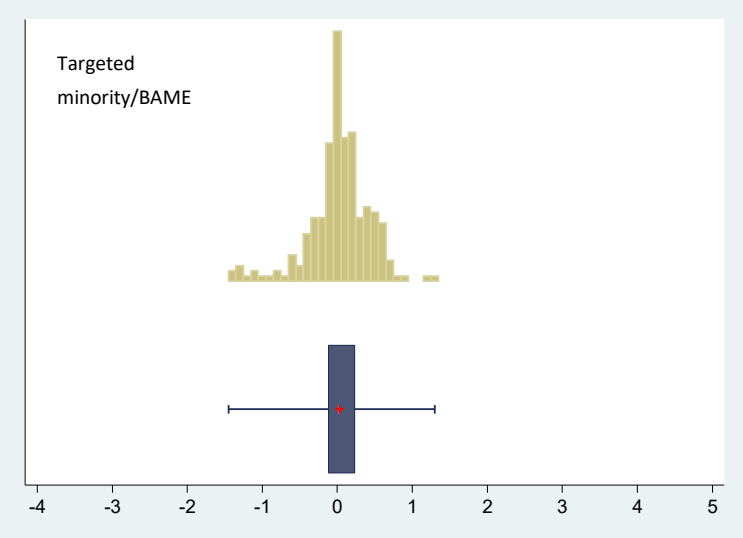
Observation	1,092
Mean	0.2874
Std. Dev.	0.6266
Variance	0.3926
Quartile 1	-0.0051
Quartile 2 (Median)	0.1334
Quartile 3	0.4092
Range	6.3699
Skewness	2.7534
Kurtosis	14.8616



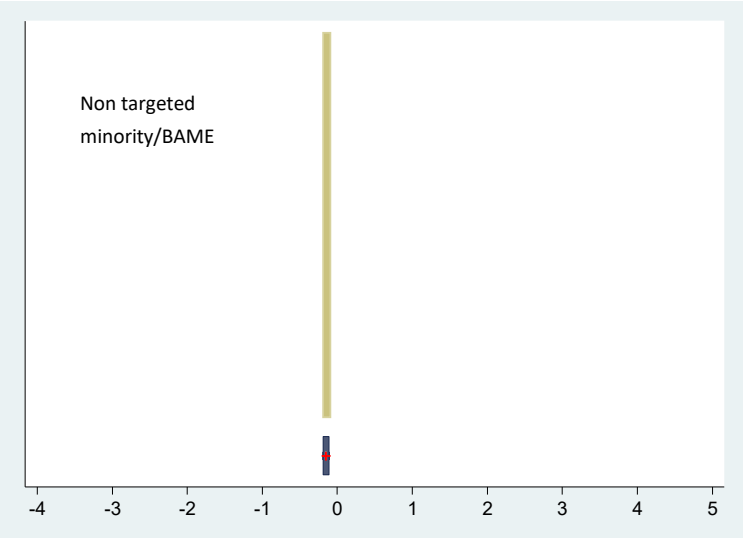
Observation	45
Mean	0.1555
Std. Dev.	0.3642
Variance	0.1326
Quartile 1	0.0519
Quartile 2 (Median)	0.2380
Quartile 3	0.2508
Range	2.4415
Skewness	0.5309
Kurtosis	8.7257

Figure 23 compares distributions of effect sizes from BAME targeted interventions against non-BAME targeted intervention, for offending and crime outcomes. The number of observations is only 2 for the non-targeted intervention, so we do not discuss this distribution. The mean ES from targeted intervention is 0.03 (SD=0.41, median =0.02), indicating no impact, and the range is 2.75. The distribution is moderately skewed to the left and has tendency for outliers.

Figure 23. Distribution of effect sizes for offending and crime outcomes, minority-targeted vs non-targeted interventions



Observation	240
Mean	0.0280
Std. Dev.	0.4125
Variance	0.1702
Quartile 1	-0.1194
Quartile 2 (Median)	0.0198
Quartile 3	0.2395
Range	2.7474
Skewness	-0.8073
Kurtosis	5.3548

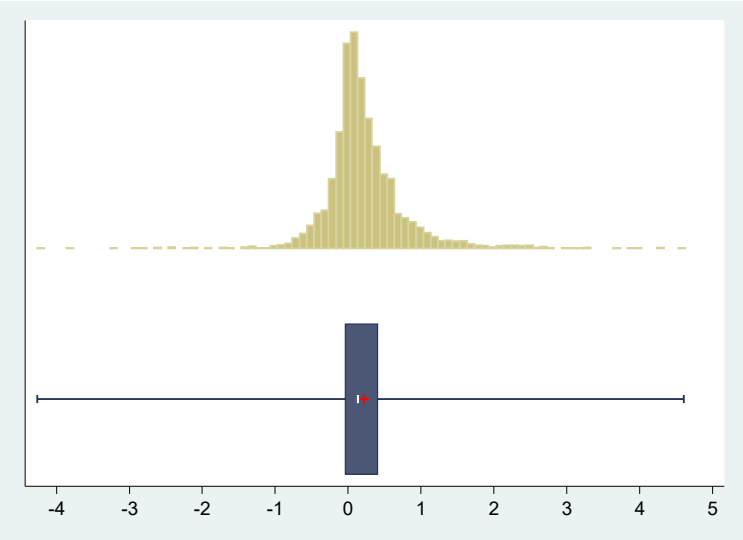


Observation	2
Mean	-0.1521
Std. Dev.	0.0552
Variance	0.0030
Quartile 1	-0.1911
Quartile 2 (Median)	-0.1521
Quartile 3	-0.1132
Range	0.0780
Skewness	0.0000
Kurtosis	1.0000

