Quantitative evidence

Primary school age (primarily)

Foster-parent delivered tutoring (Teach Your Children Well) (FP-TYCW) vs Wait List (WL)

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Word reading mean score post-intervention: assessed using the Wide Range Achievement Test Fourth Edition (WRAT-4)								
1 (Flynn 2012)	Parallel RCT	77	MD 2.54 (-1.22 to 6.30)[1]	Very serious[2]	N/A	Serious[3]	NE[4]	Very low
Spelling mean score post-intervention: assessed using the WRAT-4[1]								
1 (Flynn 2012)	Parallel RCT	77	-1.2 (-8.26 to 5.86)[1]	Very serious[2]	N/A	Serious[3]	NE[4]	Very low
Maths mean score post-intervention: assessed using the WRAT-4[1]								
1 (Flynn 2012)	Parallel RCT	77	5.8 (1.58 to 10.02)[1]	Very serious[2]	N/A	Serious[3]	NE[4]	Very low
Sentence comprehension mean score post-intervention: assessed using the WRAT-4[1]								
1 (Flynn 2012)	Parallel RCT	77	4.53 (0.41 to 8.65)[1]	Very serious[2]	N/A	Serious[3]	NE[4]	Very low

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Reading composite mean score post-intervention: assessed using the WRAT-4[1]								
1 (Flynn 2012)	Parallel RCT	77	3.79 (-0.60 to 8.18)[1]	Very serious[2]	N/A	Serious[3]	NE[4]	Very low

1. Adjusted for pre-intervention (baseline) means for these scores. Confidence intervals calculated by reviewer using reported mean values and p values.
2. Downgrade 2 levels for very serious risk of bias: few baseline variables reported, so difficult to assess success of randomisation process; unclear if allocation concealment; unclear if deviations from intended intervention; Per-protocol analysis and >30% dropped out on the intervention arm; Large loss to follow up and unclear how much missing data otherwise. Missing data imputed but unclear how much and if appropriate method used. Outcome assessors were likely unblinded and outcome may be influenced by knowledge of intervention received (but not likely). Unclear and insufficient detail provided about certain aspects of conducting trial e.g. approach to loss to follow up.
3. Downgrade 1 level for serious indirectness since study was based in Canada
4. Downgraded twice as imprecision was not estimable

Volunteer-delivered tutoring (Teach Your Children Well) (V-TYCW) vs Wait List (WL)

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Word reading mean score post-intervention: assessed using the WRAT-4								
1 (Harper 2012)	Parallel RCT	68	4.45 (1.75 to 7.15)[1]	Very serious[2]	N/A	Very serious[3]	NE[4]	Very low
Spelling mean score post-intervention: assessed using the WRAT-4								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
1 (Harper 2012)	Parallel RCT	68	7.89 (2.71 to 13.07)[1]	Very serious[2]	N/A	Very serious[3]	NE[4]	Very low
Maths mean score post-intervention: assessed using the WRAT-4								
1 (Harper 2012)	Parallel RCT	68	3.2 (p value=ns)[1]	Very serious[2]	N/A	Very serious[3]	NE[4]	Very low
Sentence comprehension mean score post-intervention: assessed using the WRAT-4								
1 (Harper 2012)	Parallel RCT	68	0.86 (p value= ns)[1]	Very serious[2]	N/A	Very serious[3]	NE[4]	Very low
Word reading mean score post-intervention: assessed using the WRAT-4[1]								
1 (Harper 2016)	Parallel RCT	101	MD 4.64 (2.01 to 7.27)	Very serious[2]	N/A	Very serious[3]	NE[4]	Very low
Spelling mean score post-intervention: assessed using the WRAT-4[1]								
1 (Harper 2016)	Parallel RCT	101	MD 3.19 (0.55 to 5.83)	Very serious[2]	N/A	Very serious[3]	NE[4]	Very low
Maths mean score post-intervention: assessed using the WRAT-4[1]								
1 (Harper 2016)	Parallel RCT	101	MD 3.84 (0.15 to 7.53)	Very serious[2]	N/A	Very serious[3]	NE[4]	Very low

