Quantitative evidence

Preschool interventions

Attachment and biobehavioural catch-up for infants (ABC-I) vs Developmental Education for Families (DEF)

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Receptive language score at 3 years of age: assessed using the Peabody Picture Vocabulary Test (third edition)								
1 (Bernard 2017)	Parallel RCT	52	MD 9.97 (1.58 to 18.36)	Very serious[1]	N/A	Serious[2]	Serious[3]	Very low
Association between being in the intervention group and receptive language score at 3 years of age: assessed using the Peabody Picture Vocabulary Test (third edition)								
1 (Bernard 2017)	Parallel RCT	52	β 9.39 (0.82 to 17.96)[4]	Very serious[1]	N/A	Serious[2]	NE[5]	Very low

1. Downgrade 2 levels for very serious risk of bias: unclear if allocation concealment; unclear how many lost to follow up and reasons why; loss to follow up could be related to outcome of interest; no blinding procedure described; no detailed protocol or original study cited
2. Downgrade 1 level for serious indirectness since study was based in USA
3. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=7.26)
4. Adjusted for gender, number of placements at baseline, low caregiver education, low caregiver income
5. Downgrade twice as imprecision was not estimable

Attachment and biobehavioural catch-up for toddlers (ABC-T) vs Developmental Education for Families (DEF)

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Attention problems score at approx. 2 years follow up: assessed using the Attention Problems Scale in the preschool version of the Child Behaviour Checklist (CBCL)								
1 (Lind 2017)	Parallel RCT	111	MD -0.90 (-1.66 to -0.14)	Very serious[1]	N/A	Serious[2]	Serious[3]	Very low
Cognitive flexibility score at approx. 2 years follow up: assessed by the Dimensional Change Card Sort (DCCS) task developed for preschoolers								
1 (Lind 2017)	Parallel RCT	111	MD 5.13 (0.51 to 9.75)	Very serious[1]	N/A	Serious[2]	Serious[4]	Very low
Receptive vocabulary (assessed at approximately 36, 48, and 60 months of age to form a composite score at 2 years of follow up): assessed using the Peabody Picture Vocabulary Test (PPVT third edition).								
1 (Raby 2019)	Parallel RCT	88	MD 7.10 (0.32 to 13.88)	Very serious[5]	N/A	Serious[2]	Serious[6]	Very low

1. Downgrade 2 levels for very serious risk of bias: unclear how randomisation was performed; unclear if allocation concealment; no discussion of approach to loss to follow up; A significant amount of missing data (>10% per arm) was observed in the final analysis - unclear how much of this was due to loss to follow up and how much due to missing outcome data; unclear reasons for loss to follow up; loss to follow up could be related to outcome of interest; study does not cite original trial or protocol; Multiple assessments were performed yearly however only selected time points were reported.
2. Downgrade 1 level for serious indirectness since study was based in USA
3. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=1.06)
4. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=6.44)

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
5. Only per-protocol analysis performed. Participants that did not complete all 10 sessions were excluded from analysis. Very large loss to follow up in both arms (approximately 20 - 25%). Very large amount of missing data. Combining numbers missing due to loss to follow up and missing outcomes, over 54% participants were missing from the ABC-T arm and 50% from the DEF arm. It is plausible that missing outcome data was related to placement changes which may be related to a child's ability to communicate/special education needs. PPVT was measured at different age points and averaged across these ages. However, PPVT scores increase with age and some children were missing scores at different annual follow ups. It is unclear if children in one intervention were older (on average) at assessment than children in the other arm after taking into account missing data. Does not link to original study or protocol. Outcome was measured at different time points. However, only composite outcomes were reported. 6. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=8.25)								

Attachment and biobehavioural catch-up for infants and toddlers (ABC-I/T) vs Developmental Education for Families (DEF)