Annex 8 Additional figures not discussed in this report

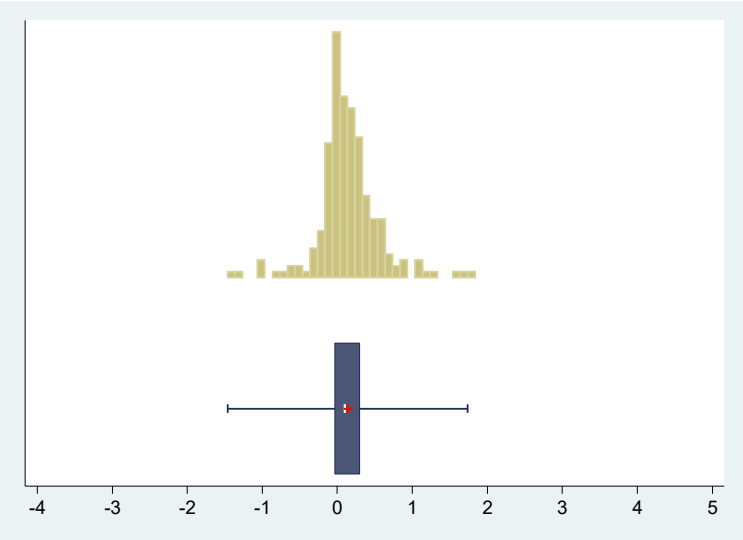
Figure A1. Distributions of other externalising behaviour, internalising behaviour and pro-social outcomes (other than SDQ) for controlled studies

a) Externalising behaviour



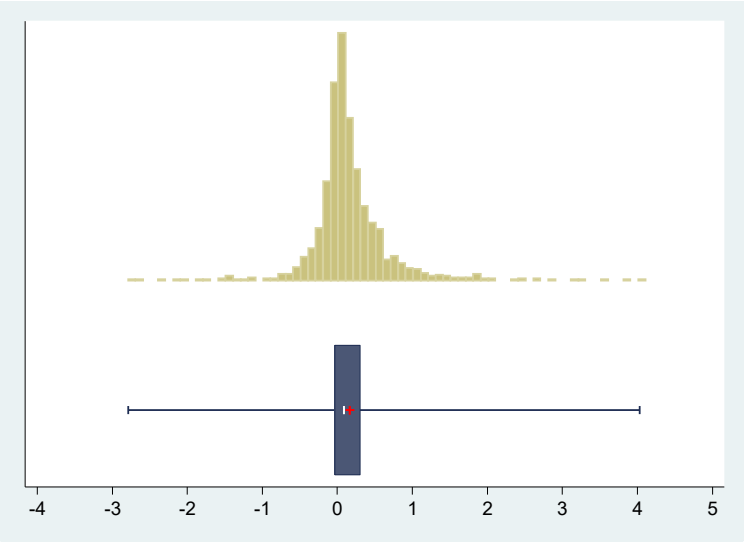
Observation	3,385
Mean	0.2175
Std. Dev.	0.5661
Variance	0.3205
Quartile 1	-0.0374
Quartile 2 (Median)	0.1350
Quartile 3	0.4060
Range	8.8713
Skewness	1.0048
Kurtosis	13.6050

b) Prosocial behaviour



Observation	294
Mean	0.1453
Std. Dev.	0.3775
Variance	0.1425
Quartile 1	-0.0190
Quartile 2 (Median)	0.1044
Quartile 3	0.2933
Range	3.2027
Skewness	0.2379
Kurtosis	7.5008

c) Internalising behaviour



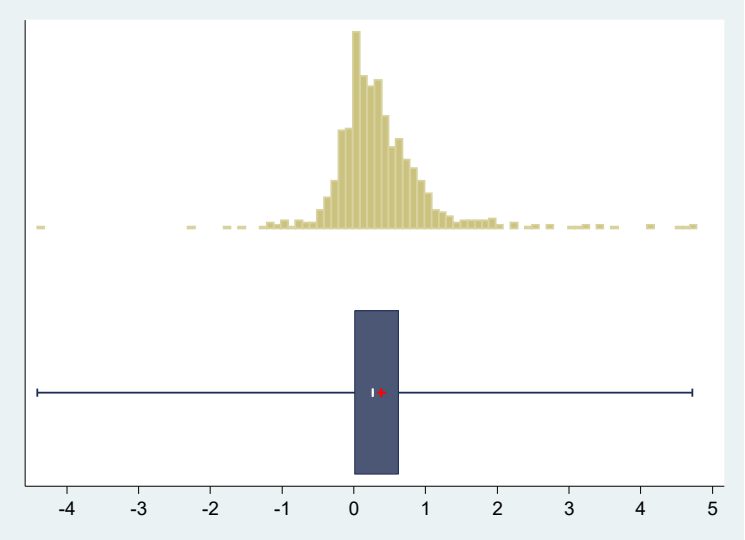
Observation	1,411
Mean	0.1712
Std. Dev.	0.5176
Variance	0.2679
Quartile 1	-0.0317
Quartile 2 (Median)	0.0907
Quartile 3	0.3146
Range	6.8204
Skewness	1.3875
Kurtosis	14.8149

Histograms and summary statistics for uncontrolled studies

The rest of this annex presents summary distributions of effect sizes for child-centred and offending and crime outcome domains (Figures B1-B2); distributions of effect sizes for SDQ measures (Figure B3), including total difficulties, externalising behaviour, internalising behaviour and pro-social behaviour; and distributions of effect sizes by age group for each outcome domain (Figures B4-B5).

