nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-018-0461-x

A systematic review and meta-analysis of effects of early life non-cognitive skills on academic, psychosocial, cognitive and health outcomes

Lisa G. Smithers[1,2,5], Alyssa C. P. Sawyer[1,2,5], Catherine R. Chittleborough[1,2], Neil M. Davies[3,4], George Davey Smith[3,4] and John W. Lynch[1,2,3*]

Success in school and the labour market relies on more than high intelligence. Associations between 'non-cognitive' skills in childhood, such as attention, self-regulation and perseverance, and later outcomes have been widely investigated. In a systematic review of this literature, we screened 9,553 publications, reviewed 554 eligible publications and interpreted results from 222 better-quality publications. Better-quality publications comprised randomized experimental and quasi-experimental intervention studies (EQIs) and observational studies that made reasonable attempts to control confounding. For academic achievement outcomes, there were 26 EQI publications but only 14 were available for meta-analysis, with effects ranging from 0.16 to 0.37 s.d. However, within subdomains, effects were heterogeneous. The 95% prediction interval for literacy was consistent with negative, null and positive effects (−0.13 to 0.79). Similarly, heterogeneous findings were observed for psychosocial, cognitive and language, and health outcomes. Funnel plots of EQIs and observational studies showed asymmetric distributions and potential for small study bias. There is some evidence that non-cognitive skills associate with improved outcomes. However, there is potential for small study and publication bias that may overestimate true effects, and the heterogeneity of effect estimates spanned negative, null and positive effects. The quality of evidence from EQIs underpinning this field is lower than optimal and more than one-third of observational studies made little or no attempt to control confounding. Interventions designed to develop children's non-cognitive skills could potentially improve outcomes. The interdisciplinary researchers interested in these skills should take a more strategic and rigorous approach to determine which interventions are most effective.

It is over 40 years since economists Bowles and Gintis[1], in their critique of the US education system, pointed to the importance of skills for labour market success beyond those captured by intelligence, abstract reasoning and academic achievement in literacy and numeracy. They used the term 'non-cognitive personality traits' (p. 116) and pointed to motivation, orientation to authority, internalization of work norms, discipline, temperament and perseverance as characteristics that influenced life success. Although it may be intuitive that there is more to success in life than high intelligence, there has been no attempt to systematically assess the research evidence on the effects of improving different types of non-cognitive skills. We recognize that there is no neat conceptual dichotomy separating cognitive from some non-cognitive skills, but for the purposes of this review, we collectively label the diverse set of factors represented in the literature as 'non-cognitive' skills (see Table 1 for a glossary of specific terms). This literature includes studies that either manipulated non-cognitive skills through randomized controlled trials (RCTs) and quasi-experimental designs, or used observed differences in non-cognitive skills through longitudinal or cross-sectional studies. In observational (correlational) data, results from comparing outcomes for higher and lower levels of non-cognitive skills are often used as evidence for their importance in the same way as results from experimental studies.

These non-cognitive skills include attention, executive function, inhibitory control, self-control, self-regulation, effortful control, emotion regulation, delay of gratification and temperament (see Table 1 for our conceptualizations of these constructs). The importance of social skills for labour market success has been demonstrated[2], but this review does not directly include improving social skills in early life as a non-cognitive ability, although the range of psychosocial outcomes includes social skill constructs. We sought to provide a systematic representation of research into non-cognitive abilities and behaviours. The need for such a systematic review is driven by the fact that these abilities are being considered by policymakers to underpin early life interventions[3], beyond cognitive abilities (intelligence or IQ) and academic achievement (literacy and numeracy).

The policy motivation

This body of research spans disciplines including psychology, sociology, economics, health and education. It is also of great policy interest to governments in many countries[3,4], who wish to sustain future economic productivity and social inclusion, by investing substantial resources to bolster the development of human capabilities in early life[5], especially for disadvantaged children. The investment logic is that children who develop cognitive and non-cognitive skills early in life have better outcomes later in life. The policy outcomes of most interest are longer term, including labour market success, welfare dependency, social relationships, better mental and physical health that ultimately lead to a more skilled, healthy and productive

[1]School of Public Health, University of Adelaide, Adelaide, South Australia, Australia. [2]Robinson Research Institute, University of Adelaide, Adelaide, South Australia, Australia. [3]Bristol Medical School: Population Health Sciences, University of Bristol, Bristol, UK. [4]Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, UK. [5]These authors contributed equally: Lisa G. Smithers, Alyssa C. P. Sawyer.
*e-mail: john.lynch@adelaide.edu.au

Table 1 | Glossary

Attention	A state of awareness in which the senses are focused selectively on aspects of the environment and the central nervous system is in a state of readiness to respond to stimuli[105].
Cognitive flexibility	This refers to a capacity for objective appraisal of and appropriate, flexible action. It involves adaptability, objectivity and fair-mindedness[106].
Conscientiousness	The tendency to be organized, responsible and hardworking, construed as a dimension of individual differences in the Big Five and Five-Factor Personality models[105].
Delay of gratification	The ability to forgo immediate reward for the sake of greater, future reward based on the original definitions by Mischel[106].
Effortful control	Includes the abilities to voluntarily manage attention (attentional regulation) and inhibit (inhibitory control) or activate (activational control) behaviour as needed to adapt, especially when the child does not particularly want to do so[107].
Emotional reactivity	The extent to which an individual experiences emotions (1) in response to a wide array of stimuli (emotion sensitivity), (2) strongly or intensely (emotion intensity) and (3) for a prolonged period of time before returning to a baseline level of arousal (emotion persistence)[108].
Emotional regulation	The ability of an individual to modulate an emotion or set of emotions. Techniques of conscious emotional regulation can include learning to construe situations differently to manage them better and recognizing how different behaviours can be used in the service of a given emotional state[105].
Executive function	Higher-level cognitive processes that organize and order behaviour, such as judgement, abstraction and concept formation, logic and reasoning, problem solving, planning and sequencing of actions[105].
Impulsivity	Behaviour characterized by little or no forethought, reflection or consideration of the consequences[105].
Inhibitory control	The ability to suppress a pre-potent response, interrupt an ongoing response and resist distraction from external stimuli[109].
Persistence	The quality or state of maintaining a course of action or keeping at a task and finishing it despite the obstacles (such as opposition or discouragement) or the effort involved[105].
Self-control	The ability to be in command of one's behaviour (overt, covert, emotional or physical) and to restrain or inhibit one's impulses[105].
Self-regulation	The control of one's own behaviour through the use of self-monitoring (keeping a record of behaviour), self-evaluation (assessing the information obtained during self-monitoring) and self-reinforcement (rewarding oneself for appropriate behaviour or for attaining a goal)[105].
Temperament	The basic foundation of personality, usually assumed to be biologically determined and present early in life, including characteristics such as energy level, emotional responsiveness, demeanour, mood, response tempo and willingness to explore[105].
Working memory	A multi-compartment model of short-term memory that has a phonological (or articulatory) loop to retain verbal information, a visuospatial scratchpad to retain visual information and a central executive to deploy attention between them[105].

This glossary has been compiled from several sources as there was no single source that contained definitions of all of the non-cognitive constructs included in the systematic review. However, there are also inconsistent definitions across different sources. We reviewed various sources and selected explanations of non-cognitive abilities that were consistent with their usage in the literature included in this systematic review.

workforce. However, data on the effects of early life cognitive skills on these kinds of later life outcomes are very limited. These generative processes are thought to involve initial investments begetting skills that enable future skills, given sustained investments. Non-cognitive skills, such as being able to sustain attention, may be especially important in this regard because they can scaffold later development of cognitive and non-cognitive abilities. It is argued that, if these skills are not developed early in life, then it can be extremely difficult and expensive to compensate later in life, and this reduces returns on later investments[6].

Diversity of non-cognitive skills

Since 2000, there has been a 400% increase in publications using keywords describing various non-cognitive skills (Supplementary Fig. 1). Several constructs comprise the set of non-cognitive skills reflected in this literature, including academic motivation[7], responsibility and persistence[8], temperament, sociability and behaviour problems[9], locus of control and self-esteem[10], and attention and socio-emotional skills[11]. Executive functions[12] or cognitive control skills (for example, aspects of how children deploy their cognitive abilities through inhibitory control and attention) may be closely related to cognitive skills, but are also distinguished from IQ, literacy and numeracy[13]. Personality traits, such as self-esteem, patterns of thoughts, feelings and behaviours that include perseverance, motivation, self-control and conscientiousness, have also been considered as non-cognitive or quasi-cognitive characteristics[14]. The term 'character skills'[15] has been used to promote the potential malleability of non-cognitive skills in contrast to the notion of personality traits that are thought to be more stable. Heckman and Kautz label these as 'soft skills'[7]. Despite the conceptual complexity and potential overlap of some constructs, many different non-cognitive or personality or character or soft skills are represented in the literature. They have been the target of interventions, especially in early life when these traits are thought to be especially malleable[16], and for disadvantaged children, who may benefit most[6]. Interventions to improve non-cognitive skills may directly improve outcomes[7,15], or indirectly, through cognitive ability or other mechanisms. For instance, our own longitudinal analyses in three large population-based cohorts in the United Kingdom and Australia showed that both cognitive abilities and non-cognitive skills were important in explaining socioeconomic inequalities in academic achievement early in life and that non-cognitive skills were only weakly associated with cognitive ability[17].

Evidence for effects of early non-cognitive skills

Non-cognitive abilities have been associated with several shorter-term and longer-term outcomes, including mental health[18,19], physical health[20], school readiness and academic achievement[21,22], crime[23], employment and income[10], and mortality[24]. Evidence from RCTs suggests that preschool interventions that improve school readiness may do so in part by increasing children's ability to self-regulate their attention, emotion and behaviour[25]. Heckman has

Table 2 | Distribution of publications by outcome domain, study type and quality[a]

	Number of publications (%)	Outcome domains			
		Academic achievement	Psychosocial	Cognitive and language	Physical health
'Better' evidence	222/554 (40%)				
RCTs	41/222 (18%)	22	27	18	2
Quasi-experimental interventions	8/222 (4%)	4	5	5	1
Twin studies (longitudinal or cross-sectional)	12/222 (5%)	4	5	6	0
Observational longitudinal	127/222 (57%)	58	52	14	23
Observational cross-sectional	34/222 (15%)	14	19	9	5
'Weak' evidence	119/554 (21%)				
Observational longitudinal	73/119 (61%)	16	49	5	13
Observational cross-sectional	46/119 (39%)	20	28	1	3
'Poor' evidence	213/554 (38%)				
RCTs	1/213 (<1%)	0	0	1	0
Observational longitudinal	79/213 (37%)	25	46	6	15
Observational cross-sectional	123/213 (62%)	29	80	28	16

$n = 554$ publications. [a]Individual publications generated multiple outcomes. For example, there were 222 publications considered as 'better' evidence that examined 293 outcomes.

argued that interventions to develop these skills, especially in disadvantaged young children, have the potential for high rates of return due to their positive effects in multiple life domains[6].