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Sentence comprehension mean score post-intervention: assessed using the WRAT-4[1]								
1 (Harper 2016)	Parallel RCT	101	1.70 (p value= ns)	Very serious[2]	N/A	Very serious[3]	NE[4]	Very low

1. Adjusted for pre-intervention (baseline) means for these scores. Confidence intervals calculated by reviewer using reported mean values and p values.
2. Downgrade 2 levels for very serious risk of bias: Unclear if deviations from intended intervention; unclear why loss to follow up; Per-protocol analysis; <10% lost to follow up; Unclear if outcome assessors were aware of a participants intervention status. It is possible that such knowledge could have impacted results; Unclear that analysis was conducted with a pre-specified plan e.g. for multivariable analysis; some evidence that multiple analyses were performed but only one reported. Raw data not reported.
3. Downgrade 2 levels for serious indirectness since study was based in Canada and the majority of participants were of aboriginal ethnicity
4. Downgrade twice as imprecision was not estimable

Volunteer-delivered tutoring (Teach Your Children Well) (Short) vs Volunteer-delivered tutoring (Teach Your Children Well) (long)

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Math Fluency score at postintervention: assessed using the Woodcock-Johnson-Third Edition (WJ-III)								
1 (Hickey 2020)	Parallel RCT	83	Beta coefficient – 3.94 (p=0.07)	Very serious[1]	N/A	Serious[2]	NE[3]	Very low

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Applied problems score at postintervention: assessed using the Woodcock-Johnson-Third Edition (WJ-III)								
1 (Harper 2016)	Parallel RCT	83	Beta coefficient – 3.07 (p=0.07)	Very serious[1]	N/A	Serious[2]	NE[3]	Very low

1. Downgrade 2 levels for very serious risk of bias: there were some significant differences observed between comparison groups, slightly more than would be expected by chance. However, these differences were not found to be associated with the outcomes of interest, according to the authors. Over 10% drop out in both arms and these results were excluded from the analysis, even where attendance of the intervention had begun. All of the variables had less than a 6% missing data rate, with the majority having less than 5% missing. Outcome assessors appeared to be unblinded, which may have influenced results
2. Downgrade 1 level for serious indirectness since study was based in Canada
3. Downgraded twice as imprecision was not estimable

Letterbox club vs Wait List

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Reading accuracy mean score 4-weeks post-intervention: assessed using the Neale Analysis of Reading Ability								
1 (Mooney 2016)	Parallel RCT	116	MD 1.00 (-4.57 to 6.57)	Not Serious	N/A	Not Serious[1]	Not serious	High
Reading comprehension mean score 4-weeks post-intervention: assessed using the Neale Analysis of Reading Ability								
1 (Mooney 2016)	Parallel RCT	116	MD -0.49 (-6.44 to 5.46)	Not Serious	N/A	Not Serious[1]	Not serious	High
Reading rate mean score 4-weeks post-intervention: assessed using the Neale Analysis of Reading Ability								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
1 (Mooney 2016)	Parallel RCT	116	MD -3.15 (-8.74 to 2.44)	Not Serious	N/A	Not Serious[1]	Serious[2]	Moderate
Recreational reading mean score 4-weeks post-intervention: assessed using the Elementary Reading Enjoyment Scale (known as the 'Garfield Test')								
1 (Mooney 2016)	Parallel RCT	116	MD -0.81 (-3.47 to 1.87)	Not Serious	N/A	Not Serious[1]	Serious[3]	Moderate
Academic reading mean score 4-weeks post-intervention: assessed using the Elementary Reading Enjoyment Scale (known as the 'Garfield Test')								
1 (Mooney 2016)	Parallel RCT	116	MD -0.67 (-3.32 to 1.98)	Not Serious	N/A	Not Serious[1]	Not Serious	High
Odds of liking school "a lot" 4-weeks post-intervention: children were asked "Do you like school?" with the option of reply "not really", "a little" or "a lot".								
1 (Mooney 2016)	Parallel RCT	116	OR 0.68 (0.31 to 1.47)[4]	Not Serious	N/A	Not Serious[1]	Very Serious[5]	Low
Like reading "a lot" 4-weeks post-intervention: children were asked "Do you like reading?" with the option of reply "not really", "a little" or "a lot".								
1 (Mooney 2016)	Parallel RCT	116	OR 0.93 (0.43 to 2.01)[4]	Not Serious	N/A	Not Serious[1]	Very Serious[5]	Low