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Theory of mind score at 4-6 years of age: assessed by the penny hiding game task								
1 (Lewis-Morrarty 2012)	Parallel RCT	37	MD 1.96 (0.84 to 3.08)	Very serious[1]	N/A	Serious[2]	Serious[3]	Very low
Cognitive flexibility score at 4-6 years of age: assessed by the Dimensional Change Card Sort task								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
1 (Lewis-Morrarty 2012)	Parallel RCT	37	MD 2.60 (1.01 to 4.19)	Very serious[1]	N/A	Serious[2]	Serious[4]	Very low

1. Downgrade 2 levels for very serious risk of bias: unclear if appropriate method used for randomisation; unclear if allocation concealment; significant differences between comparison groups across several domains: age; gender; ethnicity; and parental financial income; insufficient information about whether appropriate analysis used; unclear number of participants analysed; no information about missing data provided; unclear if blinding performed; original study or protocol not clearly cited; unclear how participants were sampled from original trial; participants were assessed annually until age 6 but it is unclear at what assessment results were reported.
2. Downgrade 1 level for serious indirectness since study was based in USA
3. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=1.26)
4. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=1.44)

Head start programme vs care as usual

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Association between being in the intervention group and assessor-rated pre-academic skills composite score at 1 year post intervention: assessed by Woodcock-Johnson III: letter-word identification, spelling, and applied problems subscales								
1 (Lipscomb 2013)	Parallel RCT	253	β 0.16 (0.02 to 0.30)[1]	Very serious[2]	N/A	Serious[3]	NE[4]	Very low
Association between being in the intervention group and teacher-rated teacher-child relationship at 1 year: assessed by student-teacher relationship scale								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
1 (Lipscomb 2013)	Parallel RCT	253	β 0.30 (0.12 to 0.48)[1]	Very serious[2]	N/A	Serious[3]	NE[4]	Very low
Association between being in the intervention group and teacher/caregiver-reported behaviour problems at 1 year: assessed by Achenbach Child Behaviour Checklist/Adjustment scales for Preschool interventions								
1 (Lipscomb 2013)	Parallel RCT	253	β -0.18 (-0.36 to 0.00)[1]	Very serious[2]	N/A	Serious[3]	NE[4]	Very low
Maths score at 5-6 years of age: assessed by the Woodcock-Johnson III Tests of Achievement, Math Reasoning (for girls)								
1 (Lee 2016a, Lee 2016b)	Parallel RCT	162	MD 4.40 (3.48 to 5.32)	Very serious[5]	N/A	Serious[3]	Not Serious	Very low
Maths score at 5-6 years of age: assessed by the Woodcock-Johnson III Tests of Achievement, Math Reasoning (for boys)								
1 (Lee 2016a, Lee 2016b)	Parallel RCT	162	MD -8.40 (-9.23 to -7.57)	Very serious[5]	N/A	Serious[3]	Not Serious	Very low
Reading score at 5-6 years of age: assessed by the Woodcock-Johnson III Tests of Achievement, Oral Comprehension (for girls)								
1 (Lee 2016a, Lee 2016b)	Parallel RCT	162	MD 4.80 (4.18 to 5.42)	Very serious[5]	N/A	Serious[3]	Not Serious	Very low
Reading score at 5-6 years of age: assessed by the Woodcock-Johnson III Tests of Achievement, Oral Comprehension (for boys)								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
1 (Lee 2016a, Lee 2016b)	Parallel RCT	162	MD -3.20 (-3.95 to -2.45)	Very serious[5]	N/A	Serious[3]	Not Serious	Very low
Association between being in the intervention group and child-teacher relationship at 5 - 6 years of age: assessed by the modified Robert Pianta scale								
1 (Lee 2016a, Lee 2016b)	Parallel RCT	162	β -0.30 (-1.01 to 0.41)[6]	Very serious[5]	N/A	Serious[3]	NE[4]	Very low
Association between being in the intervention group and caregiver-rated positive approach to learning at 5 - 6 years of age: assessed by Achenbach /Edelbrock/Howell score								
1 (Lee 2016a, Lee 2016b)	Parallel RCT	162	β 0.11 (-0.01 to 0.23)[6]	Very serious[5]	N/A	Serious[3]	NE[4]	Very low
Association between being in the intervention group and teacher-rated aggressive score at 5 - 6 years of age: assessed by Adjustment Scales for Preschool Intervention								
1 (Lee 2016a, Lee 2016b)	Parallel RCT	162	β -1.57 (-1.41 to 4.55)[6]	Very serious[5]	N/A	Serious[3]	NE[4]	Very low
Association between being in the intervention group and teacher-rated hyperactive score at 5 - 6 years of age: assessed by Adjustment Scales for Preschool Intervention								
1 (Lee 2016a, Lee 2016b)	Parallel RCT	162	β -3.28 (-6.26 to -0.30)[6]	Very serious[5]	N/A	Serious[3]	NE[4]	Very low