It is widely accepted that children's cognitive ability (that is, intelligence or IQ) associates with academic achievement and later success in adulthood[26–29]. However, the HighScope Perry Preschool Program, which started in 1962, suggests other mechanisms may be involved[30,31]. The intervention provided an active learning programme based on Piagetian principles, for disadvantaged 3.5-year-old African-American children who had IQ scores on the Stanford Binet Test of <85 (ref. [32]). In analysing the long-term outcomes of the trial, Heckman et al.[31] reported that, although initially the intervention increased IQ, these increases were not maintained to 7–8 years of age. Despite this, children who received the intervention went on to enjoy more successful lives in adulthood, including greater labour market success, reduced crime involvement and better health[30,33,34]. Although we can find no evidence that the Perry Preschool Program deliberately set out to influence non-cognitive abilities, Heckman and colleagues argued that the intervention resulted in better outcomes for the participants not as a result of increasing their intelligence, but through fostering non-intelligence-based socio-emotional 'personality' skills[31] (p. 2,503). It should be noted that the programme also improved maths, reading and language through age 14 and adult literacy, so there may be an array of mechanisms operating through non-cognitive processes as well as IQ and/or aspects of academic achievement. Nevertheless, the argument proposed as to why the Perry Preschool Program 'worked' is not dissimilar to the observations of Bowles and Gintis[1] 40 years ago. They argued that schooling does not make children more intelligent, rather, it socializes them into, and rewards, certain characteristics and behaviours that are valued in the labour market.

The aim of this review was to systematically assess all published evidence concerning the effects of non-cognitive skills among children up to 12 years of age on later outcomes. We do not review intervention studies that did not specifically aim to improve non-cognitive skills. Thus, some interventions, such as the Perry Preschool[30] and Abecedarian[35] programmes, are not formally reviewed here because we could find no documented evidence that these programmes specifically set out to improve non-cognitive abilities and so were not eligible.

We screened eligible publications and report results on associations between non-cognitive skills up to 12 years of age. We grouped publications into four outcome domains—academic achievement (including literacy, numeracy and school readiness), cognitive and language development (including intelligence and language), psychosocial well-being (including mental health problems, such as internalizing and externalizing problems, hyperactivity, social skills and classroom behaviour) and health (including anthropometry and injury). In this paper, we only report results from those publications that we judged to be 'better' evidence derived from RCTs and quasi-experimental studies grouped as experimental and quasi-experimental intervention studies (EQIs), and observational studies that made reasonable attempts to control for confounding (endogeneity) bias. However, all eligible publications were fully reviewed and, for completeness, are presented in Supplementary Tables 7 and 8.

Results

The systematic search identified 9,553 articles from electronic and handsearched sources after removing duplicates. After assessing eligibility, 554 articles were included and presented in a PRISMA[36] flowchart (Fig. 1). There were 49 (9%) publications involving RCTs and non-randomized quasi-experimental interventions that reported 85 outcomes, 69% of which were in the academic achievement and psychosocial outcome domains (Table 2). Below, we report this group of studies as EQIs. Observational studies (including twin studies) accounted for the other 91% of all publications, also dominated by publications in the academic achievement and psychosocial outcome domains. Individual studies and publications may have reported multiple outcomes across the domains.

Table 2 shows that, of the 554 eligible studies, only 40% ($n = 222$) were rated as 'better' evidence, 21.5% classified as weak and 38.5% as poor, where there was effectively no attempt to control confounding. The better evidence category does not imply that all of the publications in this category would be considered 'strong' evidence in terms of their design and analysis. For example, some of the EQIs included in better evidence did not receive high-quality ratings according to the Risk of Bias Tool (Supplementary Table 6). We extracted and reported results separately for EQIs and observational publications included in the 222 better-quality evidence publications (Supplementary Tables 2–5). This information is summarized in Fig. 2 and Supplementary Figs. 2a–19b and 24–31,

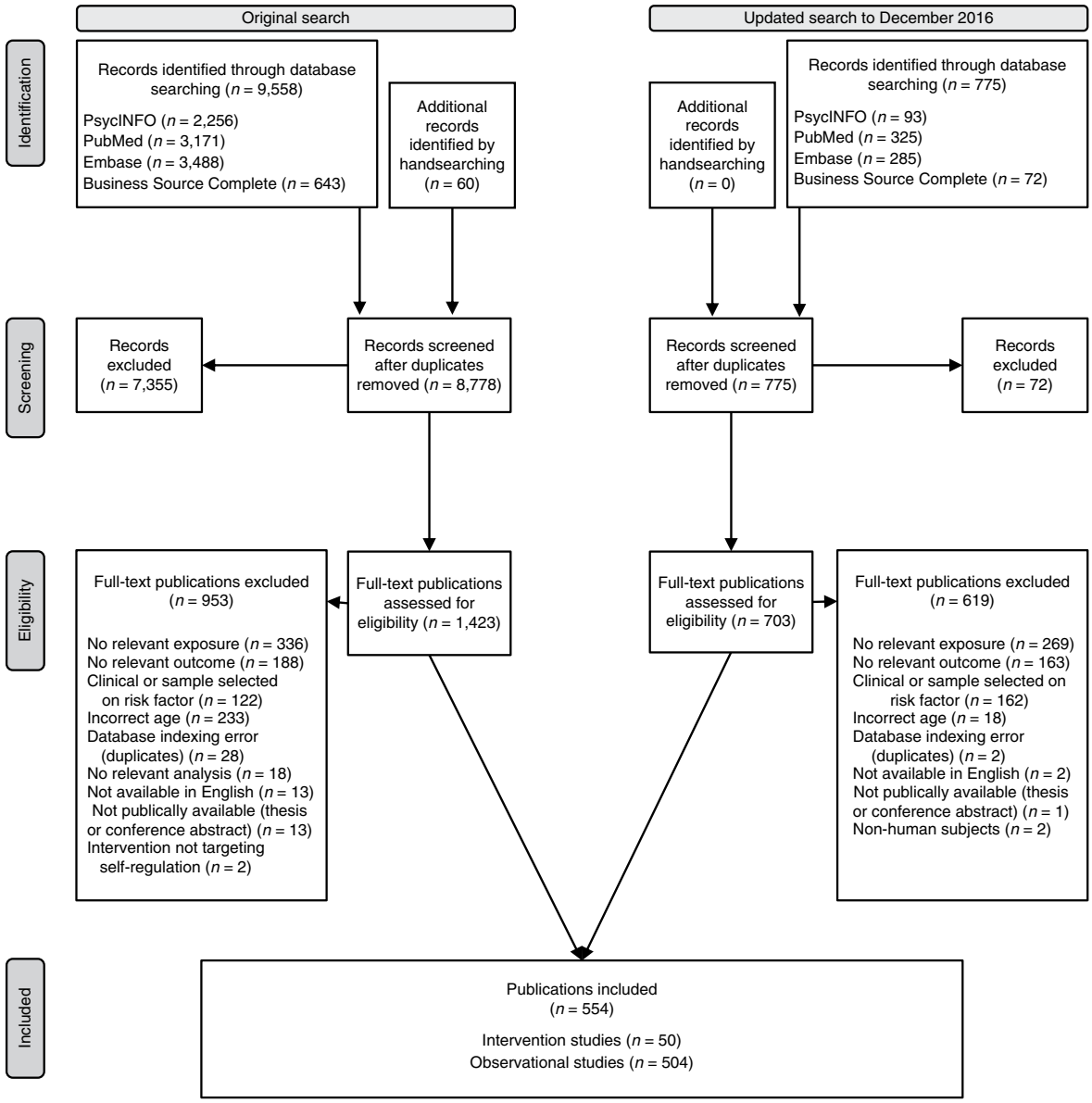

Fig. 1 | Flow of publications through different stages of the systematic review.

which display all studies in which an effect size and standard error could be calculated.

Academic achievement outcomes. Academic achievement outcomes mostly comprised reading, writing and numeracy and were most commonly measured by the Woodcock Johnson psychoeducational battery. For EQIs, Fig. 2a shows that effect sizes ranged from 0.16 s.d. (95% CI: −0.02 to 0.34) for academic achievement and school readiness to 0.37 s.d. (95% CI: 0.16–0.57) for numeracy. The 95% prediction interval for the 11 literacy studies available for meta-analysis was consistent with negative, null and positive effects (−0.13 to 0.79). For observational studies, Fig. 2b shows that effect sizes ranged from 0.16 s.d. (95% CI: 0.12–0.20) for literacy to 0.22 s.d. (95% CI: 0.14–0.31) for academic achievement and school readiness. Prediction intervals were consistent with negative, null and positive effects, ranging from −0.01 to 0.33 for literacy and −0.07 to 0.52 for school readiness. Details of these publications are presented in Supplementary Table 2. Meta-analysis and forest plots are presented in Supplementary Figs. 2a–4b. Supplementary Figs. 24 and 25 graph effect size, age and length of follow-up.

EQIs. There were 26 publications reporting 10 cluster (school or class) RCTs, 11 individual RCTs, 1 study in which the unit of randomization was unclear and 4 quasi-experimental intervention studies. These EQIs involved interventions delivered in usual preschool classes, special classes and groups additional to usual curriculum, at home or a combination of these. Interventions ranged from training specific abilities (for example, executive functions) to interventions that included several components. The interventions included teacher-delivered curriculum, teacher training to improve classroom behavioural management and training parents in game-based activities. There was about twice as many EQI publications concerning teacher-delivered curricula than EQIs including both parent and teacher components. The median age at the time of intervention was 4.5 years. The median follow-up time was under 1 year. The oldest age at follow-up was 20 years, from an intervention conducted in 1962, but no effect sizes were reported. The four largest cluster RCTs for literacy and numeracy ranged in effect sizes from 0.09 to 0.49 s.d. (Supplementary Figs. 2a–4b). The individually randomized trials were generally smaller and demonstrated effect sizes up to 0.81 s.d. but were more heterogeneous with a 95%