1. UK-based study
2. Downgrade 1 level for serious imprecision since confidence intervals crossed 1 line of MID (defined as 0.5*SD in the control group=6.53)
3. Downgrade 1 level for serious imprecision since confidence intervals crossed 1 line of MID (defined as 0.5*SD in the control group=3.34)
4. Reviewer calculated/imputed odds ratios using percentages reported in the study
5. Downgrade 2 levels for very serious imprecision since confidence intervals crossed 2 lines of MID (defined as 0.8 and 1.25 for odds ratios)

Paired-reading intervention

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Reading age pre- vs post-intervention: assessed using the Salford test								
1 (Osbourne 2010)	Uncontrolled BA study	35	MD 1.00 (0.24 to 1.76)	Very serious[1]	N/A	Not serious[2]	Serious[3]	Very low

1. Downgrade 2 levels for very serious risk of bias: No contemporary comparison group used; Participants who were unable to adhere to the intervention were likley to have had poorer results, but were not included in this study (missing data); Participants with missing data are likely to be those who would have had poorer responses to intervention; A validated measure was used but assessors were aware of intervention status (pre/post).
2. UK-based
3. Downgrade 1 level for serious imprecision since confidence intervals crossed 1 line of MID (defined as 0.5*SD in the control group=0.83)

Secondary school-age (primarily)

Take Charge intervention (coaching and mentoring) vs Usual Care

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Mean number of hours spent doing homework post-intervention: assessed by self-report								
1 (Geenen 2012)	Parallel RCT	120	MD 0.51 (0.08 to 0.94) hours	Very serious[1]	N/A	Serious[2]	Serious[3]	Very low
Mean number of hours spent doing homework at 9-month follow up: assessed by self-report								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
1 (Geenen 2012)	Parallel RCT	120	MD 0.14 (-0.24 to 0.52) hours	Very serious[1]	N/A	Serious[2]	Serious[4]	Very low
Mean youth knowledge and engagement in educational planning score post-intervention: assessed using the student version of the Educational Planning Assessment								
1 (Geenen 2012)	Parallel RCT	120	MD 2.45 (0.98 to 3.92)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low
Mean youth knowledge and engagement in educational planning score at post-intervention follow up: assessed using the parent version of the Educational Planning Assessment								
1 (Geenen 2012)	Parallel RCT	120	MD 2.81 (-0.94 to 6.56)	Very serious[1]	N/A	Serious[2]	Serious[5]	Very low
Mean youth knowledge and engagement in educational planning score post-intervention: assessed using the teacher version of the Educational Planning Assessment								
1 (Geenen 2012)	Parallel RCT	120	MD 2.51 (-0.35 to 5.37)	Very serious[1]	N/A	Serious[2]	Serious[6]	Very low
Mean youth knowledge and engagement in educational planning score at 9-month follow up: assessed using the student version of the Educational Planning Assessment								
1 (Geenen 2012)	Parallel RCT	120	MD 2.68 (-0.23 to 5.59)	Very serious[1]	N/A	Serious[2]	Serious[7]	Very low
Mean youth knowledge and engagement in educational planning score at 9-month follow up: assessed using the parent versions of the Educational Planning Assessment								
1 (Geenen 2012)	Parallel RCT	120	MD 3.22 (0.32 to 6.12)	Very serious[1]	N/A	Serious[2]	Serious[8]	Very low

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Mean youth knowledge and engagement in educational planning score at 9-month follow up: assessed using the teacher versions of the Educational Planning Assessment								
1 (Geenen 2012)	Parallel RCT	120	MD 2.77 (-0.23 to 5.77)	Very serious[1]	N/A	Serious[2]	Serious[9]	Very low
Student self-attribution of accomplishments score post-intervention: youth were asked to list all their educational accomplishments for the past 6 months and a total count was gathered at each time point.								
1 (Geenen 2012)	Parallel RCT	120	MD 0.80 (0.33 to 1.27)	Very serious[1]	N/A	Serious[2]	Serious[10]	Very low
Student self-attribution of accomplishments score at 9-months follow up: youth were asked to list all their educational accomplishments for the past 6 months and a total count was gathered at each time point.								
1 (Geenen 2012)	Parallel RCT	120	MD 0.24 (-0.22 to 0.70)	Very serious[1]	N/A	Serious[2]	Serious[11]	Very low