Figure B1 presents the overall distribution of ES for UCS. The mean effect size is 0.38 (SD=0.74, median=0.26), which is more than double that of controlled studies. The value ranges from -4.41 to 4.72, and the distribution is highly skewed to the right and has a great tendency for outliers.

Figure B1. The overall distribution of effect sizes in uncontrolled studies



Observation	788
Mean	0.3824
Std. Dev.	0.7359
Variance	0.5415
Quartile 1	0.0128
Quartile 2 (Median)	0.2617
Quartile 3	0.6247
Range	9.1358
Skewness	1.7070
Kurtosis	13.9817

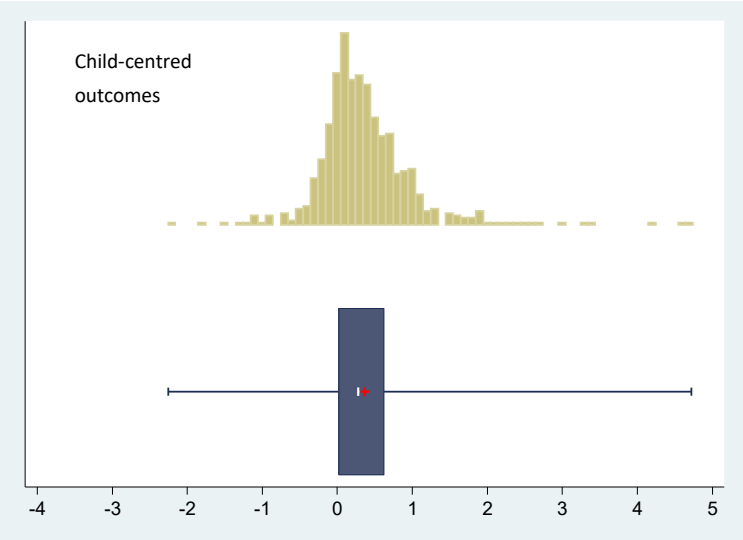
Figure B2 shows the distribution of effect sizes by outcome domain. Contrary to CS, offending and crime outcomes have a higher mean value of 0.60 which is large impact (SD=1.40, median=0.07) compared to child-centred outcomes (M=0.36; SD=0.64; median = 0.28, which is also a large ES). The range is wider for offending and crime outcomes 9.13, while the range for child-centred outcomes is 6.98. The offending and crime outcomes have less skewed to the right, and Kurtosis is high for both outcome domain suggesting greater tendency for outliers.

Figure B3 shows distribution of effect sizes for outcomes measured using SDQ. "Total difficulties score has a high mean value of 0.26 (SD=0.54) and median of 0.42. For the subscales, externalizing behaviour score and internalising behaviour score, the mean is 0.24 (SD=0.43, median=0.20) and 0.21 (SD=0.47, median=0.22) respectively. The mean ES for pro social behaviour is close to zero (SD=0.49, median=-0.06). The range is 1.5 for total difficulties score, 1.50 for externalising behaviour score, 2.02 for internalising behaviour and 1.45 for prosocial behaviour score. The distribution for total difficulties score is moderately skewed to the left, and internalising behaviour is moderately skewed to the right. Apart from internalising behaviour score, the distributions have light tails suggesting lack of outliers.

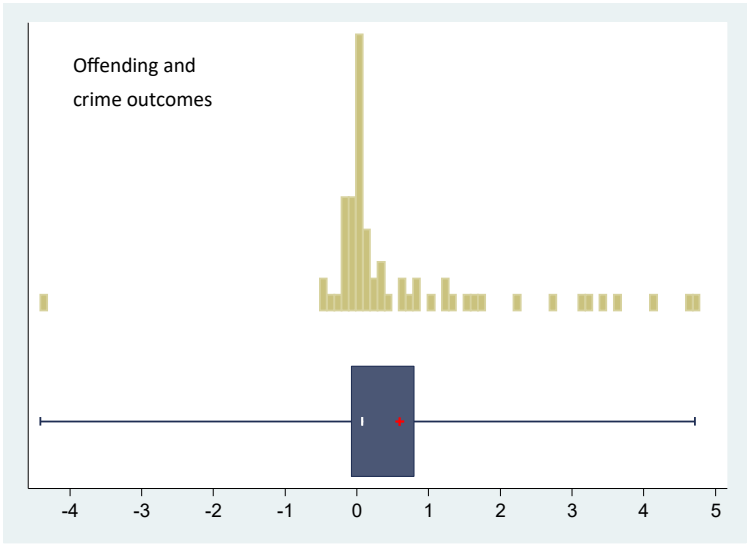
Figure B4 shows the distribution of effect sizes for child-centred outcomes by age group. The effect sizes are higher than for CS, a high average effect size is observed for 15+year (M=0.46, SD=0.86, median =0.47) and the mean ES for 0-3 year is also high (M=0.47, SD=0.64, median =0.46). The mean ES for 4-9 years is 0. 30 (SD=0.85, median=0.19) and for 10-14 year, the most frequently reported ES, mean is 0.34 (SD=0.54, median = 0.26), which are also high. The range of the ES is 5.06 for 0-3 year, 5.47 for 4-9 year, 6.81 for 10-14 year and 3.75 for 15+ year. The distributions for 4-9 years and 10-14 year are skewed to the right. The distributions have considerable tendency for outliers.