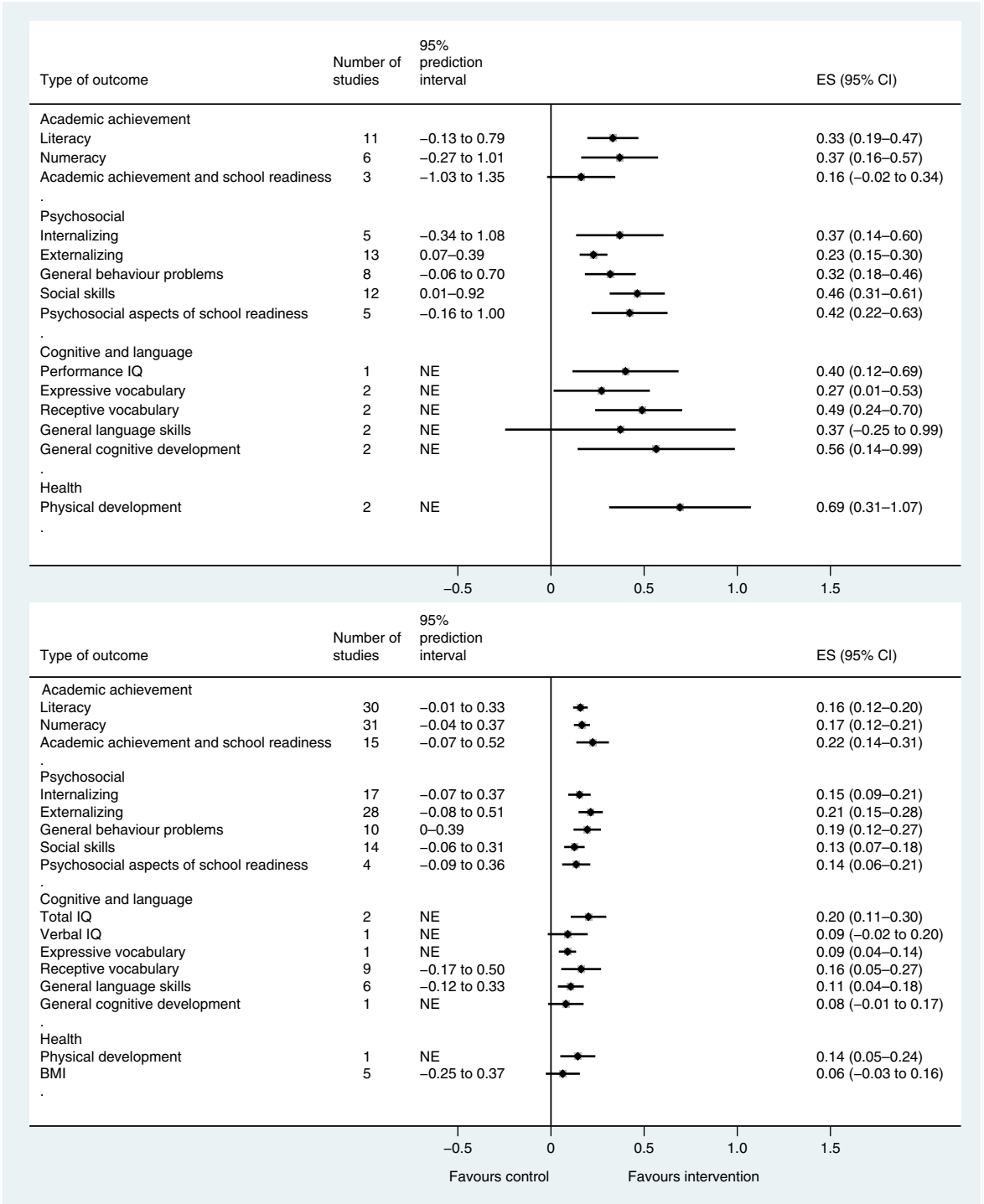

Fig. 2 | Effect sizes from studies presenting 'better-quality' evidence according to outcome. a, EQIs. **b**, Observational studies. BMI, body mass index. NE, not estimable. Effect sizes (ES) were calculated from random effects meta-analysis with inverse variance weighting (see Methods, Data synthesis).

prediction interval for literacy, consistent with negative and positive effects ranging from −0.91 to 1.79.

Observational. There were 4 publications of twin studies, 58 longitudinal (including 4 fixed-effects analysis) and 14 cross-sectional publications, with 3 publications reporting results from multiple cohort studies. Non-cognitive abilities were measured at a median age of 5.0 years and a median follow-up of 1.5 years. The oldest age at follow-up was 16 years. Study sizes ranged from 41 to 21,260. The measures of non-cognitive abilities included attention, executive function, inhibitory control, self-control, self-regulation and effortful control assessed via teacher report, parent report and objective tests, such as the Continuous Performance Task, Head Toes Knees Shoulders (HTKS) task and Stroop-like tasks. Effect sizes across observational publications were generally smaller than EQIs. Supplementary Figs. 2a–4b show effect sizes ranging from negative

effects (−0.57 s.d.), to null, to 0.77 s.d. for numeracy and similarly for literacy up to 0.80 s.d. However, 95% prediction intervals were generally narrower than for EQIs (for example, −0.04 to 0.37 for numeracy). There was little evidence to conclude that any one measurement tool, measurement method (objective or subjective) or underlying non-cognitive construct was consistently associated with academic achievement.

Psychosocial outcomes. Psychosocial outcomes included mental health problems (internalizing and externalizing behaviour), social skills and aspects of school readiness, such as learning engagement. For EQIs, Fig. 2a shows that effect sizes ranged from 0.23 s.d. (95% CI: 0.15–0.30) for externalizing behaviour to 0.46 s.d. (95% CI: 0.31–0.61) for social skills. For observational studies, Fig. 2b shows that effect sizes ranged from 0.13 s.d. (95% CI: 0.07–0.18) for social skills to 0.21 s.d. (95% CI: 0.15–0.28) for externalizing behaviour. The 95% prediction interval for psychosocial outcomes was consistent with negative, null and positive effects. For example, the 95% prediction interval for externalizing behaviour was −0.08 to 0.51. Details of these publications are presented in Supplementary Table 3. Meta-analysis and forest plots are presented in Supplementary Figs. 5a–9b. Supplementary Figs. 26 and 27 graph effect size, age and length of follow-up. Studies were not consistent in scoring of psychosocial outcomes, that is, higher scores could indicate worse or better functioning. To aid reader's interpretation of the results, we have converted all effects to be in the same direction so that positive effects indicate better psychosocial outcomes. However, Supplementary Table 3 presents the results as originally reported in individual publications.

EQIs. There were 32 publications reporting 15 cluster RCTs in classrooms, 12 individual RCTs and 5 quasi-experimental intervention studies in which the intervention was delivered in schools, sports classes, at home or in community-based settings. The content of the interventions was diverse and included teacher-delivered curriculum, sometimes specifically targeting self-regulatory abilities, parent–teacher engagement, teacher training to improve classroom behaviour, training parents in game-based activities, parental motivational interviewing and behaviour management, and martial arts. The median age at the time of intervention was 4.5 years with a median follow-up time of less than 1 year. The oldest age at follow-up was 13.5 years from a non-randomized intervention. Intervention groups ranged in size from $n=16$ to $n=314$ for the individually randomized trials and $n=20$ to $n=$~3,350 for cluster RCTs (the largest RCT did not report the exact intervention number). For externalizing outcomes, the 95% prediction interval for cluster RCTs was 0.10–0.37 s.d. and −0.15 to 0.61 s.d. for individual RCTs. Across RCTs, there was no consistent evidence favouring one mode of intervention delivery over another. The three largest cluster RCTs that trialled well-known interventions (PATHS, ParentCorps and Incredible Years) and had both a teacher and a parent engagement component[37–39] only reported effects in which $P \leq 0.05$ for 3 of the 11 outcomes studied.

Observational. There were 5 publications of twin studies, 52 longitudinal and 19 cross-sectional publications. The 5 reasonably sized twin studies that combined monozygotic and dizygotic twins (n ranged from 209 to 410 pairs) of children ~2–8 years of age reported phenotypic correlations between non-cognitive abilities and internalizing problems of 0 to −0.3 and −0.1 to −0.6 for fewer externalizing problems. The longitudinal studies ranged in size from 49 to 12,158, and cross-sectional studies ranged from 42 to 2,978. Non-cognitive skills were measured at a median age of 5.0 years with a median follow-up of 8.2 years. The oldest age at follow-up was 19.5 years. Exposures included attention, executive function, inhibitory control, self-regulation, emotion regulation, delay of gratification, effortful control, impulsivity, self-control and temperament, and were assessed by teacher report, parent report and objective tests. Supplementary Figs. 5a–9b show effects from observational studies consistent with ~0.1–0.2 s.d., but all 95% prediction intervals included the null.

Observational studies of psychosocial outcomes were the most heterogeneous in terms of measuring exposures and outcomes, complicating interpretations of overall effect estimates. There was little evidence that attention, executive function and delay of gratification affected psychosocial outcomes. For inhibitory control, self-regulation, emotional regulation, impulsivity, self-control and temperament, there was some evidence of effects (0.1–0.7 s.d.) on social skills and mental health problems. For effortful control, evidence was mixed, ranging from null to 0.85 s.d. on externalizing behaviour.

Cognitive and language outcomes. Cognitive and language outcomes were typically assessed by measures of overall intelligence (such as the Wechsler suite of intelligence tests), verbal and performance intelligence and language development, including expressive and receptive vocabulary (such as the Peabody Picture Vocabulary Test). For EQIs, Fig. 2a shows that the effect sizes ranged from 0.27 s.d. (95% CI: 0.01–0.53) for expressive vocabulary to 0.56 s.d. (95% CI: 0.14–0.99) for general cognitive development. No 95% prediction intervals could be calculated as there were fewer than three studies in each subdomain. For observational studies, Fig. 2b shows that effect sizes ranged from 0.08 s.d. (95% CI: −0.01 to 0.17) for general cognitive development to 0.20 s.d. (95% CI: 0.11–0.30) for total IQ. The 95% prediction interval could only be calculated for receptive vocabulary (−0.17 to 0.50) and general language skills (−0.12 to 0.33). Details of these publications are presented in Supplementary Table 4. Meta-analysis and forest plots are presented in Supplementary Figs. 10a–16b. Supplementary Figs. 28 and 29 graph effect size, age and length of follow-up.

EQIs. There were 23 publications reporting 18 RCTs (2 interventions were reported in 6 publications) and 5 quasi-experimental intervention studies. Of the RCTs, 6 were cluster (school or class) RCTs, 1 in which the unit of randomization was unclear and 11 individual RCTs, involving programmes delivered in schools or classrooms, at home, in a laboratory setting or a combination of classes and home. Three quasi-experimental interventions involved preschool programmes and two involved computerized working memory and inhibitory control training. The content of the interventions was diverse in both delivery and specific focus on non-cognitive ability. Interventions ranged from narrow-focused computer-based training to broader content and delivery by teachers in schools plus home visiting with parents. The median age at intervention was 4.3 years, with a median follow-up of less than 1 year, extending to 16 years. One RCT inconsistently reported effects of 0.15 and 0.25 s.d. on the same language outcome, using the same sample at 5 years of age[40,41] and an effect of 0.10 s.d. at 6 years of age in a different publication[42].

Observational. There were 6 publications of twin studies, 14 longitudinal (including one fixed effect) and 9 cross-sectional publications. The 6 twin studies that combined monozygotic and dizygotic twins (n ranged from 40 to 901 pairs) reported phenotypic correlations between non-cognitive abilities and intelligence of −0.36 to 0.23 s.d. The longitudinal and cross-sectional publications ranged in effect size from −0.38 s.d. (cross-sectional convenience sample: $n=77$ examining attention) to 0.56 s.d. (cross-sectional convenience sample: $n=80$ examining executive function). Exposure was measured at a median age of 4.5 years. The median duration of follow-up for the longitudinal studies was less than 1 year and the longest follow-up was to 12.4 years. Exposures included attention, executive function, self-regulation, effortful control, inhibitory control and

temperament, assessed via parent and teacher report questionnaires, such as the Child Behaviour Questionnaire, and objective tests, such as the Continuous Performance Task and the HTKS task. There was no compelling evidence of effects of attention on cognitive and language outcomes from observational studies. For executive function, effects ranged from a detrimental −0.36 to 0.52 s.d., but the evidence is predominantly from convenience samples. There were too few studies to make any judgements about the effects of effortful control and temperament. Most studies of self-regulation used the HTKS task and showed some effects on vocabulary.

Health outcomes. There were 2 small RCTs, 1 quasi-experimental intervention, 23 longitudinal and five cross-sectional publications that ranged in size from 105 to >26,000. Details of these publications are presented in Supplementary Table 5. Meta-analysis and forest plots are presented in Supplementary Figs. 17a–19b. Outcomes included anthropometry, injury, diet, substance use and health behaviours.

EQIs. There were three publications reporting one cluster RCT, one individual RCT and one quasi-experimental intervention study assessing effects on physical development, teen parenthood and anthropometry. One quasi-experimental study reported an effect of 0.79 s.d., but this effect is difficult to interpret because of an inadequate description of the control group and the outcome. The median age at intervention was 4.4 years. The median follow-up time was less than 1 year, with the oldest age at follow-up of 17 years.