1. Adjusted for baseline preacademic skills, baseline behaviour problems, age, special education needs, gender, family income to needs ratio, authoritarian caregiving, parent child reading, change in caregiver over prior year.

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality

2. Downgrade 2 levels for very serious risk of bias: Study did not provide information about differences between comparison groups for baseline characteristics other than for age and ethnicity; no information regarding whether any participants deviated from their planned intervention; no information about the approach to missing data or loss to follow up; unclear whether there was significant missing data and how this varied between comparison groups; outcomes could have been influenced by knowledge of the intervention group; unclear that blinding was performed; insufficient information provided about methods and analysis plan; no explanation of why certain covariables were included in the final model.
3. Downgrade 1 level for serious indirectness since study was based in USA
4. Downgraded twice as imprecision was not estimable
5. Downgrade 2 levels for very serious risk of bias: unclear how randomisation was performed; unclear if allocation concealment; no-shows accounted for 15 and 20 percent of the full randomly assigned Head Start sample; crossovers accounted for 17 and 14 percent of the randomly assigned control group; unclear how much missing data for participants included in this study; The "reading score" test was a test of oral comprehension (understanding of a spoken passage and ability to provide a missing word based on clues); Several other educational outcomes were available for analysis according to the full report, but were not reported in this study.
6. Adjusted for age, gender, special education needs, lower cognitive skills at baseline, ethnicity, education, family income, relative care, parental book reading.

Entering primary school-age education

Therapeutic playgroups vs care as usual

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Foster parent-rated social competence at 2 weeks follow up: assessed by Child Behavior Checklist								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
1 (Pears 2007)	Parallel RCT	20	MD 1.53 (0.63 to 2.43)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low
Foster parent-rated externalising behaviours at 2 weeks follow up: assessed by Child Behavior Checklist								
1 (Pears 2007)	Parallel RCT	20	MD -2.20 (-5.59 to 1.19)	Very serious[1]	N/A	Serious[2]	Serious[3]	Very low
Foster parent-rated internalising behaviours at 2 weeks follow up: assessed by Child Behavior Checklist								
1 (Pears 2007)	Parallel RCT	20	MD 1.30 (-2.52 to 5.12)	Very serious[1]	N/A	Serious[2]	Very Serious[4]	Very low
Teacher-rated social problems at 1 month following the start of school: assessed by Teacher Report Form								
1 (Pears 2007)	Parallel RCT	20	MD 0.00 (-2.72 to 2.72)	Very serious[1]	N/A	Serious[2]	Very Serious[5]	Very low
Teacher-rated externalising behaviours at 1 month following the start of school: assessed by Teacher Report Form								
1 (Pears 2007)	Parallel RCT	20	MD 0.90 (-7.12 to 8.92)	Very serious[1]	N/A	Serious[2]	Very Serious[6]	Very low
Teacher-rated internalising behaviours at 1 month following the start of school: assessed by Teacher Report Form								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
1 (Pears 2007)	Parallel RCT	20	MD 0.10 (-6.71 to 6.91)	Very serious[1]	N/A	Serious[2]	Very Serious[7]	Very low
Foster parent-rated emotional regulation at 2 weeks follow up: assessed by Emotion Regulation Checklist								
1 (Pears 2007)	Parallel RCT	20	MD -0.03 (-0.20 to 0.14)	Very serious[1]	N/A	Serious[2]	Very Serious[8]	Very low
Foster parent-rated emotional lability at 2 weeks follow up: assessed by Emotion Regulation Checklist								
1 (Pears 2007)	Parallel RCT	20	MD -0.14 (-0.34 to 0.06)	Very serious[1]	N/A	Serious[2]	Serious[9]	Very low
Assessor-rated emotional lability at 2 weeks follow up: assessed by Emotion Regulation Checklist								
1 (Pears 2007)	Parallel RCT	20	MD -0.41 (-0.65 to -0.17)	Very serious[1]	N/A	Serious[2]	Serious[10]	Very low
Teacher-rated emotional regulation at 1 month following the start of school: assessed by Emotion Regulation Checklist								
1 (Pears 2007)	Parallel RCT	20	MD -0.18 (-0.69 to 0.33)	Very serious[1]	N/A	Serious[2]	Serious[11]	Very low
Teacher-rated emotional lability at 1 month following the start of school: assessed by Emotion Regulation Checklist								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
1 (Pears 2007)	Parallel RCT	20	MD 0.22 (-0.26 to 0.70)	Very serious[1]	N/A	Serious[2]	Very Serious[12]	Very low