Figure B5 shows the distribution of effect sizes for offending and crime outcome by age group. There was no observation for 0-3 years. The effect sizes are higher than for CS, the highest average effect size is observed for 15+year (N=6, M=1.40, SD=1.24, median =1.41) and the mean ES for 10-14 year is also high (M=0.66, SD=1.56, median =0.07). The mean ES for 4-9 years (N=4) is close to zero (M=0.03, SD=0.35, median=0.08). The range of the ES is 0.79 for 4-9 year, 9.13 for 10-14 year and 3.22 for 15+ year. The distributions are approximately symmetric. And the distributions for 10-14 year have a greater tendency for outliers.

Figure B2. Distribution of effect sizes by outcome domain in uncontrolled studies

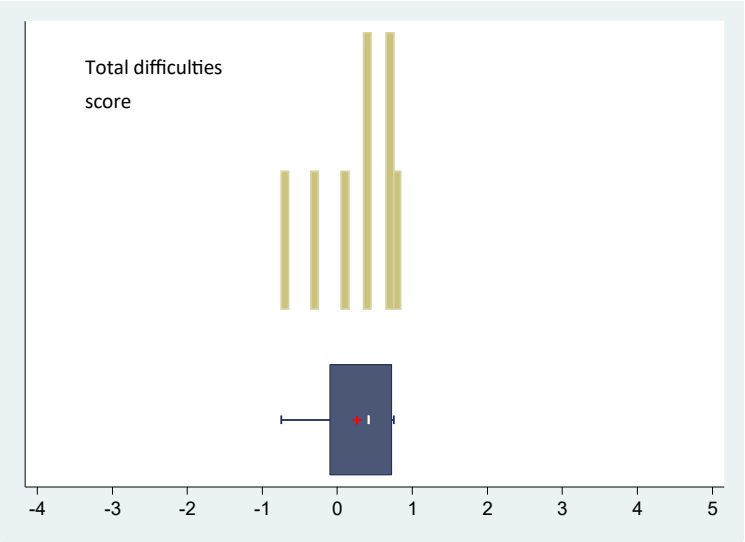


Observation	720
Mean	0.3623
Std. Dev.	0.6363
Variance	0.4048
Quartile 1	0.0205
Quartile 2 (Median)	0.2808
Quartile 3	0.6188
Range	6.9766
Skewness	1.7414
Kurtosis	12.4731

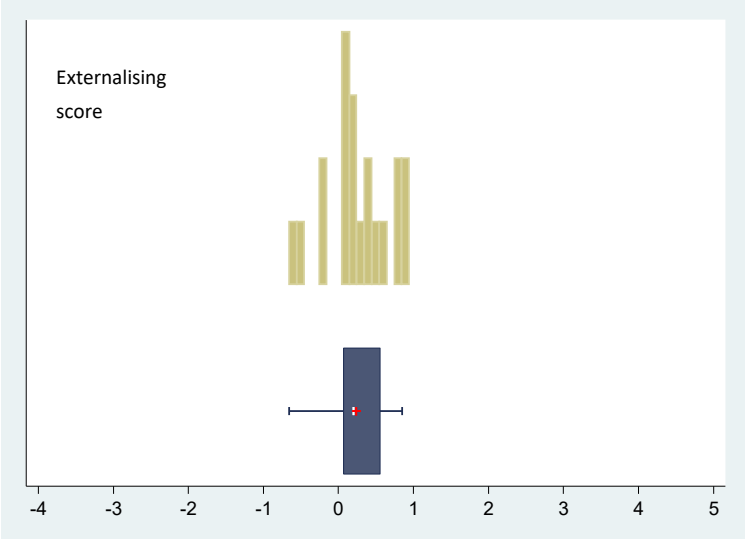


Observation	68
Mean	0.5952
Std. Dev.	1.4023
Variance	1.9664
Quartile 1	-0.0756
Quartile 2 (Median)	0.0745
Quartile 3	0.8036
Range	9.1276
Skewness	0.7936
Kurtosis	6.3486

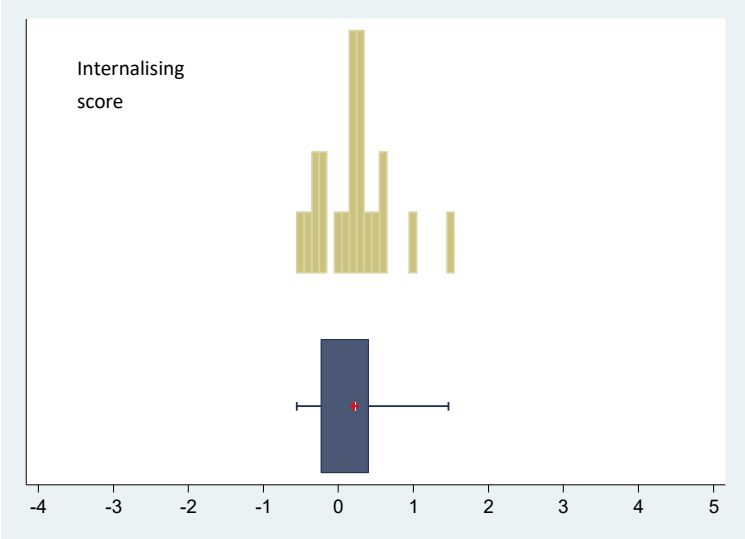
Figure B3. Distribution of effect sizes for SDQ in uncontrolled studies