Observational. Of the observational studies, the median age at exposure was 9.3 years. The median follow-up time was 4.2 years and the oldest age at follow-up was 55 years. Of the 28 observational studies, 12 involved various outcomes related to substance use, but it is difficult to summarize these because studies either did not report effect sizes or reported unstandardized effects or odds ratios. Observational studies showed little evidence for associations with any of these outcomes.

Assessment of small study (publication) bias. The funnel plots in Supplementary Figs. 20a–23b depict effect sizes for experimental and observational studies separately, according to the standard error of the effect size. These include all publications in which effect sizes were reported or able to be calculated, and reported exact P values or $P < 0.05$ (ref. [43]). Thus, all studies that reported a P value greater than some threshold were excluded. Funnel plots for both experimental and observational studies were positively skewed and consistent with smaller studies having larger effects. Egger regression coefficient P values were all $P < 0.01$. There was little evidence for differential small study bias comparing EQIs and observational studies.

Fade out. Supplementary Fig. 32 attempted to examine fade-out effects[44] by graphing reported effects at the end of intervention (or as close to end line as was reported) and at later follow-up. There were only four studies that could be included in this analysis, so interpretive caution is warranted with no clear pattern to support evidence of fade-out effects.

Discussion

We reviewed 554 publications and provided interpretation of 222 (40%) better-quality publications comprising RCTs, quasi-experimental (EQIs), fixed effects (including twin studies), longitudinal and some cross-sectional designs (observational studies). We set out to systematically examine the published literature on the effects of non-cognitive skills up to 12 years of age on outcomes as they have been presented in the literature. We put no time limit on when outcomes were measured and we grouped them in domains of academic achievement, psychosocial, cognitive and language, and health. This review can say little about longer-term effects that are of central policy interest, such as the effects of non-cognitive skills on labour market experience, because studies eligible for this review do not have data on longer-term outcomes or do not report it. Nor can this review say anything about the importance of non-cognitive skills on later outcomes that are developed as part of normal social interaction and/or the hidden curriculum of more general interventions in which children indirectly develop various non-cognitive skills and behavioural styles.

We were limited to reporting what might be termed 'proxy' or 'intermediate' outcomes. Although outcomes such as academic achievement are clearly related to employment and labour market experience, this review cannot directly inform the role of non-cognitive ability on important outcomes later in life. Despite the policy enthusiasm and discussion of the importance of non-cognitive skills, the current body of evidence is severely limited given median follow-up periods for EQIs of only about 1 year. We must search elsewhere for evidence on longer-term outcomes because it is precisely in the realm of the labour market that non-cognitive skills may be most beneficial and rewarded.

Overall, there is evidence from published EQIs supporting a role for non-cognitive skills in better academic achievement, psychosocial, and cognitive and language outcomes ranging from approximately 0.2 to 0.5 s.d. depending on outcome as shown in Fig. 2a. We urge some caution in interpreting our results. Analysis of funnel plots clearly demonstrates asymmetry of effect size and the potential of small study bias[43]. In addition, forest plots and 95% prediction intervals show large heterogeneity of reported effect sizes generally including the null. This suggests that the overall meta-analysed effects from EQIs reported here may be overestimates that include a null effect.

Presenting the analysis in Fig. 2 by separating EQIs (Fig. 2a) and observational publications (Fig. 2b) shows larger effects from EQIs than found in higher-quality observational studies, which ranged from approximately 0.06 to 0.22 s.d. This is the opposite of what is often seen, in which observational studies overestimate effects found in large, well-designed RCTs. This overestimation is often due to residual and/or unmeasured confounding introduced by using observations of exposures rather than experimental manipulation of exposures[45]. Furthermore (as pointed out by a reviewer), the effect sizes from EQIs and observational studies would only be comparable if the EQI induced a s.d. change in the particular non-cognitive skill. In reality, effects of interventions on the target non-cognitive skill might be closer to 0.2–0.5 s.d. So, at 0.25 and with no bias, effects found in observational data would be expected to be four-times larger than experimental effects.

Franco et al.[46] found that among rigorously reviewed social science publications in the Time-Sharing in the Social Sciences National Sciences Foundation database, 'strong' results were 40 percentage points more likely to be published than null results and 60 percentage points more likely to be written up. They argued that this provided direct evidence of publication bias when researchers choose which results should be written up and presented for publication. It is possible that the published EQIs favour stronger statistically significant results if these are selected by researchers based on P values. If the published EQIs are dominated by smaller studies with lower power, the overall EQI evidence may provide inflated meta-analysed effect estimates. However, we found little evidence of differences in potential small study and publication bias between EQIs and observational studies. Nevertheless, in academic achievement and psychosocial outcome domains, larger sample cluster RCTs tended to generate smaller effects than individually randomized small RCTs. A recent meta-analysis of observational studies of over 14,000 children[47] showed a mean effect size of 0.27 for inhibitory control on academic achievement.

However, this meta-analysis did not exclude poor-quality studies and did not explore potential for small study bias. We deliberately selected higher-quality observational studies with more stringent controls for confounding, so it is possible that true effects of non-cognitive skills are actually closer to those from higher-quality observational studies that may include a null effect.

Main findings. *Academic achievement outcomes.* Intervention studies focused on improving children's non-cognitive skills at around 4 years of age with a median follow-up of under 1 year. These studies were generally consistent with 0.2–0.4 s.d. short-term effects on academic achievement, but effects were heterogeneous with 95% prediction intervals including negative, null and positive effects. Larger, higher-quality RCTs showed effects from 0 to 0.3 s.d.[25,48–50]. These larger, higher-quality RCTs spanned child-focused interventions on specific domains of non-cognitive skills (for example, Tools of the Mind), to more teacher-focused curricula (for example, Chicago School Readiness), to more multidimensional content interventions that included parent, child and teacher (for example, PATHS). Observational studies on academic achievement generally showed effects around 0.2 s.d., but all 95% prediction intervals included the null. This is consistent with one higher-quality observational publication[11] that examined six different cohorts with a longer follow-up of 5.5 years and reported effects from 0 to 0.2 s.d. Overall, there was insufficient evidence on which to base a conclusion about the relative effectiveness of different modes and mechanisms of intervention on non-cognitive skills. Even within the same study, effect sizes differed according to which aspects of academic achievement were measured. For example, one RCT showed an effect on numeracy but not literacy[49]. Similarly, another RCT showed that effects on literacy depended on the component of literacy that was measured[40,41] and effects on some outcomes faded after 1 year[42].

Psychosocial outcomes. For psychosocial outcomes, the evidence from RCTs was dominated by studies of externalizing problems, with fewer RCTs on social skills and internalizing problems. The average age at the time of intervention was around 4 years, with a median follow-up time under 1 year. Effects on externalizing problems for EQIs was 0.23 (95% CI: 0.15–0.30), with a 95% prediction interval of 0.07–0.39. Higher-quality RCTs that examined externalizing outcomes reported positive[51,52] and null effects in the largest of the RCTs[37]. These variable effects could be due to differences in the focus of intervention, mode of delivery (parent, teacher or both) or problems with implementation fidelity in larger trials. Similarly, inconsistent results were reported for EQIs with social skills outcomes. The heterogeneity of effects is mirrored in the twin, longitudinal and cross-sectional studies. A good example of this is the inconsistent results reported in five publications that all used the same data source[53–57]. Across these five publications, interpretation of the effects of self-regulatory abilities depended on how the exposure (attention, delayed gratification and inhibitory control) and outcome (social skills, withdrawal and aggression) were measured. The different measures of attention had different associations with the same social skills outcome. Inhibitory control was associated with social skills and aggression but not social withdrawal, whereas the effects of delayed gratification on social skills depended on whether the outcomes were directly observed or from maternal report.

The psychosocial outcome studies were the most diverse in interventions (ranging from martial arts to motivational interviewing and Tools of the Mind) and exposure and outcome measurement. This diversity reflected different approaches to improving children's psychosocial outcomes, such as supporting parents or helping teachers to manage classroom behaviour. Each approach points to different conceptualizations of where psychosocial problems arise and for how, where and whom to intervene (for example, teachers, psychologists, community nurses or social workers).

Cognitive and language outcomes. The relatively small number of studies in this outcome domain ($n = 23$) produced a wide range of effects. Three reasonably sized cluster RCTs provided the best estimate of the effect of non-cognitive skills on language and cognitive outcomes[25,49]. They found effects of ~0.1–0.2 s.d. The largest effect sizes were from a well-designed regression discontinuity study (0.44 s.d.)[58], a non-randomized intervention (0.55–0.73 s.d.)[59] and a small, low-quality randomized trial[60]. However, all of these studies were small (range: $n = 12$–64) and reported effects that attenuated over time or were inconsistent at different ages. The observational studies provide little evidence that the effects are likely to be bigger than ~0.1 s.d, with 7 of 9 longitudinal studies showing few differences and cross-sectional studies reporting mixed effects (−0.38 to 0.56 s.d.). The longitudinal studies were dominated by non-cognitive skills measured using the HTKS and the Woodcock Johnson Picture Vocabulary as the outcome, and despite the popularity of these measurement tools, the results indicate no effects on vocabulary outcomes. Thus, non-cognitive abilities seem to have effects on cognitive and language outcomes of $\leq$0.2 s.d.

Physical health outcomes. It is difficult to draw conclusions for physical health outcomes. There were only three EQIs reporting diverse outcomes. Outcomes reported across the 28 better-quality observational studies were diverse, ranging from anthropometry, to injury to physiological characteristics and were consistent with effects ranging from 0.06 to 0.14 s.d.

Limitations of this review. The compilation of 554 publications was systematic, but our assessment of the quality of the evidence is based on our judgement of the potential for bias. Here, we follow the approach of others who have argued for limiting systematic reviews and meta-analyses to higher-quality evidence[61,62]. We a priori created criteria for bias based on well-established procedures, including quality appraisal tools, evidence hierarchies, directed acyclic graphs and content knowledge about potential sources of confounding and selection bias. Although this involves an element of subjective judgement, we are confident that any other reasonable assessment of the quality of evidence would not change the overall conclusions presented here. In the interests of transparency, we have disclosed all of the subjective choices that we have made in the Supplementary materials and text.

It is possible that some relevant articles were not included in this review, even though we undertook an extensive search that included multiple databases, numerous search terms, contacting authors of potentially relevant papers and handsearching reference lists of published papers. Studies of systematic review methods have shown that the most difficult to find articles are in the 'grey literature', sometimes smaller, of poorer quality and the results unlikely to unduly influence the findings in an already large systematic review[62].

The value of a systematic review. Although there have been reviews of some aspects of non-cognitive skills[3,4,14,15,47,63], none has been systematic in covering the entire literature or included screening for evidence quality. It has long been recognized in health and medical research that non-systematic reviews of research enable the selective use of evidence to support a particular argument[62]. For evidence consumers, who are often not evidence-quality specialists, competing claims about effects of non-cognitive abilities based on particular studies are hard to reconcile without the safety net of a systematic review. We have paid particular attention in this review to issues of quality of the primary evidence. There is little point in summarizing evidence that includes obviously flawed studies that can only distort the overall results and reduce the value of the systematic nature of the review[61,62,64].