1. Downgrade 2 levels for very serious risk of bias: randomisation process not described; unclear if allocation concealment; reasons for participant attrition and missing data not provided; >10% lost to follow up or missing data; teachers and assessors were blinded to the intervention but foster parents were not; unclear that trial was analysed with a pre-specified plan (lots of missing information).
2. Downgrade 1 level for serious indirectness since study was based in USA
3. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=1.94)
4. Downgrade 2 levels for very serious imprecision since confidence intervals crossed two lines of minimum important effect (half the standard deviation of the control arm=1.25)
5. Downgrade 2 levels for very serious imprecision since confidence intervals crossed two lines of minimum important effect (half the standard deviation of the control arm=2.02)
6. Downgrade 2 levels for very serious imprecision since confidence intervals crossed two lines of minimum important effect (half the standard deviation of the control arm=5.05)
7. Downgrade 2 levels for very serious imprecision since confidence intervals crossed two lines of minimum important effect (half the standard deviation of the control arm=3.90)
8. Downgrade 2 levels for very serious imprecision since confidence intervals crossed two lines of minimum important effect (half the standard deviation of the control arm=0.08)
9. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=0.12)
10. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=0.26)
11. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=0.32)
12. Downgrade 2 levels for very serious imprecision since confidence intervals crossed two lines of minimum important effect (half the standard deviation of the control arm=0.28)