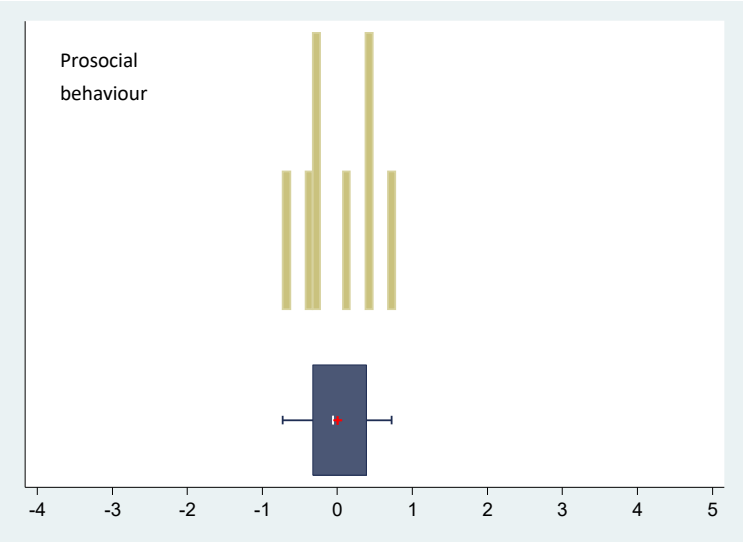
Observation	8
Mean	0.2614
Std. Dev.	0.5419
Variance	0.2937
Quartile 1	-0.0979
Quartile 2 (Median)	0.4192
Quartile 3	0.7218
Range	1.5021
Skewness	-0.8320
Kurtosis	2.4099



Observation	20
Mean	0.2377
Std. Dev.	0.4301
Variance	0.1850
Quartile 1	0.0681
Quartile 2 (Median)	0.2023
Quartile 3	0.5574
Range	1.5019
Skewness	-0.3033
Kurtosis	2.4695

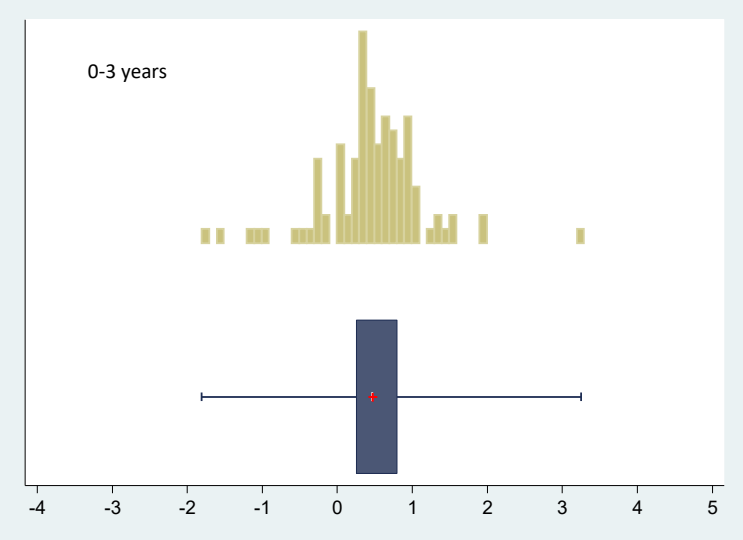


Observation	22
Mean	0.2072
Std. Dev.	0.4715
Variance	0.2223
Quartile 1	-0.2301
Quartile 2 (Median)	0.2224
Quartile 3	0.4027
Range	2.0209
Skewness	0.6510
Kurtosis	3.6882

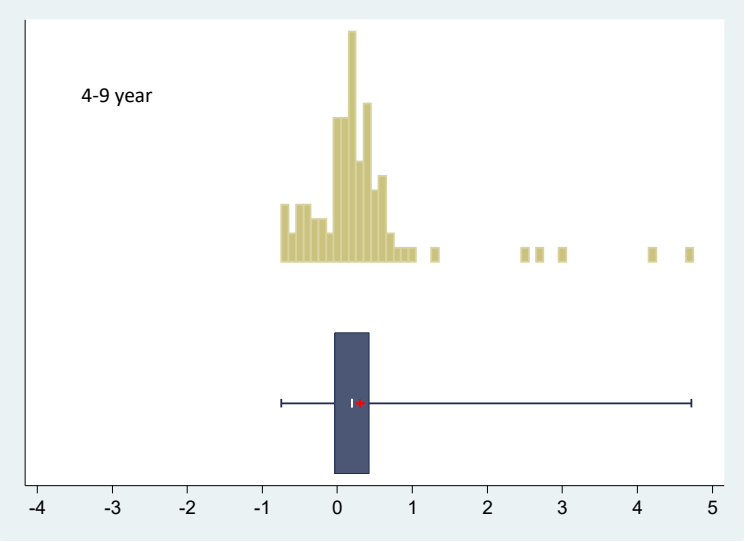


Observation	8
Mean	0.0026
Std. Dev.	0.4853
Variance	0.2355
Quartile 1	-0.3263
Quartile 2 (Median)	-0.0564
Quartile 3	0.3932
Range	1.4512
Skewness	0.0303
Kurtosis	1.8604