1. Downgrade 2 levels for very serious risk of bias: Some considerable differences between comparison groups for length of time in foster care, speech and language disability, autism, and emotional/behavioural needs; unclear if any deviations from intended interventions; unclear if intention to treat analysis used (but most likely); Just over 10% with missing data post randomisation; unclear whether any further missing outcome data; unclear reasons for drop out; unclear how drop out varied between groups; It is possible that missingness of data is related to outcomes; It is unclear how assessments were performed (by whom). Unclear if facilitators were aware of intervention status of participants. Measurements used are often crude indicators of the phenomenon of interest; unclear that analysis was conducted according to a pre-specified protocol. Data not provided for certain non-significant results. Evidence of multiple analyses used for different outcomes.
2. Downgrade 1 level for serious indirectness since study was based in USA
3. Downgrade 1 level for serious imprecision since confidence intervals crossed 1 line of MID (defined as 0.5*SD in the control group=0.56)
4. Downgrade 1 level for serious imprecision since confidence intervals crossed 1 line of MID (defined as 0.5*SD in the control group=0.48)
5. Downgrade 1 level for serious imprecision since confidence intervals crossed 1 line of MID (defined as 0.5*SD in the control group=6.45)
6. Downgrade 1 level for serious imprecision since confidence intervals crossed 1 line of MID (defined as 0.5*SD in the control group=4.03)
7. Downgrade 1 level for serious imprecision since confidence intervals crossed 1 line of MID (defined as 0.5*SD in the control group=4.58)
8. Downgrade 1 level for serious imprecision since confidence intervals crossed 1 line of MID (defined as 0.5*SD in the control group=4.07)

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
9. Downgrade 1 level for serious imprecision since confidence intervals crossed 1 line of MID (defined as 0.5*SD in the control group=4.45) 10. Downgrade 1 level for serious imprecision since confidence intervals crossed 1 line of MID (defined as 0.5*SD in the control group=0.60) 11. Downgrade 1 level for serious imprecision since confidence intervals crossed 1 line of MID (defined as 0.5*SD in the control group=0.62)								

Multidimensional treatment foster care for adolescents (MTFC-A) vs Usual Care

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Odds of higher scholastic/language skills at 12 months follow up: assessed by a domain of the Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA)								
1 (Green 2014)	Parallel RCT	34	OR 0.6 (0.15 to 2.4)	Very serious[1]	N/A	Not serious[2]	Serious[3]	low
Odds of higher school attendance score at 12 months follow up: assessed by a domain of the Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA)								
1 (Green 2014)	Parallel RCT	34	OR 2.5 (0.48 to 13.1)	Very serious[1]	N/A	Not serious[2]	Serious[3]	low

1. Downgrade 2 levels for very serious risk of bias: Unclear if/why participants did not receive allocated intervention; Significant deviations apparent since 8/20 in the treatment group did not receive their interventions; In the intervention group 15-20% had missing data; it was also unclear how much other data was missing since some outcomes were imputed; Unclear if appropriate imputation methods used; reasons for missing data not given; Missingness of data may well be related to the result of the outcomes reported.
2. UK-based
3. Downgrade 2 levels for very serious imprecision since confidence intervals crossed 2 lines of MID (defined as 0.8 and 1.25 for odds ratios)