This review covers the entire published inter-disciplinary research field, describing intentional efforts and observational analogues

of interventions to improve the development of non-cognitive skills, albeit with most evidence coming from rich countries, especially the United States. The scope of the review should minimize 'cherry picking' of results to bolster a particular concept, theory or intervention. This is necessary to advance knowledge given the multidisciplinary nature of this field and is central to informing interventions to boost life chances for disadvantaged children. In health sciences, major advances have been made by coming to agreement and attempting, where possible, to harmonize methods for measuring exposures, outcomes, synthesizing and reporting of outcomes. This work includes collaborative efforts such as the EQUATOR network (http://www.equator-network.org/). Such efforts are needed to reduce waste in research[64,65] and improve reproducibility of scientific findings[66–68].

Implications for future research. *What are the active ingredients of non-cognitive skills?.* Research that has examined non-cognitive skills in childhood has spanned many disciplines and research traditions, leading to a large number of constructs and tasks being investigated that are sometimes similar in their definition and operationalization[69,70]. In 1927, Kelley labelled this the 'jangle' fallacy[71] (p. 64) in which constructs are given different names but are in fact virtually identical. This idea has been recently raised in regard to the construct validity of the concept of 'grit'[72]. It was not uncommon for the same objective tasks to be used as measures of different conceptualizations of non-cognitive abilities. For example, the Continuous Performance Task (also known as the 'Go/No Go' task) is used in executive functioning research as a measure of sustained attention and inhibitory control, but it is also used as a measure of effortful control[70]. Similarly, the HTKS task has been used to measure both behavioural self-regulation[73] and executive functioning[74]. The interventions that we reviewed attempted to influence many different facets of non-cognitive skills. Policymakers and researchers ideally need to know what the 'active ingredients' are, to enhance children's non-cognitive skills and, ultimately, the relative effectiveness of different interventions and different intervention doses. There are no strong scientific reasons to favour a specific skill over another, but nevertheless, it remains important to better understand what the active ingredient(s) underlying non-cognitive skills might be, if we want to support their development.

Mechanisms of action. Theoretically, we might expect that interventions involving both parents and teachers might have larger effects on children's outcomes. However, a recent meta-analysis of early childhood education programmes found little evidence that those with parenting involvement produced larger effects, unless it involved a high dose of home visits[75]. Of the academic outcomes reviewed here, over half involved only preschool teachers. In our review, there is little evidence to decide which mode of delivery is best and we can find no evidence of attempts at purposive testing of which way to intervene (for example, teacher, student, parent or various combinations). Purposive testing of delivery mode has been usefully deployed in the design of an RCT in regard to the nurse home-visiting literature, showing better effects using trained nurses than using para-professionals[76]. Interventions that trained children in more specific skills, such as executive function, generally showed small effects (for example, Tools of the Mind)[49]. Other studies imply that non-specific interventions seem to generate better generalized outcomes[31], which may suggest that more holistic programmes, including multidimensional content, may better support overall child development and broad-based benefits.

Head-to-head comparisons of interventions. Comparative effectiveness research has been widely promoted in health and medical science as an important contribution to knowing which treatments are the most effective[77,78]. The potential for interventions on non-cognitive skills to influence outcomes may be enhanced by similar approaches. We could find almost no evidence of these sorts of purposeful comparative studies in this field. Exceptions were Barnett et al.[49] and Blair and Raver[79] who examined effects of Tools of the Mind intervention in cluster RCTs and both found effects of ~0.1 s.d. for vocabulary. This exception highlights the potential value of these comparisons.

Designing studies for effect modification. In assessing the potential importance of non-cognitive skills for improving life chances, it is obvious that a combination of both high cognitive and non-cognitive ability would be desirable. If that expectation is correct, then the effects of interest lie in a test of effect measure modification or interaction depending on what effect is of interest[80]. We found no publications attempting to test this theory, despite its obvious importance for judging how non-cognitive skills might influence later life outcomes. It is also of interest to test for differential effects of developing non-cognitive skills according to different characteristics of children, such as age or socioeconomic background, and of intervention type and setting. However, we urge some caution in investigating differential effects in subgroups when the basic evidence for effects of non-cognitive skills on outcomes, such as academic achievement and psychosocial outcomes, is already highly heterogeneous and consistent with null effects.

Long-term follow-up. There is a paucity of literature with long-term follow-up. Studies typically began at 4–5 years of age, with median follow-up of about 1 year, and with very few studies with follow-up beyond 10 years of age, there is very little evidence addressing effects on medium-to-longer-term outcomes. This is no doubt owing to funding constraints. However, it is frequently argued that non-cognitive skills developed in childhood have major effects on long-term adult outcomes[6]. Thus, interventions that have short-term effects but few detectable long-term effects are unlikely to be cost-effective. Thus, longer-term follow-up of RCTs is especially important and is being supported by several funding agencies in education and elsewhere. Nevertheless, until such longer-term studies are reported, many of the claims in the literature that early interventions on a specific trait or with a particular intervention have major long-term effects are supported by very little empirical evidence.

Fade out. Recent concerns about the fade out of initially promising effects is crucial to consider in regard to the likely value of interventions early in life. Bailey et al.[44] argue that interventions should target what they term 'trifecta skills' (p. 8). These skills are malleable, fundamental and would not have developed eventually in the absence of the intervention. There were only four studies in which we could assess evidence for fade out and results were inconclusive. This seems another important facet to develop within the research portfolio around non-cognitive skills; studies could be specifically designed to test the fade-out hypothesis in rigorous ways.

Small study effects. Larger effects observed in smaller RCTs may be due to several factors, including publication bias favouring positive results, true heterogeneity of effects due to differing baseline risks in different intervention populations, implementation difficulties in maintaining intervention intensity and fidelity in larger community settings, and poorer methodological design of smaller studies[81]. If smaller studies were better able to implement the intervention, then larger effects might be real owing to greater fidelity to the intervention as designed. Conversely, publication bias favouring more positive results would mean that larger effects from smaller studies would bias true effects upward. This is an important issue for practice and policy, as it suggests that effects found in RCTs of small convenience samples may be overestimated or even non-existent. For example, when studies are scaled-up, the results can be inconsistent

or attenuate towards the null, perhaps suggesting that fidelity is harder because an expert is no longer delivering the intervention and/or that larger-scale studies are unable to deliver as intensive interventions as small studies. A useful framework for considering such variation in intervention effects across different scales, contexts and population groups is presented in Weiss et al.[82].

Heterogeneity of effects. This review clearly demonstrated large between-study heterogeneity from 95% prediction intervals that were consistent with negative, null and positive effects among subdomains, such as literacy and numeracy. It is possible to argue that this was inevitable in a field in which there are many dimensions of non-cognitive skills being investigated in largely convenience samples, against a wide variety of measures of broad constructs such as literacy and numeracy. Perhaps that is so, but the field is nevertheless presented somewhat monolithically in the application of this science to broad-scale intervention and policy practice[3]. Quantifying the amount of heterogeneity is valuable in providing a baseline from which future research can investigate potential sources of this heterogeneity. For instance, we sought to examine whether studies that used more representative population-based samples tended to generate smaller effect estimates, but the number of population-based samples in this field is actually rather small. For example, of the 11 literacy EQIs that were able to be included in the meta-analysis, only 3 were population based. For externalizing behaviours, of 13 EQIs, only 1 was in a population-based sample.

Evidence quality. *Citing practices.* We reviewed recent RCTs to count the number of previous RCTs they cited. There were seven RCTs published from 2014 to 2016. There were 27 previous RCTs of non-cognitive skills on academic achievement and psychosocial outcome domains available to be cited. The highest number of citations of previous RCTs referenced in any of the RCTs published from 2014 to 2016 was four[79]. Several RCTs published between 2014 and 2016 referenced no previous RCTs. It could perhaps be argued that these RCTs intervened on different non-cognitive skills, so should not necessarily cite studies of other non-cognitive skills. Nevertheless, attention regulation and self-control were common ingredients of almost all of these interventions (Supplementary Tables 2–5), so the impression is that new RCTs were not being explicitly justified on the basis of what was already known from existing RCTs.

Quality of RCTs. The quality of RCTs was not ideal and reporting of some details was poor or even absent. No RCTs had a formal pre-registered protocol and two-thirds did not explicitly identify primary outcomes (See Supplementary Table 6 on Risk of Bias Tool[83]). This can allow cherry picking of significant results within studies rather than focus on a single or small number of pre-stated main outcome(s) that the intervention is theoretically or empirically (based on previous evidence) meant to most influence[84]. Over one-quarter of RCTs may have had other potential biases, for example, differential participation in the control and intervention groups, and unclear processes for selection of control participants. Ninety-two per cent of RCTs did not adequately report randomization procedures, 81% did not report concealment of allocation processes and participant flow, and most failed to address missing data. It was common for cluster RCTs to have too few clusters to achieve balance between intervention and control groups, and in some, it was unclear whether clustering was adequately dealt with in the analysis[85]. Poor reporting made it difficult to fully assess study quality, and we strongly encourage researchers, journal editors and reviewers to use tools such as the CONSORT statement (http://www.consort-statement.org/) for reporting and for RCTs to be pre-registered. These are now mandatory requirements for publishing in most leading health and medical science journals. However, it is possible that research practice regarding pre-registration is already changing and those pre-registered studies are yet to be published.

Quality of observational studies. More than 90% of all research in this field comes from observational studies. Of the 504 observational studies reviewed here, 66% were judged as 'weak' or 'poor' quality. Of all observational studies, 42% made little or no attempt to adjust for even basic confounding, that is, common causes of non-cognitive ability and the outcome. Problems of endogeneity and confounding are well known and may result in substantial bias of the association of non-cognitive skills and later outcomes.

One regrettable consequence of the relatively low quality of much of the research effort in this field is that it is not able to shed much light on the question of whether improving non-cognitive skills positively influences outcomes. To advance understanding of non-cognitive skills in children and their effects on outcomes later in life, there is little point in amassing more small-scale[86], biased observational or experimental studies that have a higher likelihood of failing to be replicated[65,66,87] and are unable to contribute to evidence triangulation, which is central for stronger causal inference[88,89]. The recommendations that we make here to improve evidence quality in this field are not controversial. A 2018 *Annual Review of Psychology* paper called for more sophisticated power analyses, better statistical practices, study design specific to addressing effect modification and better disclosure of non-significant as well as significant findings[90].

Implications of suboptimal reporting practices of effect sizes and *P* values. To be included in a meta-analysis, studies needed to report or have the information available to calculate an effect size and the standard error. When standard errors were not available, we calculated the standard error from an exact P value, or when the P value was reported as $P<p$, we assumed that $P=p$. We were unable to calculate effect sizes in several cases, and in others, P values were reported as $P>p$. Consistent with recommended practice, this meant that studies were excluded when an effect size and/or a standard error could not be calculated[91]. Excluding studies that reported $P>p$ provides a more conservative estimate of the precision of studies. These exclusions were on top of excluding studies in which an effect size was either not reported or could not be calculated. We illustrate the effect of this two-layer exclusion for literacy outcomes. The literature reported 49 literacy-related outcomes in 17 EQIs. Excluding outcomes in which an effect size could not be calculated reduced the number of available literacy outcomes to 42 outcomes from 14 EQIs. Further excluding results in which the P value was reported as $P>p$ meant that the meta-analysis and funnel plots could only include 33 literacy outcomes from 11 EQIs. Thus, this two-layer exclusion of reported results (owing to suboptimal reporting practices) meant that we could only include 67% of the literacy outcomes actually presented in the literature. This also meant that the meta-analysed effect size for literacy increased from 0.22 (including studies with $P>p$) to 0.33 (excluding studies with $P>p$) for EQIs because of the exclusion of studies with smaller effect sizes.