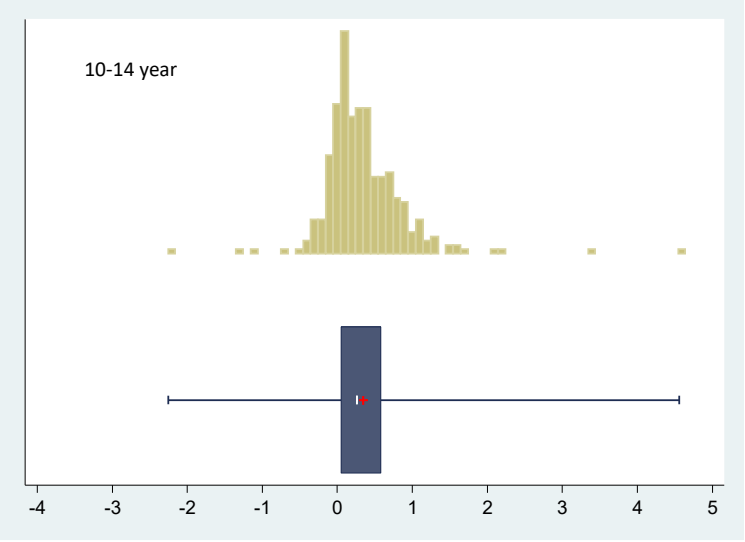
Figure B4. Distribution of effect sizes for child-centred outcomes by age in uncontrolled studies



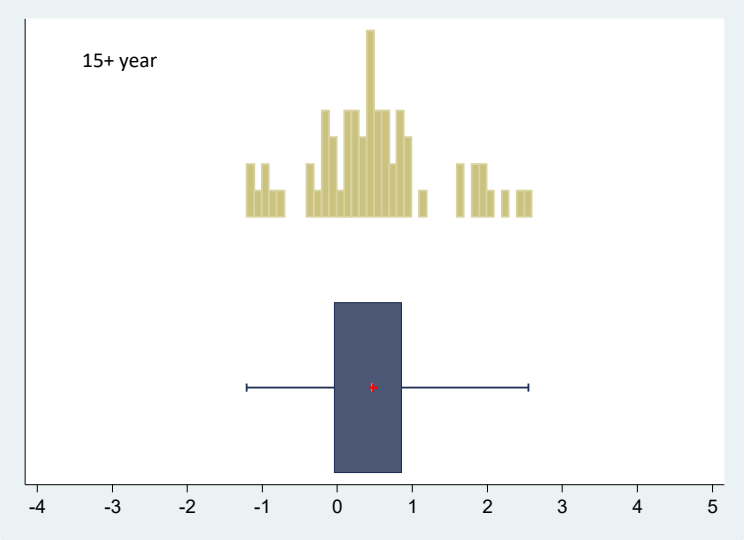
Observation	115
Mean	0.4674
Std. Dev.	0.6397
Variance	0.4092
Quartile 1	0.2318
Quartile 2 (Median)	0.4626
Quartile 3	0.7848
Range	5.0574
Skewness	-0.0178
Kurtosis	7.2013



Observation	98
Mean	0.3041
Std. Dev.	0.8452
Variance	0.7144
Quartile 1	-0.0343
Quartile 2 (Median)	0.1940
Quartile 3	0.4274
Range	5.4702
Skewness	3.0778
Kurtosis	14.7732

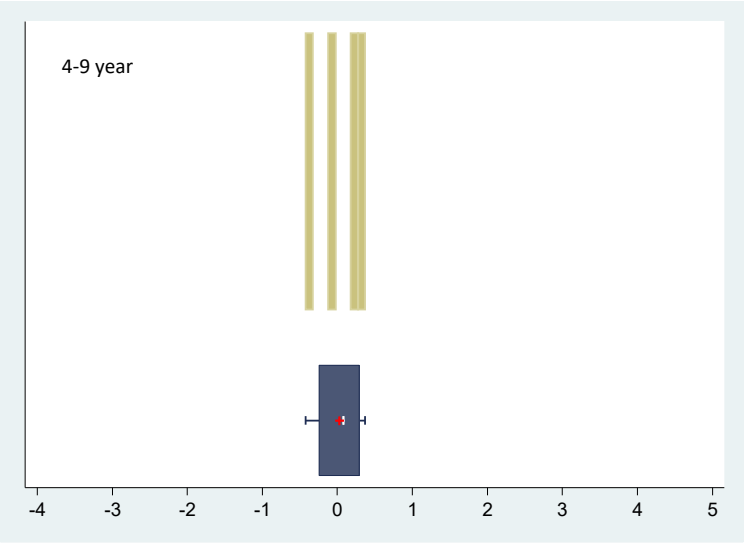


Observation	343
Mean	0.3431
Std. Dev.	0.5352
Variance	0.2864
Quartile 1	0.0510
Quartile 2 (Median)	0.2618
Quartile 3	0.5788
Range	6.8133
Skewness	1.9723
Kurtosis	18.2898