Kids in Transition to School (KITS) programme vs care as usual

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Initial sound fluency score following intervention: assessed by subtest of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS)								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	MD 0.81 (-1.22 to 2.84)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low
Letter naming fluency following intervention: assessed by subtest of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS)								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	MD 0.23 (-2.81 to 3.27)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low
Concepts about print score following intervention: assessed by the Concepts About Print test								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	MD 0.65 (-0.37 to 1.67)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Caregiver rating of pre-reading skills following intervention: caregivers asked and scored on whether their child could recognise the letters of the alphabet and write his/her first name								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	MD -0.13 (-0.37 to 0.11)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low
Association between being in the intervention group and early literacy skills following intervention before starting school: assessed by a composite of standardised means from indicators of early literacy skills above (initial sound fluency, letter naming fluency, concepts about print, and pre-reading skills).								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	β 0.10 $P<0.05$[3]	Very serious[1]	N/A	Serious[2]	NE[4]	Very low
Prosocial skills score following intervention: assessed by Preschool Penn Interactive Peer Play Scale (PIPPS) score								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	MD -0.05 (-0.17 to 0.07)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Social competence score following intervention: assessed by the Child Behaviour Checklist								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	MD -0.10 (-0.67 to 0.47)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low
Emotional understanding score following intervention: assessed by matching vignettes to correct emotional state								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	MD -0.21 (-1.01 to 0.59)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low
Association between being in the intervention group and prosocial skills following intervention before starting school: assessed by composite of indicators of prosocial skills, above (prosocial skills score, social competence score, and emotional understanding score)								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	β 0.4 P>0.05[5]	Very serious[1]	N/A	Serious[2]	NE[4]	Very low
Inhibitory control score following intervention: assessed by a composite score from the Inhibitory Control subscale and the Attentional Focusing subscale (of the Children's Behavior Questionnaire), the Inhibit subscale from the Brief Rating Inventory of Executive Function–								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Preschool Version, and two computer-administered tasks shown to activate specific regions of the prefrontal cortex and anterior cingulate gyrus								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	MD 0.03 (-0.18 to 0.24)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low
Behavioural regulation score following intervention: assessed by a composite score of the Activity Level subscale and Impulsivity subscale (of the Childrens Behaviour Questionnaire), the Externalizing subscale (of the Child Behaviour Checklist), and the Lability subscale of the Emotion Regulation Checklist (ERC)								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	MD 0.14 (-0.11 to 0.39)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low
Emotional regulation score following intervention: assessed by a composite score from the anger subscale and the reactivity/soothability subscale (of the Children’s Behaviour Questionnaire), the Emotion Regulation scale (of the Emotion Regulation Checklist), and the Emotion Control subscale (of the BRIEF–P)								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	MD 0.00 (-0.22 to 0.22)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Association between being in the intervention group and self-regulatory skills following intervention before starting school: assessed by composite of indicators of self-regulation, above (inhibitory control, behavioural regulation, emotional regulation)								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	β 0.11 $P<0.05$[6]	Very serious[1]	N/A	Serious[2]	NE[4]	Very low
Teacher-reported aggressive behaviour at the end of kindergarten year: assessed by the aggressive behavior subscales of the Teacher Report Form								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	MD -1.84 (-4.81 to 1.13)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low
Teacher-reported delinquent behaviour at the end of kindergarten year: assessed by the delinquent behavior subscales of the Teacher Report Form								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	MD -0.58 (-1.21 to 0.05)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low
Teacher-reported oppositional behaviour at the end of kindergarten year: assessed by the oppositional subscale of the Conners' Teacher Ratings Scales-Revised: Short version (CTRS:S)								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	MD -0.81 (-1.78 to 0.16)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low
Association between being in the intervention group and child oppositional and aggressive behaviours at the end of kindergarten year: assessed by composite of indicators of oppositional and aggressive behaviours, above (aggressive behaviour, delinquent behaviour, and oppositional behaviour).								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	β -0.17 $P<0.05$[7]	Very serious[1]	N/A	Serious[2]	NE[4]	Very low
Days free from internalising symptoms over 12 months of kindergarten: assessed by symptom reports from caregivers on the Child Behavior Checklist (CBCL) to create days that had significant internalizing symptoms								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	MD 26.00 (0.05 to 51.95)	Very serious[1]	N/A	Serious[2]	Serious[8]	Very low
Days free from externalising problems over 12 months of kindergarten: assessed by symptom reports from caregivers on the Child Behavior Checklist (CBCL) to create days that had significant externalizing behaviors								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	MD 26.60 (-2.76 to 55.96)	Very serious[1]	N/A	Serious[2]	Serious[9]	Very low
Positive attitudes towards alcohol at 9 years of age: assessed by questions adapted from the Monitoring the Future National Survey Questionnaire								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	MD -0.30 (-0.50 to -0.10)	Very serious[1]	N/A	Serious[2]	Serious[10]	Very low
Positive attitudes towards antisocial behaviours at 9 years of age: assessed based on responses to two questions - "What are some of the things you think teenagers do for fun with their friends?" and "What are some of the things you think teenagers do when their moms or dads are not there?"								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	MD -0.09 (-0.27 to 0.09)	Very serious[1]	N/A	Serious[2]	Serious[11]	Very low
Involvement with deviant peers at 9 years of age: assessed by responses to questions about whether "none", "some", or "all" of their friends were involved in five rule-breaking or deviant behaviors								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	MD -0.19 (-0.44 to 0.06)	Very serious[1]	N/A	Serious[2]	Not Serious	Very low
Self-competence at 9 years of age: assessed by six questions on the Global Self-Worth Scale of the Self-Perception Profile for Children.								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	MD 1.91 (0.82 to 3.00)	Very serious[1]	N/A	Serious[2]	Serious[12]	Very low
Association between being in the intervention group and positive attitudes towards alcohol at 9 years of age: assessed by questions adapted from the Monitoring the Future National Survey Questionnaire								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	β -0.34 P<0.05[13]	Very serious[1]	N/A	Serious[2]	NE[4]	Very low
Association between being in the intervention group and positive attitudes towards antisocial behaviour at 9 years of age: assessed based on two questions - "What are some of the things you think teenagers do for fun with their friends?" and "What are some of the things you think teenagers do when their moms or dads are not there?"								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	β -0.11 P<0.05[13]	Very serious[1]	N/A	Serious[2]	NE[4]	Very low
Association between being in the intervention group and self-competence at 9 years of age: assessed based on the Global Self-Worth Scale of the Self-Perception Profile for Children								
1 (Pears 2012, Pears (2013), Pears (2016), Lynch (2017))	Parallel RCT	192	β 1.95 P<0.01[13]	Very serious[1]	N/A	Serious[2]	NE[4]	Very low