Multidimensional treatment foster care (MTFC) vs Group Care control

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Mean homework completion score at 3-6 months post-intervention: composite score using the number of days in the last week that the girls spent at least 30 min/day on homework; and whether or not the girls did homework that day.								
1 (Leve 2007)	Parallel RCT	81	MD 0.64 (0.16 to 1.12)	Very serious[1]	N/A	Serious[2]	Serious[3]	Very low
Mean homework completion score at 12 months post-intervention: composite score using caregiver and girl report of the number of days in the last week that the girls spent at least 30 min/day on homework; and whether or not the girls did homework that day.								
1 (Leve 2007)	Parallel RCT	81	MD 1.44 (0.59 to 2.29)	Very serious[1]	N/A	Serious[2]	Serious[4]	Very low
Mean school attendance score at 12 months post baseline: composite score using caregivers and girls reports of how often the girls attended school.								
1 (Leve 2007)	Parallel RCT	81	MD 0.61 (0.15 to 1.07)	Very serious[1]	N/A	Serious[2]	Serious[5]	Very low

1. Downgrade 2 levels for very serious risk of bias: Unclear how randomisation was performed or if allocation concealment; Unclear if all participants assigned to their groups received their interventions as allocated. Intention to treat analysis used; Over 10% lost to follow up. Unclear how much additional missing outcome data or if this differed between comparison groups; Quite crude measures used for homework completion and school attendance. Unclear if outcome assessors were aware of intervention group. Possibility that reporting of outcomes was affected by knowledge of intervention group; Insufficient information to convince that trial was conducted according to a prespecified plan that was finalised before unblinded outcome data was available.
2. Downgrade 1 level for serious indirectness since study was based in USA
3. Downgrade 1 level for serious imprecision since confidence intervals crossed 1 line of MID (defined as 0.5*SD in the control group=0.57)
4. Downgrade 1 level for serious imprecision since confidence intervals crossed 1 line of MID (defined as 0.5*SD in the control group=1.06)
5. Downgrade 1 level for serious imprecision since confidence intervals crossed 1 line of MID (defined as 0.5*SD in the control group=0.67)

ESTEP tutoring programme vs No ESTEP tutoring

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Mean letter-word identification score at approximately 26 months follow up: assessed by Woodcock–Johnson Tests of Achievement III								
1 (Zinn 2014/Courtney 2008)	Parallel RCT	529	MD 2.10 (-2.25 to 6.45)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low
Mean calculation score at approximately 26 months follow up: assessed by Woodcock–Johnson Tests of Achievement III								
1 (Zinn 2014/Courtney 2008)	Parallel RCT	529	MD -0.30 (-4.22 to 3.62)	Very serious[1]	N/A	Serious[2]	Not serious	Very low
Mean passage comprehension score at approximately 26 months follow up: assessed by Woodcock–Johnson Tests of Achievement III								
1 (Zinn 2014/Courtney 2008)	Parallel RCT	529	MD -0.20 (-4.33 to 3.93)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low
Mean highest grade level completion at approximately 26 months follow up: self-report								
1 (Zinn 2014/Courtney 2008)	Parallel RCT	529	MD 0.00 (-0.19 to 0.19)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low
Mean grade point average at follow up at approximately 26 months follow up: Participants reported their school grades they had received across four core subjects during their previous full semester of school. Responses were scored based on a standard 4-point scale, and an overall GPA was computed by taking the average of these.								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
1 (Zinn 2014/Courtney 2008)	Parallel RCT	529	MD 0.00 (-0.18 to 0.18)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low
School behaviour score: youths were asked to indicate how often they had had "trouble" completing the following five tasks during their last full semester of school attendance: (1) getting along with your teachers, (2) paying attention in school, (3) getting your homework done, (4) getting along with other students, and (5) arriving on time for class. Response options ranged from "never" (0) to "every day" (5). School behaviour was then operationalized or defined as the mean of these five items.								
1 (Zinn 2014/Courtney 2008)	Parallel RCT	529	MD -0.02 (-0.25 to 0.21)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low
Achieving high school diploma or general equivalency diploma at approximately 26 months follow up: self-report								
1 (Zinn 2014/Courtney 2008)	Parallel RCT	529	OR 0.79 (0.41 to 1.52)	Very serious[1]	N/A	Serious[2]	Very Serious[3]	Very low