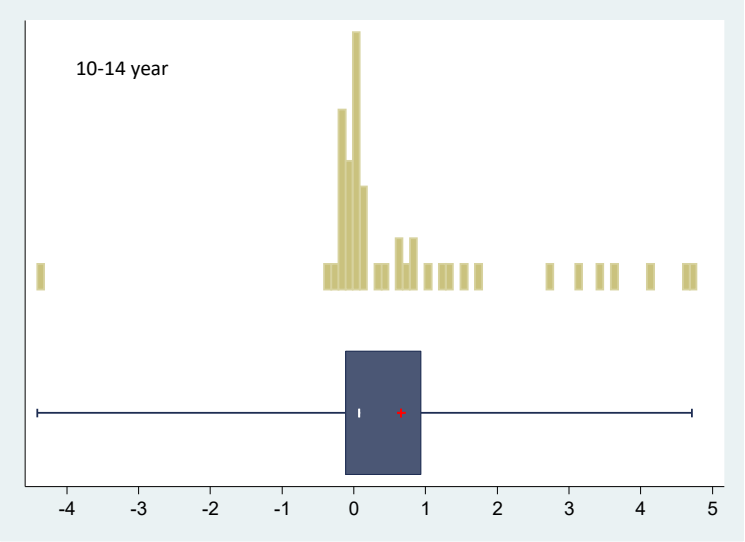


Observation	65
Mean	0.4631
Std. Dev.	0.8621
Variance	0.7432
Quartile 1	-0.0639
Quartile 2 (Median)	0.4708
Quartile 3	0.8600
Range	3.7540
Skewness	0.3860
Kurtosis	3.0339

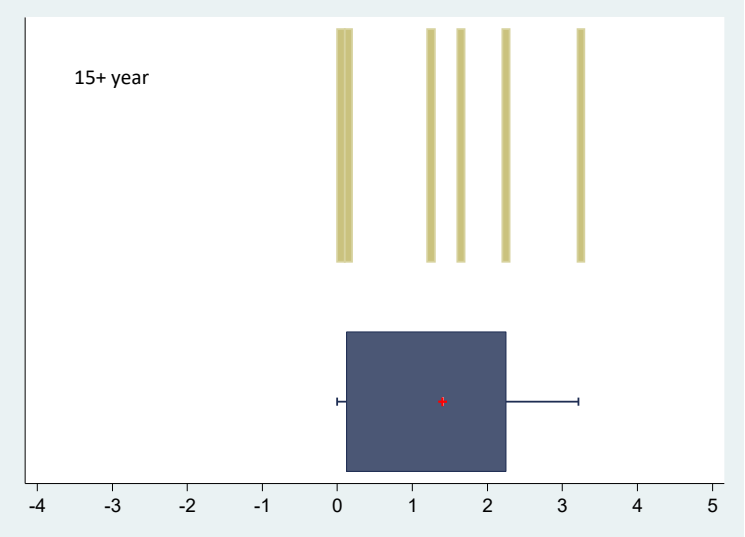
Figure B5. Distribution of effect sizes for offending and crime outcomes by age in uncontrolled studies



Observation	4
Mean	0.0265
Std. Dev.	0.3501
Variance	0.1226
Quartile 1	-0.2450
Quartile 2 (Median)	0.0806
Quartile 3	0.2981
Range	0.7902
Skewness	-0.3906
Kurtosis	1.6561



Observation	48
Mean	0.6600
Std. Dev.	1.5620
Variance	2.4398
Quartile 1	-0.1115
Quartile 2 (Median)	0.0745
Quartile 3	0.9351
Range	9.1276
Skewness	0.5855
Kurtosis	5.5900



Observation	6
Mean	1.4005
Std. Dev.	1.2410
Variance	1.5401
Quartile 1	0.1211
Quartile 2 (Median)	1.4064
Quartile 3	2.2538
Range	3.2152
Skewness	0.2015
Kurtosis	1.8116

References

Allcott, H. (2015). Site selection bias in program evaluation. The Quarterly Journal of Economics, 130 (3), 1117-1165.

Beelman, A., & Lösel, F. (2021). A comprehensive meta-analysis of randomized evaluations of the effect of child social skills training on antisocial development. *Journal of Developmental and Lifecourse Criminology, 7*(1), 41 – 65. https://doi.org/10.1007/s40865-020-00142-8

Brantingham, P. J., & Faust, F. L. (1976). A conceptual model of crime prevention. *Crime and Delinquency, 22,* 284-296.

Burdsal, C., & Buel, C. L. (1980). A short term community based early stage intervention program for behavior problem youth. *Journal of Clinical Psychology*, *36*(1), 226-241.

Farrington, D. P., Gaffney, H., & White, H. (2022). Effectiveness of 12 types of interventions in reducing juvenile offending and antisocial behaviour. *Canadian Criminal Justice Association, 64*(4), 47 – 68. https://doi.org/10.3138/cjccj.2022-0022

Farrington, D. P., & McGee, T. R. (2019). The Integrated Cognitive Antisocial Potential (ICAP) theory; New empirical tests. In In D. P. Farrington, L. Kazemian, & A. R. Piquero (Eds.), *The Oxford Handbook of Developmental and Lifecourse Criminology,* (pp. 173 – 192). Oxford University Press.

Gaffney, H., Farrington, D. P., & White, H. (2021a). *Social skills training (Toolkit technical report)*. Youth Endowment Fund.

Gaffney, H., Farrington, D. P., & White, H. (2021b). *Pre-court diversion (Toolkit technical report)*. Youth Endowment Fund.

Gaffney, H., Farrington, D. P., & White, H. (2021c). *Boot Camps (Toolkit technical report)*. Youth Endowment Fund.

Gillis, H. L., & Gass, M. A. (2010). Treating juveniles in a sex offender program using adventure-based programming: A matched group design. *Journal of Child Sexual Abuse*, *19*(1), 20-34.

Gillis, H. L., Gass, M. A., & Russell, K. C. (2008). The effectiveness of project adventure's behavior management programs for male offenders in residential treatment. *Residential Treatment for Children & Youth*, *25*(3), 227-247.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Journal of child psychology and psychiatry*, 38(5), 581-586.

Higgins, J., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M., & Welch, V. (2021). Cochrane Handbook for Systematic Reviews of Interventions. Cochrane: London.

Howell, J. C. (2013). Mark Lipsey's contribution to evidence-based services for juvenile offenders: What works across juvenile justice systems. *Criminology & Public Policy, 13*(1), 15 – 19. https://doi.org/10.1111/1745-9133.12068

Jolliffe, D., Farrington, D. P., Brunton-Smith, I., Loeber, R., Ahonen, L., & Palacios, A. P. (2019). Depression, anxiety and delinquency: Results from the Pittsburgh Youth Study. *Journal of Criminal Justice, 62,* 42 – 49. https://doi.org/10.1016/j.jcrimjus.2018.08.004

Lipsey, M. W. (1999). Can intervention rehabilitate serious delinquents? *The ANNALS of the American Academy of Political and Social Science, 564,* 142 – 166.

Lipsey, M. W., & Wilson, D. B. (1998). Effective intervention for serious juvenile offenders: A synthesis of research. In R. Loeber & D. P. Farrington (Eds.), *Serious and Violent Juvenile Offenders: Risk factors and successful interventions*. Sage.

Lipsey, M. W., & Wilson, D. B. (2001). Practical meta-analysis. SAGE publications, Inc.

Minor, K. I. (1988) An Evaluation of an Intervention Program for Juvenile Probationers. Dissertations. 2169. https://scholarworks.wmich.edu/dissertations/2169

Petrosino, A., Petrosino, C., Guckenburg, S., Terrell, J., Fronius, T. A., & Choo, K. (2019). The effects of juvenile system processing on subsequent delinquency outcomes. In D. P. Farrington, L. Kazemian, & A. R. Piquero (Eds.), *The Oxford Handbook of Developmental and Lifecourse Criminology,* (pp. 553 – 575). Oxford University Press.

Piquero, A. R., Farrington. D. P., Welsh, B. C., Tremblay, R., & Jennings, W. G. (2009). Effects of early family/parent training programs on antisocial behavior and delinquency. *Journal of Experimental Criminology, 5,* 83 – 120.

Skinner, G.C.M., & Farrington, D. P. (2021). Health of convicted persons in the third generation of the longitudinal Cambridge Study in Delinquent Development. *International Journal of Offender Therapy and Comparative Criminology,* 1 – 26. https://doi.org/10.1177/0306624X211066837

Sutton, A., Cherney, A., White, R., & Clancey, G. (2021). *Crime prevention: Principles, perspectives and practices, 3rd edition.* Cambridge University Press.

Sutton, C. E., Monaghan, M., Case, S., Greenhalgh, J., & Wright, J. (2022). Contextualising youth justice interventions: Making the case for realist synthesis. *Sustainability, 14*(2), 854. https://doi.org/10.3390/su14020854

Tonry, M., & Farrington, D. P. (1995). Strategic approaches to crime prevention. *Crime and Justice, 19,* 1 – 20. https://www.jstor.org/stable/1147594

Waddington, H., White, H., Snilstveit, B., Hombrados, J. G., Vojtkova, M., Davies, P., ... & Tugwell, P. (2012). How to do a good systematic review of effects in international development: a tool kit. *Journal of development effectiveness, 4*(3), 359-387.

Weisburd, D., Farrington, D. P., & Gill, C. (2017). What works in crime prevention and rehabilitation. *Criminology & Public Policy, 16*(2), 363 – 671. https://doi.org/10.1111/1745-9133.12298

White, H. (2014). Current challenges in impact evaluation. European Journal of Development Research, 26 (1), 18-30.

White, H., Saran, A., Verma, A., Oprea, E., & Babudu, P. (2021). Evidence and gap map of interventions to prevent children getting involved in violence.

Wilson, D. B., MacKenzie, D. L., & Mitchell, F. N. (2008). Effects of correctional boot camps on offending. *Campbell Systematic Reviews, 1*. https://doi.org/10.4073/csr.2005.6

Wright, A.N. (1982). *Therapeutic potential of the Outward Bound process: An evaluation of a treatment program for juvenile delinquents*. The Pennsylvania State University.

Youth Endowment Fund (2022). Core measurement guidance Strengths and Difficulties Questionnaire (SDQ) Guidance on the implementation and analysis of the SDQ. https://youthendowmentfund.org.uk/wp-content/uploads/2022/04/18.-YEF-SDQ-guidance-April-2022.pdf.

Ziven, H. S. (1988). *The effects of the challenge group treatment program on psychiatrically hospitalized adolescents*. Massachusetts School of Professional Psychology.