Interpreting effect sizes. We have avoided labelling effect sizes as 'small (~0.2 s.d.)', 'medium (~0.5 s.d.)' or 'large (~0.8 s.d.)' according to Cohen's suggestions[92]. Even though these metrics are widely, often ritualistically, used as reference points, Cohen did not intend them to be used as absolutes. He cautioned that such generic application of sizes of effects to all research fields was "an operation fraught with many dangers"[92] (p. 12). Deciding whether an effect is 'big' is not straightforward in any field. Effect sizes are nothing more than mean differences between intervention (exposed) and control (unexposed) groups on some scale of outcome measurement divided by the standard deviation of the outcome. The use of such standardized effect measures has been criticized in several

disciplines. In epidemiology, Greenland et al.[93] have argued that the process of standardizing effects, rather than making them more comparable across studies, simply serves to confound that comparison by the observed standard deviation, which is often an artefact of the study sample, particularly for homogenous convenience samples. In political science, King[94] argued that if apples and oranges cannot be meaningfully compared on the original outcome measurement scale, then this lack of comparability is not improved by comparing standardized fruit. The size of effects must be judged within the context of the field, the methods used in the study[95] and, importantly, linked to some normative understanding of what weak or strong effects look like in a particular field. For example, if the best interventions available to improve a particular outcome found reliable effects of 0.2 s.d. when trialled in large population-based samples, then a novel intervention finding the same effect might be considered large. Another way of norming effect size may be to consider the size of intervention effects against secular change in an outcome over time. Lipsey et. al.[96] present a sophisticated understanding of interpreting effect sizes. For example, they show that the secular growth in reading from kindergarten to grade one in the United States is estimated to be about 1.5 s.d. By grades 4–5, this growth has declined to about 0.4 s.d. per year. How should an effect of a non-cognitive skills intervention in kindergarten on reading in grade one of 0.2 s.d. be judged? Such an intervention has generated about 13% greater improvement than the natural growth in reading during that time. Deciding whether an intervention is worth implementing will depend not only on its benefits but also on its costs, discount rate, scalability and a range of other potential considerations. Interventions that have small effects on average across the population and that are cheap could be very cost-effective, particularly if they influence long-term outcomes in adulthood. Thus, the traditional labelling of an intervention as having 'small' effects (~0.2 s.d.) is inappropriate because it fails to consider the research, policy and practice context within which the intervention is situated.

Conclusion

So, after all the voluminous research included in this systematic review and meta-analysis, do intentional (from EQI evidence) or implied (from observational evidence) efforts to improve early life non-cognitive skills influence outcomes? Overall, yes, there is some evidence supporting a role for non-cognitive skills in better academic achievement, psychosocial, cognitive and language domains, but these effects are highly heterogeneous as they relate to the shorter-term outcomes examined in this review.

We urge caution in interpreting this overall finding as unequivocally positive, given the potential for small study (publication) bias that may overestimate the true effects and the underlying heterogeneity of effect estimates, as shown in 95% prediction intervals that were generally consistent with negative, null and positive effects. Thus, a true null effect of non-cognitive skills on these outcomes cannot be ruled out. We urgently need more robust evidence about which skills may be the active ingredient(s) and which outcomes they affect in the longer term. That may come from studies that are funded for long-term follow-up of some of the more promising interventions reviewed here. These results suggest profitable pathways forward to help to improve influences on life success beyond the traditional focus on reading, writing and arithmetic, and IQ. However, the research community interested in these diverse aspects of non-cognitive skills needs higher-quality, adequately powered studies and a strategically integrated, rigorous scientific focus to help to answer the policy-relevant questions[97].

Methods

The systematic review protocol was pre-registered with the International Prospective Register for Systematic Reviews (PROSPERO; CRD42013006566) in December 2013 and is available at: http://www.crd.york.ac.uk/PROSPERO/. This original protocol included children up to 8 years of age. Reviewers suggested extending this to 12 years of age, hence the protocol was updated in September 2017.

Inclusion criteria. Publications were eligible if they involved non-cognitive abilities of children up to 12 years of age, including executive function (working memory, cognitive flexibility, inhibitory control and attention), effortful control, emotional regulation (emotional reactivity), persistence, conscientiousness, attention, self-control, impulsivity and delay of gratification. See Table 1 for a glossary of terms. Interventions that had general developmental goals were included if they specifically stated an aim related to improving any non-cognitive abilities. Only publications that reported original research were included. Publications involving non-cognitive characteristics in clinical subgroups (for example, those already diagnosed with problems such as attention-deficit/hyperactivity disorder) were excluded because we were interested in the effects of non-cognitive characteristics among developmentally normal healthy children.

Literature search. We searched four electronic databases for articles published from database conception until December 2016: PubMed, PsycINFO, Embase and Business Source Complete. These databases were chosen because of their broad coverage of psychological, education, health and economic literature. The search strategy for each database is included in Supplementary Table 1. Search terms were tailored to each database and pilot tested. Study outcomes were not included as search terms to capture all published outcomes associated with non-cognitive abilities. Searches were not restricted by language. Authors of non-English articles were contacted for details or translations. Authors of conference abstracts, editorials and theses were contacted to obtain full-text articles. Handsearching of relevant reviews[16,98–100], our own libraries and references cited in all RCTs and quasi-experimental interventions were conducted to identify further studies.

Screening. The titles and abstracts of all articles were screened for eligibility (by A.C.P.S., L.G.S., C.R.C. and T. Nuske). To ensure consistency of searching, the first 300 references were searched as a group by all authors and subsequent references were searched independently (Kappa values for agreement were >0.80). When eligibility was not able to be determined by the title or abstract, the full text was reviewed, and when eligibility was unclear, this was resolved by group consensus.

Data extraction. The following information was systematically extracted from each article using a standardized form created by the authors. It included: study design, population-based or convenience sample, age of participants at exposure and outcome measurement, sample size and loss to follow-up, measurement of exposure and outcome, type of intervention and comparison group, confounding adjustment and results. To be categorized as a population-based study, the publication needed to report some intent and procedure to sample from a defined population base. For studies that did not report age but did report school grade, ages were approximated on the basis of knowledge of school attendance age in the country of interest. L.G.S., J.W.L., C.R.C., A.C.P.S., T. Goodwin and T. Nuske extracted data from articles. N.M.D. independently (that is, blinded to assessments of other authors) reviewed the data extraction for 15% of all studies, including all intervention studies, and consensus was reached for the very small number of discrepancies.

Where possible, we extracted a standardized 'beta' coefficient or standardized effect size to have a unit-free way of comparing effects across exposures and outcomes (that is, the difference in s.d. units between intervention and control groups, or the effect of a 1 s.d. increase in exposure on an outcome in observational studies). When unstandardized coefficients were reported, where possible, we calculated standardized effect size to allow comparability of effects across the studies. When a standardized effect size could not be calculated (that is, standard deviations for exposure and outcome were not reported), we reported the unstandardized effect sizes.

Screening to assess risk of bias. The authors J.W.L., L.G.S. and A.C.P.S. reviewed all eligible studies and rated their evidence quality as 'better, weak or poor' on the basis of study design and confounding adjustment (Table 2). For RCTs, the risk of bias was assessed using the Cochrane Collaboration Risk of Bias Assessment Tool[83] (Supplementary Table 6). We adopted a 'potential outcomes approach' to conceptualizing confounding, in which the interpretation of a 'causal' effect of an exposure estimated from observational data relies on several assumptions[101]. One of the key assumptions is conditional exchangeability between exposed and unexposed. This corresponds to the idea that the estimate is reasonably free from 'confounding' by poorly measured or unmeasured characteristics. This is called endogeneity bias in economics. Thus, our assessment of better-quality evidence relied on a subjective judgement of the risk of bias from confounding. Publications that made no attempt to statistically control for common causes of exposure and outcome were rated as 'poor' because the likelihood of confounding (endogeneity) bias was high, and so these publications could not inform any assessment of likely causal effects of non-cognitive skills on outcomes. Conversely, observational studies using fixed-effects regression (that is, twins, siblings and within-individual change) or adjustment for strong common causes of the exposure–outcome association (including proxies for these, such as baseline measures of the outcome or a child's

cognitive ability) were rated as better evidence. Here, we only report results from studies that met the definition of 'better evidence'. However, all weak and poor evidence studies were reviewed and appear in Supplementary Tables 7 and 8.

Data synthesis. *Meta-analysis and forest plots.* We used effect sizes as reported in the original study or, where possible, used information presented to calculate effect sizes as Hedges' g. This may mean that some differences exist in how different studies calculated effect sizes in terms of how they included information on standard deviations of the outcome. We synthesized the information on effect sizes by undertaking random effects meta-analysis using inverse variance weighting. When no measure of variance was reported, we calculated confidence intervals from *P* values[102]. It was common for studies to not report variance or exact *P* values. To overcome this problem for conducting meta-analyses using inverse variance weighting, we were forced to make assumptions about *P* values to calculate confidence intervals. If the *P* value was reported as less than a specific value, we assumed the *P* value equalled that value, for example, if the *P* value was reported as $P < 0.01$, we assumed $P = 0.01$ for the purpose of calculating confidence intervals. When the *P* value was reported as greater than a specific value, we followed the Cochrane Review Handbook, which recommends removing any estimates where the *P* value is reported as greater than some value[91]. The main summary of results is shown in Fig. 2a (EQIs) and Fig. 2b (observational studies). We show the meta-analysed average effect size (and its 95% confidence interval) in each subdomain of academic achievement, psychosocial, cognitive and language, and health outcome. The 95% confidence interval informs how precisely the mean effect size has been estimated. On unlimited repetitions of sampling, and assuming that there is no effect (that is, the null is true), then 95% of all of the confidence intervals calculated would include the true population mean—in this case, the effect size. We also present the 95% prediction interval, which indicates the heterogeneity of effects across the population of studies that generated the meta-analysis effect size. The prediction interval estimates where the true effects are to be expected for 95% of similar studies that might be conducted in the future[103,104].

More detailed analyses showing individual publications in each of the subdomains (for example, literacy) are presented in Supplementary Figs. 2a–19b according to study design (EQIs versus observational, and then by cluster, individual, quasi-experimental, longitudinal and cross-sectional). To reduce bias that may have arisen from studies reporting multiple measures of the same outcome, we obtained an overall estimate across all of the reported measures. For example, if a publication reported three different measures of literacy, we meta-analysed those three estimates to get an overall effect. These are the estimates shown in Supplementary Figs. 2a–19b. These figures show the meta-analysed effect size (95% confidence interval), Tau[2] (a measure of variation in true effects among studies), the I^2 statistic, which describes the proportion of observed variability that can be attributed to among-study heterogeneity[104], and the 95% prediction intervals.

Funnel plots and Egger regression. We examined asymmetry of the published evidence by generating funnel plots of effect size against the inverse of study size separately for EQIs and observational studies (Supplementary Figs. 20a–23b) and calculated the summary Egger regression coefficient and *P* value indicating the degree of asymmetry[81]. The coefficient from the Egger regression tests whether the *y* intercept is zero. The expectation is that the *y* intercept is zero if there is an even spatial spread of studies within the funnel. The coefficient is the effect size normalized by dividing by the standard error (*x* axis) against the reciprocal of the standard error of the estimate (*y* axis). Small *P* values on the Egger regression coefficient suggest the presence of small study bias that may produce larger effects.