1. Downgrade 2 levels for very serious risk of bias: randomisation process not described; unclear if allocation concealment; there was significant missing data "ranging from 0 - 40%" across measures; unclear how different outcomes were affected by missing data; reasons for missing data not outlined; unclear how quantity of missing data differed between intervention groups; insufficient information to confirm pre-specified protocol/no cited protocol; Composite outcomes were frequently created from the results of multiple (separate) scales, these subscales were not reported separately. There was also no cited protocol to show that methods of analysing data had been pre-agreed.
2. Downgrade 1 level for serious indirectness since study was based in USA
3. Adjusted for general cognitive ability at baseline and early literacy skills at baseline
4. Downgraded twice as imprecision was not estimable
5. Adjusted for gender, kinship foster care, prosocial skills at baseline
6. Adjusted for gender, Latino ethnicity, self-regulatory skills at baseline, day-care attendance
7. Adjusted for oppositional and aggressive behaviours at baseline, gender, overall level of disruptiveness in classroom
8. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=50.75)

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 one line of minimum important effect (half the standard deviation of the control arm=52.30) 10. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=0.41) 11. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=0.16) 12. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=2.09) 13. Adjusted for gender, general cognitive ability at baseline, kinship foster care, child oppositional and aggressive behaviour at baseline, placement changes during study, other psychological/educational services								