1. Downgrade 2 levels for very serious risk of bias: No information about randomisation process or whether allocation was concealed; 12% of randomised participants were excluded immediately following randomisation; While intention to treat analysis was used, there was significant deviations from the intended treatment in both groups. 38.2% of those assigned to the E-STEP group did not receive E-STEP services and 12.3% of those in the control group did receive ESTEP services; other than the 12% who were excluded immediately following randomisation, there was also >10% who did not respond to the follow up surveys. The reasons for this are unclear and may be associated with having poorer school outcomes; Unclear if assessors were blinded to intervention status. It is possible that they may influence some of the outcomes; Insufficient information provided to convince that trial was conducted according to a pre-specified plan; study authors note that approximately equal proportions of ESTEP and control groups received some form of tutoring (58.4% vs 60.8%); Only results from second follow up were reported.
2. Downgrade 1 level for serious indirectness since study was based in USA
3. Downgrade 2 levels for very serious imprecision since confidence intervals crossed 2 lines of MID (defined as 0.8 and 1.25 for odds ratios)

Animal-assisted psychotherapy vs residential care as usual

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Mean change in self-rated school maladjustment (pre- vs post-intervention): measured as part of the Spanish version of the Behavior Assessment System for Children.[1]								
1 (Balluerka 2015)	Non-randomised controlled trial	67	MD -0.63 (-5.48 to 4.22)	Very serious[2]	N/A	Serious[3]	NE[4]	Very low
Mean change in teacher-rated school maladjustment (pre- vs post-intervention): measured as part of the Spanish version of the Behavior Assessment System for Children.[1]								
1 (Balluerka 2015)	Non-randomised controlled trial	67	MD -3.19 (-6.93 to 0.55)	Very serious[2]	N/A	Serious[3]	NE[4]	Very low
Mean change in teacher-rated behavioural symptoms (pre- vs post-intervention): measured as part of the Spanish version of the Behavior Assessment System for Children.[1]								
1 (Balluerka 2015)	Non-randomised controlled trial	67	MD -1.39 (-5.92 to 3.14)	Very serious[2]	N/A	Serious[3]	NE[4]	Very low
Mean change in teacher-rated adaptive skills (pre- vs post-intervention): measured as part of the adaptive skills composite of the Teacher Rating Scale.[1]								
1 (Balluerka 2015)	Non-randomised controlled trial	67	MD 5.88 (1.61 to 10.15)	Very serious[2]	N/A	Serious[3]	NE[4]	Very low
Mean change in negative attitude towards school score (pre- vs post-intervention): attitude to school scale of the Self-Report of Personality[1]								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
1 (Muela 2017)	Non-randomised controlled trial	87	MD -0.03 (-4.28 to 4.22)	Very serious[2]	N/A	Serious[3]	NE[4]	Very low
Mean change in negative attitude towards teachers score (pre- vs post-intervention): attitude to teachers scale of the Self-Report of Personality[1]								
1 (Muela 2017)	Non-randomised controlled trial	87	MD -2.69 (-4.73 to -0.65)	Very serious[2]	N/A	Serious[3]	NE[4]	Very low

1. Confidence intervals calculated by reviewer using reported mean values and p values.
2. Downgrade 2 levels for very serious risk of bias: Matching methods used. Unclear how matching criteria were measured. Similarity between groups was not reported in detail; Large amounts of missing data for various outcomes reported, no reason for missing data provided; Teachers/caregivers/residential care staff were unaware of intervention status. However, self-report outcomes were completed with knowledge of intervention status; various subscales reported (often if significant) but not others.
3. Downgrade 1 level for serious indirectness since study was based in Spain
4. Downgrade two levels as imprecision was not estimable

All ages

Child advocate volunteers vs care as usual

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Pass all courses by year 1 (%):[1] **unclear how school indicators were measured/reported**								
1 (Waxman 2009)	Non-randomised controlled trial	581	OR 3.05 (2.09 to 4.45)	Very serious[2]	N/A	Serious[3]	Not Serious	Very low

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Poor conduct by year 1 (%)[1] unclear how school indicators were measured/reported								
1 (Waxman 2009)	Non-randomised controlled trial	581	OR 0.35 (0.25 to 0.49)	Very serious[2]	N/A	Serious[3]	Not Serious	Very low
Expelled by year 1 (%)[1] unclear how school indicators were measured/reported								
1 (Waxman 2009)	Non-randomised controlled trial	581	OR 0.51 (0.25 to 1.06)	Very serious[2]	N/A	Serious[3]	Serious[4]	Very low
Pass all courses by year 2 (%)[1] unclear how school indicators were measured/reported								
1 (Waxman 2009)	Non-randomised controlled trial	581	OR 1.55 (0.97 to 2.48)	Very serious[2]	N/A	Serious[3]	Serious[4]	Very low
Poor conduct by year 2 (%)[1] unclear how school indicators were measured/reported								
1 (Waxman 2009)	Non-randomised controlled trial	581	0.84 (0.60 to 1.18)	Very serious[2]	N/A	Serious[3]	Serious[4]	Very low
Expelled by year 2 (%)[1] unclear how school indicators were measured/reported								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
1 (Waxman 2009)	Non-randomised controlled trial	581	OR 0.92 (0.55 to 1.53)	Very serious[2]	N/A	Serious[3]	Very Serious[5]	Very low

1. Calculated using percentages reported in study
2. Downgrade 2 levels for very serious risk of bias: Participants were only matched for gender, age, and type of abuse. However, there are several other relevant factors e.g. behaviour, special education needs, and mental health problems; Unclear if intervention had already begun at the start of observation period. Children still in advocate system may be those with more stable placements. Therefore, starting observation midway through the treatment may ignore those who received treatment with worse outcomes; Unclear how often advocates met with youth, or the placement types of those youth. Treatment children received double the amount of counselling ?as a direct result of the intervention but not necessarily; Unclear level of interaction youth had with the advocate. Only assignment of treatment tested. Unclear if deviations from intended intervention, however drop out was high; By year 2, there was a 10-15% loss to follow up. Also there was substantial missing data which was >50% in some cases. Unclear reasons for missing data and how reasons differed between groups; Interviewers were the advocates (the treatment givers) in the intervention group. Therefore, different personnel were used to carry out interviews for different comparison groups."Not all measures were administered to all children" but no further information provided.
3. Downgrade 1 level for serious indirectness since study was based in USA
4. Downgrade 1 level for very serious imprecision since confidence intervals crossed 1 lines of MID (defined as 0.8 and 1.25 for odds ratios)
5. Downgrade 2 levels for very serious imprecision since confidence intervals crossed 2 lines of MID (defined as 0.8 and 1.25 for odds ratios)

Evolve Interagency Services vs care as usual

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Problems with scholastic or language skills score: assessed using a subscale of the Health of the Nations Outcome Scale for Children and Adolescents								
1 (Klag 2010)	Uncontrolled before and after study	255	MD -0.64 (-0.87 to -0.41)	Very serious[2]	N/A	Serious[3]	Not Serious	Very low

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Poor school attendance score: assessed using a subscale of the Health of the Nations Outcome Scale for Children and Adolescents								
1 (Klag 2010)	Uncontrolled before and after study	249	MD -0.54 (-0.29 to -0.79)	Very serious[2]	N/A	Serious[3]	Not Serious	Very low

1. Calculated using percentages reported in study
2. Downgrade 2 levels for very serious risk of bias: Participants were only matched for gender, age, and type of abuse. However, there are several other relevant factors e.g. behaviour, special education needs, and mental health problems; Unclear if intervention had already begun at the start of observation period. Children still in advocate system may be those with more stable placements. Therefore, starting observation midway through the treatment may ignore those who received treatment with worse outcomes; Unclear how often advocates met with youth, or the placement types of those youth. Treatment children received double the amount of counselling, possibly as a direct result of the intervention but this is not clear; Unclear level of interaction youth had with the advocate. Only assignment of treatment tested. Unclear if deviations from intended intervention, however, drop out was high; By year 2, there was a 10-15% loss to follow up. Also, there was substantial missing data which was >50% in some cases. Unclear reasons for missing data and how reasons differed between groups; Interviewers were the advocates (the treatment givers) in the intervention group. Therefore, different personnel were used to carry out interviews for different comparison groups. "Not all measures were administered to all children" but no further information provided.
3. Downgrade 1 level for serious indirectness since study was based in Australia