Length of follow-up. To include information on the length of follow-up, we graphed each publication according to the length of follow-up, effect size and study size (Supplementary Figs. 24–31). The size of the icon in Supplementary Figs. 24–31 corresponds with small ($n < 100$), medium ($n = 100$–500) and large ($n > 500$) studies. The length of the line displays the duration of follow-up. Supplementary Fig. 32 specifically compares end of intervention (or as closely as we could approximate) and follow-up effects for studies in which it could be calculated.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data used to undertake this systematic review and meta-analysis are freely available from our BetterStart website (https://health.adelaide.edu.au/betterstart/).

Received: 3 August 2016; Accepted: 28 September 2018;
Published online: 5 November 2018

References

1. Bowles, S. & Gintis, H. *Schooling in Capitalist America: Educational Reform and the Contradictions of Economic Life* (Basic Books, New York, 1976).
2. Deming, D. J. The growing importance of social skills in the labor market. *Q. J. Econ.* **132**, 1593–1640 (2017).
3. *Skills for Social Progress: The Power of Social and Emotional Skills* (OECD, Paris, 2015).
4. Institute of Education *The Impact of Non-cognitive Skills for Young People* (UK Cabinet Office, 2013).
5. Allen, G. *Early Intervention: the Next Steps. An Independent Report to Her Majesty's Government* (UK Cabinet Office, 2011).
6. Heckman, J. J. Skill formation and the economics of investing in disadvantaged children. *Science* **312**, 1900–1902 (2006).
7. Heckman, J. J. & Kautz, T. Hard evidence on soft skills. *Labour Econ.* **19**, 451–464 (2012).
8. Lindqvist, E. & Vestman, R. The labor market returns to cognitive and noncognitive ability: evidence from the Swedish enlistment. *Am. Econ. J. Appl. Econ.* **3**, 101–128 (2011).
9. Cunha, F., Heckman, J. J. & Schennach, S. M. Estimating the technology of cognitive and non-cognitive skill formation. *Econometrica* **78**, 883–931 (2010).
10. Heckman, J. J., Stixrud, J. & Urzua, S. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *J. Labor Econ.* **24**, 411–482 (2006).
11. Duncan, G. J. et al. School readiness and later achievement. *Dev. Psychol.* **43**, 1428–1446 (2007).
12. Hendry, A., Jones, E. J. H. & Charman, T. Executive function in the first three years of life: precursors, predictors and patterns. *Dev. Rev.* **42**, 1–33 (2016).
13. Diamond, A., Barnett, W. S., Thomas, J. & Munro, S. Preschool program improves cognitive control. *Science* **318**, 1387–1388 (2007).
14. Borghans, L., Duckworth, A. L., Heckman, J. J. & Ter Weel, B. The economics and psychology of personality traits. *J. Hum. Resour.* **43**, 972–1059 (2008).
15. Heckman, J. J. & Kautz, T. *Fostering and Measuring Skills: Interventions that Improve Character and Cognition* (National Bureau of Economic Research, 2013).
16. Diamond, A. & Lee, K. Interventions shown to aid executive function development in children 4 to 12 years old. *Science* **333**, 959–964 (2011).
17. Pearce, A. et al. Do early life cognitive ability and self-regulation skills explain socio-economic inequalities in academic achievement? An effect decomposition analysis in UK and Australian cohorts. *Soc. Sci. Med.* **165**, 108–118 (2016).
18. Eisenberg, N. et al. Relations among maternal socialization, effortful control, and maladjustment in early childhood. *Dev. Psychopathol.* **22**, 507–525 (2010).
19. Fergusson, D. M., Boden, J. M. & Horwood, L. Childhood self-control and adult outcomes: results from a 30-year longitudinal study. *J. Am. Acad. Child Adolesc. Psychiatry* **52**, 709–717.e1 (2013).
20. Evans, G. W., Fuller-Rowell, T. E. & Doan, S. N. Childhood cumulative risk and obesity: the mediating role of self-regulatory ability. *Pediatrics* **129**, e68–e73 (2012).
21. Blair, C. & Razza, R. P. Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Dev.* **78**, 647–663 (2007).
22. Mischel, W., Shoda, Y. & Peake, P. K. The nature of adolescent competencies predicted by preschool delay of gratification. *J. Pers. Soc. Psychol.* **54**, 687–696 (1988).
23. Moffitt, T. E. et al. A gradient of childhood self-control predicts health, wealth, and public safety. *Proc. Natl Acad. Sci. USA* **108**, 2693–2698 (2011).
24. Kern, M. L. & Friedman, H. S. Do conscientious individuals live longer? A quantitative review. *Health Psychol.* **27**, 505–512 (2008).
25. Raver, C. C. et al. CSRP's Impact on low-income preschoolers' preacademic skills: self-regulation as a mediating mechanism. *Child Dev.* **82**, 362–378 (2011).
26. Deary, I. J., Whiteman, M. C., Starr, J. M., Whalley, L. J. & Fox, H. C. The impact of childhood intelligence on later life: following up the Scottish mental surveys of 1932 and 1947. *J. Pers. Soc. Psychol.* **86**, 130–147 (2004).
27. Fergusson, D. M., John Horwood, L. & Ridder, E. M. Show me the child at seven II: childhood intelligence and later outcomes in adolescence and young adulthood. *J. Child Psychol. Psychiatry* **46**, 850–858 (2005).
28. Kuh, D., Richards, M., Hardy, R., Butterworth, S. & Wadsworth, M. E. Childhood cognitive ability and deaths up until middle age: a post-war birth cohort study. *Int. J. Epidemiol.* **33**, 408–413 (2004).
29. Whalley, L. J. & Deary, I. J. Longitudinal cohort study of childhood IQ and survival up to age 76. *BMJ* **322**, 819–822 (2001).
30. Schweinhart, L. J. et al. *Lifetime Effects: The High/Scope Perry Preschool Study through Age 40* (High/Scope Press, Ypsilanti, 2005).
31. Heckman, J. J., Pinto, R. & Savelyev, P. Understanding the mechanisms through which an early childhood program boosted adult outcomes. *Am. Econ. Rev.* **103**, 2052–2086 (2013).
32. Weikert, D. P. *Comparative Study of Three Preschool Curricula* Report No. F244 (Bureau of Elementary and Secondary Education, 1969).

33. Schweinhart, L. J., Weikart D. P. & Barnes, H. V. *Significant Benefits: The High/Scope Perry Preschool Study Through Age 27 (Monographs of the High/Scope Educational Research Foundation)* (High/Scope Press, Ypsilanti, 1993).
34. Heckman, J., Moon, S. H., Pinto, R., Savelyev, P. & Yavitz, A. Analyzing social experiments as implemented: a reexamination of the evidence from the HighScope Perry Preschool Program. *Quant. Econom.* **1**, 1–46 (2010).
35. Campbell, F. & Ramey, C. Effects of early intervention on intellectual and academic achievement: a follow-up study of children from low-income families program title: Carolina Abecedarian Project. *Child Dev.* **65**, 684–698 (1994).
36. Liberati, A. et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* **339**, b2700 (2009).
37. Webster-Stratton, C., Jamila Reid, M. & Stoolmiller, M. Preventing conduct problems and improving school readiness: evaluation of the incredible years teacher and child training programs in high-risk schools. *J. Child Psychol. Psychiatry* **49**, 471–488 (2008).
38. Conduct Problems Prevention Research Group Initial impact of the Fast Track prevention trial for conduct problems: I. The high-risk sample. *J. Consult. Clin. Psychol.* **67**, 631–647 (1999).
39. Dawson-McClure, S. et al. A population-level approach to promoting healthy child development and school success in low-income, urban neighborhoods: impact on parenting and child conduct problems. *Prev. Sci.* **16**, 279–290 (2015).
40. Nix, R. L., Bierman, K. L., Domitrovich, C. E. & Gill, S. Promoting children's social-emotional skills in preschool can enhance academic and behavioral functioning in kindergarten: findings from Head Start REDI. *Early Educ. Dev.* **24**, 1000–1019 (2013).
41. Bierman, K. L. et al. Promoting academic and social-emotional school readiness: the Head Start REDI program. *Child Dev.* **79**, 1802–1817 (2008).
42. Bierman, K. L. et al. Effects of Head Start REDI on children's outcomes 1 year later in different kindergarten contexts. *Child Dev.* **85**, 140–159 (2014).
43. Egger, M. & Smith, G. D. Misleading meta-analysis. *BMJ* **310**, 752–754 (1995).
44. Bailey, D., Duncan, G., Odgers, C. & Yu, W. Persistence and fadeout in the impacts of child and adolescent interventions. *J. Res. Educ. Eff.* **10**, 7–39 (2017).
45. Fewell, Z., Davey Smith, G. & Sterne, J. A. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *Am. J. Epidemiol.* **166**, 646–655 (2007).
46. Franco, A., Malhotra, N. & Simonovits, G. Publication bias in the social sciences: unlocking the file drawer. *Science* **345**, 1502–1505 (2014).
47. Allan, N. P., Hume, L. E., Allan, D. M., Farrington, A. L. & Lonigan, C. J. Relations between inhibitory control and the development of academic skills in preschool and kindergarten: a meta-analysis. *Dev. Psychol.* **50**, 2368–2379 (2014).
48. Brotman, L. M. et al. Cluster (school) RCT of parentcorps: impact on kindergarten academic achievement. *Pediatrics* **131**, e1521–e1529 (2013).
49. Barnett, W. S. et al. Educational effects of the Tools of the Mind curriculum: a randomized trial. *Early Child. Res. Q.* **23**, 299–313 (2008).
50. Ialongo, N. S. et al. Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression, and antisocial behavior. *Am. J. Commun. Psychol.* **27**, 599–641 (1999).
51. Raver, C. C. et al. Targeting children's behavior problems in preschool classrooms: a cluster-randomized controlled trial. *J. Consult. Clin. Psychol.* **77**, 302–316 (2009).
52. Shelleby, E. C. et al. Behavioral control in at-risk toddlers: the influence of the family check-up. *J. Clin. Child Adolesc. Psychol.* **41**, 288–301 (2012).
53. NICHD Early Child Care Research Network Do children's attention processes mediate the link between family predictors and school readiness? *Dev. Psychol.* **39**, 581–593 (2003).
54. Ramani, G. B., Brownell, C. A. & Campbell, S. B. Positive and negative peer interaction in 3- and 4-year-olds in relation to regulation and dysregulation. *J. Genet. Psychol.* **171**, 218–250 (2010).
55. Runions, K. C. & Keating, D. P. Anger and inhibitory control as moderators of children's hostile attributions and aggression. *J. Appl. Dev. Psychol.* **31**, 370–378 (2010).
56. Mintz, T. M., Hamre, B. K. & Hatfield, B. E. The role of effortful control in mediating the association between maternal sensitivity and children's social and relational competence and problems in first grade. *Early Educ. Dev.* **22**, 360–387 (2011).
57. Booth-Laforce, C. & Oxford, M. L. Trajectories of social withdrawal from grades 1 to 6: prediction from early parenting, attachment, and temperament. *Dev. Psychol.* **44**, 1298–1313 (2008).
58. Weiland, C. & Yoshikawa, H. Impacts of a pre kindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Dev.* **84**, 2112–2130 (2013).
59. Bradley, R. T., Galvin, P., Atkinson, M. & Tomasino, D. Efficacy of an emotion self-regulation program for promoting development in preschool children. *Glob. Adv. Health Med.* **1**, 36–50 (2012).
60. Ford, R. M., McDougall, S. J. & Evans, D. Parent-delivered compensatory education for children at risk of educational failure: improving the academic and self-regulatory skills of a Sure Start preschool sample. *Br. J. Psychol.* **100**, 773–797 (2009).
61. Slavin, R. E. Best evidence synthesis: an intelligent alternative to meta-analysis. *J. Clin. Epidemiol.* **48**, 9–18 (1995).
62. Egger, M., Juni, P., Bartlett, C., Holenstein, F. & Sterne, J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technol. Asses.* **7**, 1–76 (2003).
63. Diamond, A. Executive functions. *Annu. Rev. Psychol.* **64**, 135–168 (2013).
64. Chalmers, I. et al. How to increase value and reduce waste when research priorities are set. *Lancet* **383**, 156–165 (2014).
65. Ioannidis, J. P. et al. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* **383**, 166–175 (2014).
66. Open Science Collaboration Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
67. Munafò, M. R. et al. A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 0021 (2017).
68. Camerer, C. F. et al. Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637–644 (2018).
69. Duckworth, A. L. & Kern, M. L. A meta-analysis of the convergent validity of self-control measures. *J. Res. Pers.* **45**, 259–268 (2011).
70. Zhou, Q., Chen, S. H. & Main, A. Commonalities and differences in the research on children's effortful control and executive function: a call for an integrated model of self-regulation. *Child Dev. Perspect.* **6**, 112–121 (2012).
71. Kelley, T. L. *Interpretation of Educational Measurement* (World Books, New York, 1927).
72. Credé, M., Tynan, M. C. & Harms, P. D. Much ado about grit: a meta-analytic synthesis of the grit literature. *J. Pers. Soc. Psychol.* **11**, 492–511 (2017).
73. Ponitz, C. C., McClelland, M. M., Matthews, J. & Morrison, F. J. A structured observation of behavioral self-regulation and its contribution to kindergarten outcomes. *Dev. Psychol.* **45**, 605–619 (2009).
74. Cameron, C. E. et al. Fine motor skills and executive function both contribute to kindergarten achievement. *Child Dev.* **83**, 1229–1244 (2012).
75. Grindal, T. et al. The added impact of parenting education in early childhood education programs: a meta-analysis. *Child. Youth Serv. Rev.* **70**, 238–249 (2016).
76. Olds, D. et al. Effects of home visits by paraprofessionals and by nurses: age 4 follow-up results of a randomized trial. *Pediatrics* **114**, 1560–1568 (2004).
77. Iglehart, J. K. Prioritizing comparative-effectiveness research—IOM recommendations. *N. Engl. J. Med.* **361**, 325–328 (2009).
78. Fiore, L. D. & Lavori, P. W. Integrating randomized comparative effectiveness research with patient care. *N. Engl. J. Med.* **374**, 2152–2158 (2016).
79. Blair, C. & Raver, C. C. Closing the achievement gap through modification of neurocognitive and neuroendocrine function: results from a cluster randomized controlled trial of an innovative approach to the education of children in kindergarten. *PLoS ONE* **9**, e112393 (2014).
80. Knol, M. J. & VanderWeele, T. J. Recommendations for presenting analyses of effect modification and interaction. *Int. J. Epidemiol.* **41**, 514–520 (2012).
81. Egger, M., Davey Smith, G., Schneider, M. & Minder, C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* **315**, 629–634 (1997).
82. Weiss, M. J., Bloom, H. S. & Brock, T. A conceptual framework for studying the sources of variation in program effects. *J. Policy Anal. Manag.* **33**, 778–808 (2014).
83. Higgins, J. P. T. et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* **343**, d5928 (2011).
84. Kaplan, R. M. & Irvin, V. L. Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS ONE* **10**, e0132382 (2015).
85. Leyrat, C., Morgan, K., Leurent, B. & Kahan, B. Cluster randomized trials with a small number of clusters: which analyses should be used? *Int. J. Epidemiol.* **47**, 321–331 (2018).
86. Smaldino, P. E & McElreath, R. The natural selection of bad science. *R. Soc. Open Sci.* **3**, 160384 (2016).
87. Gertler, P., Galiani, S. & Romero, M. How to make replication the norm. *Nature* **554**, 417–419 (2018).
88. Munafo, M. & Davey Smith, G. Repeating experiments is not enough. *Nature* **553**, 399–401 (2018).
89. Lawlor, D. A., Tilling, K. & Davey Smith, G. Triangulation in aetiological epidemiology. *Int. J. Epidemiol.* **45**, 1866–1886 (2016).
90. Shrout, P. E. & Rodgers, J. Psychology, science and knowledge construction: broadening perspectives from the replication crisis. *Annu. Rev. Psychol.* **69**, 487–510 (2018).