Entering secondary school-age education

Middle school success intervention vs care as usual

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
Association between being in the intervention group and foster parent and girl reported internalising problems at 6 months: assessed by Parent Daily Report Checklist								
1 (Kim 2011, Smith 2011)	Parallel RCT	100	β -0.28 P<0.01[1]	Very serious[2]	N/A	Serious[3]	NE[4]	Very low
Association between being in the intervention group and foster parent and girl reported externalising problems at 6 months: assessed by Parent Daily Report Checklist								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
1 (Kim 2011, Smith 2011)	Parallel RCT	100	β -0.21 P<0.01[5]	Very serious[2]	N/A	Serious[3]	NE[4]	Very low
Association between being in the intervention group and foster parent and girl reported prosocial behaviour at 6 months: assessed by Parent Daily Report Checklist								
1 (Kim 2011, Smith 2011)	Parallel RCT	100	β 0.15 P>0.05[6]	Very serious[2]	N/A	Serious[3]	NE[4]	Very low
Prosocial behaviour score at 6/12 months follow up: assessed by a subscale from the Parent Daily Report Checklist								
1 (Kim 2011, Smith 2011)	Parallel RCT	100	MD 0.06 (0.01 to 0.11)	Very serious[2]	N/A	Serious[3]	Serious[7]	Very low
Caregiver-reported Internalising/externalising symptoms score at 12/24 months follow up: assessed by the Achenbach System of Empirically Based Assessment								
1 (Kim 2011, Smith 2011)	Parallel RCT	100	MD 0.27 (-3.03 to 3.57)	Very serious[2]	N/A	Serious[3]	Not Serious	Very low
Self-reported association with delinquent peers score at 12 months follow up: assessed by a modified version of the general delinquency scale from the Self-Report Delinquency Scale								
1 (Kim 2011, Smith 2011)	Parallel RCT	100	**Beta -0.21 SE 0.09 P<0.05**	Very serious[2]	N/A	Serious[3]	NE[4]	Very low
Delinquent behaviour score at 3 years follow up: assessed using the Self-Report Delinquency Scale								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
1 (Kim 2011, Smith 2011)	Parallel RCT	100	MD -0.65 (-1.43 to 0.13)	Very serious[2]	N/A	Serious[3]	Serious[8]	Very low
Association with delinquent peers score at 3 years follow up: assessed by a modified version of the general delinquency scale from the Self-Report Delinquency Scale								
1 (Kim 2011, Smith 2011)	Parallel RCT	100	MD -0.34 (-0.71 to 0.03)	Very serious[2]	N/A	Serious[3]	Serious[9]	Very low
Substance use score at 3 years follow up (composite): girls were asked how many times in the past year they had (a) smoked cigarettes or chewed tobacco, (b) drank alcohol (beer, wine, or hard liquor), and (c) used marijuana. The response scale ranged from 1 (never) through 9 (daily).								
1 (Kim 2011, Smith 2011)	Parallel RCT	100	MD -0.74 (-1.33 to -0.15)	Very serious[2]	N/A	Serious[3]	Serious[10]	Very low
Tobacco use score at 3 years follow up (composite): girls were asked how many times in the past year they had smoked cigarettes or chewed tobacco. The response scale ranged from 1 (never) through 9 (daily).								
1 (Kim 2011, Smith 2011)	Parallel RCT	100	MD -0.87 (-1.69 to -0.05)	Very serious[2]	N/A	Serious[3]	Serious[11]	Very low
Alcohol use score at 3 years follow up (composite): girls were asked how many times in the past year they had drank alcohol (beer, wine, or hard liquor). The response scale ranged from 1 (never) through 9 (daily).								
1 (Kim 2011, Smith 2011)	Parallel RCT	100	MD -0.31 (-0.78 to 0.16)	Very serious[2]	N/A	Serious[3]	Serious[12]	Very low
Marijuana use score at 3 years follow up (composite): girls were asked how many times in the past year they had used marijuana. The response scale ranged from 1 (never) through 9 (daily).								

No. of studies	Study design	Sample size	Effect size (95% CI)	Risk of bias	Inconsistency	Indirectness	Imprecision	Quality
1 (Kim 2011, Smith 2011)	Parallel RCT	100	MD -1.04 (-1.74 to -0.34)	Very serious[2]	N/A	Serious[3]	Serious[13]	Very low

1. Adjusted for age, maltreatment history, pubertal development, internalising behaviours at baseline
2. Downgrade 2 levels for very serious risk of bias: unclear if allocation concealment; approximately 10% loss to follow up by 2 years; analysis of outcomes at various time points appeared to be decided post-hoc; results (apart from results for substance use and delinquency) appear to have been selected on the basis of results across multiple time points.
3. Downgrade 1 level for serious indirectness since study was based in USA
4. Downgraded 2 levels as imprecision was not estimable
5. Adjusted for age, maltreatment history, pubertal development, externalising behaviours at baseline
6. Adjusted for age, maltreatment history, pubertal development, prosocial behaviours at baseline
7. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=0.07)
8. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=1.35)
9. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=0.51)
10. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=0.97)
11. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=1.25)
12. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=0.73)
13. Downgrade 1 level for serious imprecision since confidence intervals crossed one line of minimum important effect (half the standard deviation of the control arm=1.22)