91. Higgins, J. P. T. & Green, S. (eds) *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 (The Cochrane Collaboration, 2011).
92. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* 2nd edn (Lawrence Erlbaum Associates, Hillsdale, 1988).
93. Greenland, S., Maclure, M., Schlesselman, J. J., Poole, C. & Morgenstern, H. Standardized regression coefficients: a further critique and review of some alternatives. *Epidemiology* **2**, 387–392 (1991).
94. King, G. How not to lie with statistics: avoiding common mistakes in quantitative political science. *Am. J. Polit. Sci.* **30**, 666–687 (1986).
95. Cheung, A. C. K. & Slavin, R. E. How methodological features affect effect sizes in education. *Educ. Res.* **45**, 283–292 (2016).
96. Lipsey, M. W. et al. *Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms* (US Department of Education, 2012).
97. Watts, D. J. Should social science be more solution-oriented? *Nat. Hum. Behav.* **1**, 0015 (2017).
98. Blair, C. & Diamond, A. Biological processes in prevention and intervention: the promotion of self-regulation as a means of preventing school failure. *Dev. Psychol.* **20**, 899–911 (2008).
99. Blair, C. & Raver, C. C. School readiness and self-regulation: a developmental psychobiological approach. *Annu. Rev. Psychol.* **66**, 711–731 (2015).
100. Diamond, A. Activities and programs that improve children's executive functions. *Curr. Dir. Psychol. Sci.* **21**, 335–341 (2012).
101. Little, R. J. & Rubin, D. B. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu. Rev. Public Health* **21**, 121–145 (2000).
102. Altman, D. G. & Bland, J. M. How to obtain the confidence interval from a *P* value. *BMJ* **343**, d2090 (2011).
103. Higgins, J. P. T., Thompson, S. G. & Spiegelhalter, D. J. A re-evaluation of random-effects meta-analysis. *J. R. Stat. Soc. Ser. A Stat. Soc.* **172**, 137–159 (2009).
104. Borenstein, M., Higgins, J. P. T., Hedges, L. V. & Rothstein, H. R. Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Res. Synth. Methods* **8**, 5–18 (2017).
105. VandenBos, G. R. (ed.) *APA Concise Dictionary of Psychology* (APA, Washington DC, 2009).
106. Corsini, R. *The Dictionary of Psychology* (Taylor Francis, Philadelphia, 1999).
107. Eisenberg, N. *Encyclopedia on Early Childhood Development* (Centre of Excellence for Early Childhood Development and Strategic Knolwedge Cluster on Early Child Development, Montreal, 2012); www.child-encyclopedia.com
108. Nock, M., Wedig, M., Holmberg, E. & Hooley, J. The emotion reactivity scale: development, evalution and relation to self-injurious thoughts and behaviours. *Behav. Ther.* **39**, 107–116 (2008).
109. Barkley, R. Behavioural inhibition, sustained attention, and executive functions: constructing a unifying theory of ADHD. *Psychol. Bull.* **121**, 65–94 (1997).

Acknowledgements

We thank J. Grant, T. Nuske and T. Goodwin for their research assistance in collecting, and initially screening eligibility, and in the preparation of tables and figures. J.L. is funded by a National Health and Medical Research Council of Australia Partnership Project Grant (1056888) and Centre of Research Excellence (1099422). N.D. is supported by the Economics and Social Research Council (ESRC) via a Future Research Leaders Fellowship (ES/N000757/1). The Medical Research Council (MRC) and the University of Bristol fund the MRC Integrative Epidemiology Unit (MC_UU_12013). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. All authors will have access to the data and will take responsibility for the integrity and accuracy of the review.

Author contributions

L.G.S., A.C.P.S., C.R.C., G.D.S. and J.W.L. conceived the study. L.G.S., A.C.P.S., C.R.C., N.M.D. and J.W.L. screened the literature and extracted the data. L.G.S., A.C.P.S., C.R.C. and N.M.D. analysed the data. J.W.L. led the drafting of the manuscript, with all authors contributing to the interpretation of the findings and writing of the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-018-0461-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.W.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): John Lynch
NATHUMBEHAV-16080342C

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	For this systematic review, the full search strategy including specific search terms and limits applied to search each database for relevant articles has been provided in the Supplementary material. Endnote software (version X7) was used to manage publications retrieved from literature searches. Microsoft Access and Excel (version 2013) were used to extract and collate data from studies included in the review, and excel was used to generate graphs. Effect sizes, prediction intervals, tau statistics, forest plots, funnel plots and Egger regression were conducted using Stata statistical software (SE, version 15).
Data analysis	We used Stata SE program and publicly available syntax for analyses in this study. This included: (1) metan command to generate effect sizes, tau statistics, forest plots and prediction intervals, (2) metafunnel command to generate funnel plots, and (3) metabias command to calculate Egger regression. No custom or open source code was used.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The methods section of our manuscript has a Data Availability statement where we provide a weblink to the data used for this study (methods section of manuscript). Specifically, we state: "The data used to undertake this systematic review and meta-analysis are freely available from our BetterStart website (https://health.adelaide.edu.au/betterstart/)".

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This article reports a systematic review and meta-analyses. The data are quantitative.
Research sample	All studies that had samples involving typically-developing children were eligible to be included in the review. The review was limited to children aged 12 years and under at the time of intervention or exposure, with no age limit posed on the age at follow-up. Studies with convenience samples and population-based samples were included in the review.
Sampling strategy	Eligible studies were sampled via database searches (Pubmed, PsycINFO, Embase and Business Source Complete) as well as through searching reference lists of reviews on the topic, references lists of randomised controlled trials that were included in the review, and our own libraries. As we intended to review all available literature on the topic, no predetermined sample size calculation was conducted.
Data collection	Data collection proceeded electronically, via searches of electronic databases for eligible articles and storage of articles in Endnote libraries. Extraction of data from articles included in the review was via a Microsoft access database and an Excel spreadsheet. Blinding was not undertaken during implementation of the review procedures as it is impractical to blind the researchers who extracted data from published articles.
Timing	Articles were eligible to be included in the review if they were published on or before December 2016.
Data exclusions	Articles published in other languages for which there was no English translation or published after December 2016 were excluded. Some studies were unable to be included in the meta-analyses if they did not report effect sizes and standard errors or reported p values greater than some value (e.g. p>0.05).
Non-participation	Where available, details of follow up (attrition) are given in the Supplementary Material. Not all studies reported this information.
Randomization	No randomisation occurred as part of the review. Details of randomisation procedures within each randomised controlled trial that was included in the review are provided in the Supplementary Material.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging