Smithers, L. G., Sawyer, A. C. P., Chittleborough, C., Davies, N., Davey Smith, G., & Lynch, J. (2018). A systematic review and meta-analysis of effects of early life non-cognitive skills on academic, psychosocial, cognitive and health outcomes. *Nature Human Behaviour*, *2*(11), 867-880. 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†], Alyssa C. P. Sawyer [1,2†], Catherine R. Chittleborough[1,2], Neil Davies[3,4], George Davey Smith[3,4], John Lynch[1,2,3*].

[†] Equal first authors.

[1] School of Public Health, University of Adelaide, Australia

[2] Robinson Research Institute, University of Adelaide, Australia.

[3] Bristol Medical School: Population Health Sciences, University of Bristol, Barley House, Oakfield Grove, Bristol, BS8 2BN, United Kingdom.

[4] Medical Research Council Integrative Epidemiology Unit, University of Bristol, BS8 2BN, United Kingdom.

*Author for correspondence:

Email: john.lynch@adelaide.edu.au

Abstract

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 the first systematic review of this literature we screened 9553 publications, reviewed 554 eligible publications, and interpreted results from 222 better quality publications. Better quality publications comprised randomised experimental and quasi-experimental 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 0.37SD. However, within sub-domains 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 over-estimate true effects, and heterogeneity of effect estimates spanned negative, null and positive effects. The quality of evidence from EQIs under-pinning this field is lower than optimal and more than a 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 inter-disciplinary researchers interested in these skills should take a more strategic and rigorous approach to determine which interventions are most effective.

INTRODUCTION

It is over forty 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. While 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 effects of improving different types of non-cognitive skills. We recognize 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. This literature includes studies that either manipulated non-cognitive skills through randomised 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 is 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 in Methods 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 skills constructs. We sought to provide the first 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 policy makers to underpin early life interventions[3], beyond cognitive abilities (intelligence or IQ) and academic achievement (literacy, numeracy).

The policy motivation to improve early life non-cognitive skills

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 significant 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 workforce. However, data on 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 a variety of non-cognitive skills (Supplementary Figure 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 (e.g., 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 UK and Australia showed 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].

Examples of the evidence for effects of early life non-cognitive skills

Non-cognitive abilities have been associated with a number of shorter 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 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 (i.e., intelligence or IQ) associates with academic achievement and later success in adulthood[26-29]. However, the HighScope Perry Preschool Program started in 1962 suggests other mechanisms may be involved[30,31]. The intervention provided an active learning program based on Piagetian principles, for disadvantaged 3.5 year old African American children who had IQ scores on the Stanford Binet Test < 85[32]. In analysing the long term outcomes of the trial, Heckman et al.[31], reported that while initially the intervention increased IQ these increases were not maintained to age 7-8 years. 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]. While 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. 2503)]. It should be noted that the program 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] forty years ago. They argued that

143 schooling does not make children more intelligent, rather, it socializes them into, and rewards,
144 certain characteristics and behaviours that are valued in the labour market.
145
146 The aim of this review was to systematically assess all published evidence concerning effects
147 of non-cognitive skills among children up to age 12 on later outcomes. We do not review
148 intervention studies that did not specifically aim to improve non-cognitive skills. Thus, some
149 interventions such as the Perry Preschool[30] and Abecedarian[35] programs are not formally
150 reviewed here because we could find no documented evidence that these programs
151 specifically set out to improve non-cognitive abilities, and so were not eligible.
152
153 We screened eligible publications and report results on associations between non-cognitive
154 skills up to age 12. We grouped publications into four outcome domains - academic
155 achievement (including literacy, numeracy and school readiness), cognitive and language
156 development (including intelligence and language), psychosocial well-being (including
157 mental health problems such as internalising and externalising problems, hyperactivity, social
158 skills, and classroom behaviour), and health (including anthropometry and injury). In this
159 manuscript we only report results from those publications we judged to be "better" evidence
160 derived from RCTs and quasi-experimental studies grouped as experimental and quasi-
161 experimental intervention studies (EQIs), and observational studies that made reasonable
162 attempts to control for confounding (endogeneity) bias. However, all eligible publications
163 were fully reviewed and for completeness are presented in Supplementary Tables 7 and 8.
164
165 **RESULTS**
166 The systematic search identified 9553 articles from electronic and hand-searched sources.
167 After removing duplicates and assessing eligibility, 554 articles were included and presented

in a PRISMA[36] flowchart (Figure 1). There were 49 (9%) publications involving RCTs and non-randomised quasi-experimental interventions that reported 85 outcomes, 69% of which were in the academic achievement and psychosocial outcome domains. (Table 1). Below we report this group of studies as Experimental/Quasi-experimental studies (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 1 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 summarised in Figure 2 and Supplementary Figures 2a-19b, and 24-31, which display all studies where 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 (WJ) psycho-educational battery. For EQIs, Figure 2 (panel A) shows effect sizes ranged from 0.16SD (95%CI -0.02 to 0.34) for academic achievement and school readiness to 0.37 (95%CI 0.16 to 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, Figure 2 (panel B) shows effect sizes ranged from 0.16 (95%CI 0.12 to 0.20) for literacy and 0.22 (95%CI 0.14 to 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 Figures 2a-4b. Supplementary Figures 24-25 graph effect size, age and length of follow up.

EQIs: There were 26 publications reporting ten cluster (school or class) RCTs, eleven individual RCTs, one study where the unit of randomisation was unclear, and four 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 (e.g. 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. Median age at the time of intervention was 4.5 y. The median follow-up time was under 1 year. The oldest age at follow up was 20 y, 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 0.49SD (Supplementary Figures 2a-4b). The individually randomised trials were generally smaller and demonstrated effect sizes up to 0.81SD but were more heterogeneous with a 95% prediction interval for literacy consistent with negative and positive effects ranging from -0.91 to 1.79.

Observational: There were four publications of twin studies, 58 longitudinal (including four fixed effects analysis) and 14 cross-sectional publications, with three publications reporting results from multiple cohort studies. Non-cognitive abilities were measured at median age 5.0 y and median follow up of 1.5 years. The oldest age at follow up was 16 y. 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 Figures 2a-4b show effect sizes ranging from negative effects (-0.57SD), to null, to 0.77SD for numeracy and similarly for literacy up to 0.80SD. However, 95% prediction intervals were generally narrower than for EQIs (e.g. -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 (internalising and externalising behaviour), social skills, and aspects of school readiness, such as learning engagement. For EQIs, Figure 2 (panel A) shows effect sizes ranged from 0.23SD (95%CI 0.15 to 0.30) for externalising behaviour to 0.46SD (95%CI 0.31 to 0.61) for social skills. For observational studies, Figure 2 (panel B) shows effect sizes ranged from 0.13SD (95%CI 0.07 to 0.18) for social skills to 0.21SD (95%CI 0.15 to 0.28) for externalizing behaviour. The 95% prediction interval for all psychosocial outcomes were consistent with negative, null and positive effects. For example, the 95% prediction interval for externalising 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 Figures 5a-9b. Supplementary Figures 26-27 graph effect size, age and length of follow-up. Studies were not consistent in scoring of psychosocial outcomes, i.e. 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 five quasi-experimental intervention studies where the intervention was delivered in schools, sports classes, at home, or in community-based settings. 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. Median age at the time of intervention was 4.5 years with median follow up time less than one year. The oldest age at follow up was 13.5 y from a non-randomised intervention. Intervention groups ranged in size from n=16 to 314 for the individually-randomised trials and n=20 to ~3,350 for cluster RCTs (the largest RCT did not report the exact intervention number). For externalising outcomes, the 95% prediction interval for cluster RCTs was 0.10 to 0.37SD and -0.15 to 0.61SD 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, Incredible Years) and had both a teacher and parent engagement component[37-39], only reported effects where $p \leq 0.05$ for three of the eleven outcomes studied.

Observational: There were five publications of twin studies, 52 longitudinal and 19 cross-sectional publications. The five reasonably-sized twin studies that combined MZ and DZ twins (n ranged from 209-410 pairs) of children aged ~2-8 reported phenotypic correlations between non-cognitive abilities and internalising problems of 0 to -0.3, and -0.1 to -0.6 for fewer externalising 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 median age 5.0 years with 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 Figures 5a to 9b show effects from observational studies consistent with ~0.1 to 0.2SD 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 to 0.7SD) on social skills and mental health problems. For effortful control, evidence was mixed, ranging from null to 0.85SD on externalizing behaviour.

Cognitive & 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, Figure 2 (panel A) shows the effect sizes ranged from 0.27SD (95%CI 0.01 to 0.53) for expressive vocabulary to 0.56SD (95%CI 0.14 to 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, Figure 2 (panel B) shows effect sizes ranged from 0.08SD (95%CI -0.01 to 0.17) for general cognitive development to 0. 20SD (95%CI 0.11 to 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 Figures 10a-16b. Supplementary Figures 28-29 graph effect size, age and length of follow up.

EQIs: There were 23 publications reporting 18 RCTs (two interventions were reported in six publications) and five quasi-experimental intervention studies. Of the RCTs, six were cluster (school or class) RCTs, one where the unit of randomisation was unclear, and eleven individual RCTs, involving programs delivered in schools or classrooms, at home, in a laboratory setting or a combination of classes and home. Three quasi-experimental interventions involved preschool programs and two involved computerised 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. Median age at intervention was 4.3 years, with median follow up of less than one year, extending to 16 years. One RCT inconsistently reported effects of 0.15 and 0.25SD on the same language outcome, using the same sample at age five[40,41] and an effect of 0.10SD at age 6 in a different publication[42].

Observational: There were six publications of twin studies, 14 longitudinal (including one fixed effect) and nine cross-sectional publications. The six twin studies that combined MZ and DZ twins (n ranged from 40-901 pairs) of children reported phenotypic correlations between non-cognitive abilities and intelligence of -0.36 to 0.23SD. The longitudinal and cross-sectional publications ranged in effect size from -0.38 (a cross-sectional convenience sample n=77 examining attention) to 0.56SD (a cross-sectional convenience-sample n=80 examining executive function). Exposure was measured at median age 4.5 years. The median duration of follow up for the longitudinal studies was less than one 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 Behavior 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 0.52SD, but the evidence is predominantly from convenience samples. There were too few studies to make any judgments 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 two small RCTs, one 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 Figures 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 0.79SD, but this effect is difficult to interpret because of an inadequate description of the control group and the outcome. Median age at intervention was 4.4 years. Median follow up time was less than one 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 summarise these because studies either did not report effect sizes, or reported unstandardised 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 Figures 20a-23b depict effect sizes for experimental and observational studies separately, according to the standard error of the effect size. These include all publications where effect sizes were reported or able to be calculated, and reported exact p values or $P<0.05$[43]. Thus, all studies that reported P 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 <0.01. There was little evidence for differential small study bias comparing EQIs and observational studies.

Fade Out

Supplementary Figure 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 effects of non-cognitive skills up to age 12 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 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 where children indirectly develop a variety of non-cognitive skills and behavioural styles.

We were limited to reporting what might be termed ‘proxy’ or ‘intermediate’ outcomes. While outcomes like 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 one 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 0.5SD depending on outcome as shown in Figure 2 Panel A.. We urge some caution in interpreting our results. Analysis of funnel plots clearly demonstrate asymmetry of effect size and the potential of small study bias[43]. Additionally, forest plots and 95% prediction intervals show large heterogeneity of reported effect sizes generally including the null. This suggests the overall meta-analysed effects from EQIs reported here may be over-estimates that include a null effect.

Presenting the analysis in Figure 2 by separating EQIs (Panel A) and observational publications (Panel B) shows larger effects from EQIs than found in higher quality observational studies which ranged from approximately 0.06 to 0.22SD. This is the opposite of what is often seen where observational studies over-estimate effects found in large, well designed RCTs. This over-estimation 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) effect sizes from EQIs and observational studies would only be comparable if the EQI induced a SD change in the particular non-cognitive skill. In reality, effects of interventions on the target non-cognitive skill might be closer to 0.2 to 0.5 SD. So at 0.25 and with no bias, effects found in observational data would be expected to be four times larger than experimental impacts.

Franco *et al*.[46] found that among rigorously reviewed social science publications in the Time-Sharing in the Social Sciences National Sciences Foundation database that "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 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 focussed on improving children's non-cognitive skills at around 4 years of age with median follow-up under one year. These studies were generally consistent with 0.2 to 0.4SD 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 0.3SD[25,48-50]. These

larger higher-quality RCTs spanned child-focussed interventions on specific domains of non-cognitive skills (e.g. Tools of the Mind), to more teacher-focussed curricula (e.g. Chicago School Readiness), to more multi-dimensional content interventions that included parent, child and teacher (e.g. PATHS). Observational studies on academic achievement generally showed effects around 0.2SD but all 95% prediction intervals included the null. This is consistent with one higher quality observational publication[11] which examined six different cohorts with longer follow-up of 5.5 years and reported effects from 0 to 0.2SD. Overall, there was insufficient evidence upon 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 one year[42].

Psychosocial outcomes: For psychosocial outcomes, the evidence from RCTs was dominated by studies of externalising problems, with fewer RCTs on social skills and internalising problems. Average age at the time of intervention was around 4 years with median follow up time under one year. Effects on externalising problems for EQIs was 0.23 (95% CI 0.15-0.30) with a 95% prediction interval of 0.07 to 0.39. Higher quality RCTs examining externalising outcomes, reported positive[5152] 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) was 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 conceptualisations of where psychosocial problems arise and for how, where and whom to intervene (e.g. 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 to 0.2SD. The largest effect sizes were from a well-designed regression discontinuity study (0.44SD)[58], a non-randomised intervention (0.55-0.73SD)[59] and a small, low quality randomised trial[60]. However, all these studies were small (ranging from n=12 to 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 ~0.1SD, with seven of nine longitudinal studies showing few differences and

493 cross-sectional studies reporting mixed effects (-0.38 to 0.56SD). The longitudinal studies
494 were dominated by non-cognitive skills measured using the HTKS and the WJ Picture
495 Vocabulary as the outcome, and despite the popularity of these measurement tools, the results
496 indicate no effects on vocabulary outcomes. Thus, non-cognitive abilities appear to have
497 effects on cognitive and language outcomes of ≤0.2SD.
498
499 *Physical health outcomes:* It is difficult to draw conclusions for physical health outcomes.
500 There were only 3 EQIs reporting diverse outcomes. Outcomes reported across the 28 better
501 quality observational studies were diverse ranging from anthropometry, to injury to
502 physiological characteristics and were consistent with effects ranging from 0.06 to 0.14SD.
503
504 **Limitations of this review**
505 The compilation of 554 publications was systematic but our assessment of the quality of the
506 evidence is based on our judgement of the potential for bias. Here we follow the approach of
507 others who have argued for limiting systematic reviews and meta-analyses to higher quality
508 evidence[61,62]. We *a priori* created criteria for bias based on well-established procedures
509 including quality appraisal tools, evidence hierarchies, directed acyclic graphs and content
510 knowledge about potential sources of confounding and selection bias. While this involves an
511 element of subjective judgement, we are confident that any other reasonable assessment of the
512 quality of evidence would not change the overall conclusions presented here. In the interests
513 of transparency we have disclosed all of the subjective choices we have made in the
514 Supplementary materials and text.
515
516 It is possible that some relevant articles were not included in this review, even though we
517 undertook an extensive search that included multiple databases, numerous search terms,

518 contacting authors of potentially-relevant papers, and hand searching reference lists of
519 published papers. Studies of systematic review methods have shown that the most difficult to
520 find articles are in the 'grey literature', sometimes smaller, of poorer quality and the results
521 unlikely to unduly influence the findings in an already large systematic review[62].
522
523 **The value of a systematic review**
524 While there have been reviews of some aspects of non-cognitive skills,[3,4,14,15,47,63] none have
525 been systematic in covering the entire literature, or included screening for evidence quality. It
526 has long been recognised in health and medical research that non-systematic reviews of
527 research enable the selective use of evidence to support a particular argument[62]. For evidence
528 consumers, who are often not evidence-quality specialists, competing claims about effects of
529 non-cognitive abilities based on particular studies are hard to reconcile without the safety net
530 of a systematic review. We have paid particular attention in this review to issues of quality of
531 the primary evidence. There is little point in summarising evidence that includes obviously
532 flawed studies that can only distort the overall results and reduce the value of the systematic
533 nature of the review[61,62,64].
534
535 This review covers the entire published inter-disciplinary research field describing intentional
536 efforts and observational analogues of interventions to improve development of non-cognitive
537 skills, albeit with most evidence coming from rich countries, especially the USA. The scope
538 of the review should minimise 'cherry picking' of results to bolster a particular concept,
539 theory or intervention. This is necessary to advance knowledge given the multidisciplinary
540 nature of this field and is central to informing interventions to boost life chances for
541 disadvantaged children. In health sciences, major advances have been made by coming to
542 agreement and attempting, where possible, to harmonise methods for measuring exposures,

outcomes, synthesising 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 operationalisation[69,70]. In 1927, Kelley labelled this the "jangle" fallacy[71 (p. 64)] where constructs are given different names but in fact are 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 conceptualisations of non-cognitive abilities. For example the Continuous Performance Task (aka "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 "Head Toes Knees Shoulders Task" has been used to measure both behavioural self-regulation[73] and executive functioning[74]. The interventions we reviewed attempted to influence many different facets of non-cognitive skills. Policy makers and researchers ideally need to know what the 'active ingredients' are, in order 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 programs 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 (e.g. 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 compared to para-professionals[76]. Interventions that trained children in more specific skills such as executive function, generally showed small effects (e.g., Tools of the Mind)[49]. Other studies imply that non-specific interventions seem to generate better generalised outcomes[31], which may suggest that more holistic programs including multi-dimensional 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 small effects of ~0.1SD 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 sub-groups 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 age 4-5, with median follow-up of about one year, and with very few studies with follow-up beyond age 10, there is very little evidence addressing effects on medium to longer-term outcomes. This is no doubt due to funding constraints. However, it is frequently argued that non-cognitive skills developed in childhood have major impacts on long-term adult outcomes[6]. Thus, interventions which have short term effects, but few detectable long-term effects are unlikely to be cost-effective. Therefore, longer-term follow-up of RCTs is especially important and are 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 due to greater fidelity to the intervention as designed. On the other hand, publication bias favouring more positive results would mean 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 sub-domains such as literacy and numeracy. It is possible to argue that this was inevitable in a field where 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 instance, of the 11 literacy EQIs able to be included in the meta-analysis only three were population based. For externalising behaviours, of 13 EQIs, only one 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 2014-2016. There were 27 previous RCTs of non-cognitive skills on academic achievement and psychosocial outcome domains available to be cited. The

668 highest number of citations of previous RCTs referenced in any of the RCTs published from
669 2014-16 was four[79]. Several RCTs published between 2014 and 2016 referenced no previous
670 RCTs. It could perhaps be argued that these RCTs intervened on different non-cognitive skills
671 so should not necessarily cite studies of other non-cognitive skills. Nevertheless, attention
672 regulation and self-control were common ingredients of almost all of these interventions
673 (Supplementary Table 2-5), so the impression is that new RCTs were not being explicitly
674 justified on the basis of what was already known from existing RCTs.
675
676 *Quality of RCTs*
677 The quality of RCTs was not ideal and reporting of some details was poor or even absent. No
678 RCTs had a formal pre-registered protocol and two thirds did not explicitly identify primary
679 outcomes (See Supplementary Table 6 on Risk of Bias Tool[83]). This can allow cherry-
680 picking of significant results within studies rather than focus on a single or small number of
681 pre-stated main outcome(s) that the intervention is theoretically, or empirically (based on
682 previous evidence) meant to most influence[84]. Over one-quarter of RCTs may have had other
683 potential biases, for example, differential participation in the control and intervention groups,
684 and unclear processes for selection of control participants. Ninety-two percent of RCTs did
685 not adequately report randomisation procedures, 81% did not report concealment of allocation
686 processes and participant flow, and most failed to address missing data. It was common for
687 cluster RCTs to have too few clusters to achieve balance between intervention and control
688 groups and in some it was unclear whether clustering was adequately dealt with in the
689 analysis[85]. Poor reporting made it difficult to fully assess study quality and we strongly
690 encourage researchers, journal editors and reviewers to use tools such as the CONSORT
691 statement (http://www.consort-statement.org/) for reporting, and for RCTs to be pre-
692 registered. These are now mandatory requirements for publishing in most leading health and

693 medical science journals. However, it is possible that research practice regarding pre-
694 registration is already changing and those pre-registered studies are yet to be published.
695
696 *Quality of observational studies*
697 More than 90% of all research in this field comes from observational studies. Of the 504
698 observational studies reviewed here, 66% were judged as 'weak' or 'poor' quality. Of all
699 observational studies, 42% made little or no attempt to adjust for even basic confounding i.e.,
700 common causes of non-cognitive ability and the outcome. Problems of endogeneity and
701 confounding are well known and may result in substantial bias of the association of non-
702 cognitive skills and later outcomes.
703
704 One regrettable consequence of the relatively low quality of much of the research effort in this
705 field, is that it is not able to shed much light on the question of whether improving non-
706 cognitive skills positively influences outcomes. To advance understanding of non-cognitive
707 skills in children and their effects on outcomes later in life, there is little point in amassing
708 more small-scale[86], biased observational or experimental studies that have higher likelihood
709 of failing to be replicated[65,66,87] and are unable to contribute to evidence triangulation which is
710 central for stronger causal inference[88,89]. The recommendations we make here to improve
711 evidence quality in this field are not controversial. A 2018 *Annual Review of Psychology*
712 paper called for more sophisticated power analyses, better statistical practices, study design
713 specific to addressing effect modification, and better disclosure of non-significant as well as
714 significant findings[90].
715
716 *Implications of Sub-optimal Reporting Practices of Effect Sizes and P-values*

In order 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. Where standard errors were not available, we calculated standard error from an exact p-value or where 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 studies were excluded where an effect size and/or a standard error could not be calculated[91]. Excluding studies reporting P>p provides a more conservative estimate of the precision of studies. These exclusions were on top of excluding studies where an effect size was either not reported or could not be calculated. We illustrate the effect of this 2-layer exclusion for literacy outcomes. The literature reported 49 literacy related outcomes in 17 EQIs. Excluding outcomes where an effect size could not be calculated reduced the number of available literacy outcomes to 42 outcomes from 14 EQIs. Further excluding results where the p-value was reported as P>p meant the meta-analysis and funnel plots could only include 33 literacy outcomes from 11 EQIs. Thus, this 2-layer exclusion of reported results (due to sub-optimal reporting practices) meant we could only include 67% of the literacy outcomes actually presented in the literature. This also meant 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 (~0.2SD)", "medium (~0.5SD)", or "large (~0.8SD)" 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 if an effect is "big" is not

742 straightforward in any field. Effect sizes are nothing more than mean differences between
743 intervention (exposed) and control (unexposed) groups on some scale of outcome
744 measurement divided by the standard deviation of the outcome. Use of such standardized
745 effect measures has been criticized in several disciplines. In epidemiology, Greenland et al.[93],
746 have argued that the process of standardizing effects, rather than making them more
747 comparable across studies, simply serves to confound that comparison by the observed
748 standard deviation, which is often an artefact of the study sample, particularly for
749 homogenous convenience samples. In political science, King argued that if apples and
750 oranges cannot be meaningfully compared on the original outcome measurement scale then
751 this lack of comparability is not improved by comparing standardized fruit[94]. Size of effects
752 must be judged within the context of the field, the methods used in the study[95] and,
753 importantly, linked to some normative understanding of what weak or strong effects look like
754 in a particular field. For example, if the best interventions available to improve a particular
755 outcome, found reliable effects of 0.2SD when trialled in large population based samples,
756 then a novel intervention finding the same effect might be considered large. Another way of
757 norming effect size may be to consider the size of intervention effects against secular change
758 in an outcome over time. Lipsey, et. al. present a sophisticated understanding of interpreting
759 effect sizes[96]. For example, they show that the secular growth in reading from kindergarten to
760 grade one in the US is estimated to be about 1.5 SD. By grades 4-5 this growth has declined
761 to about 0.4 SD per year. How should an effect of a non-cognitive skills intervention in
762 kindergarten on reading in grade one of 0.2 SD be judged? Such an intervention has generated
763 about 13% greater improvement than the natural growth in reading during that time. Deciding
764 whether an intervention is worth implementing will depend not only on its benefits, but also
765 its costs, discount rate, scalability and a range of other potential considerations. Interventions
766 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. Therefore, the traditional labelling of an intervention as having "small" effects (~0.2SD) 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 efforts (from observational evidence) 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, 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 over-estimate 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 which are funded for long-term follow-up of some of the more promising interventions reviewed here. These results suggest profitable pathways forward to help 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

792 a strategically integrated, rigorous scientific focus to help answer the policy-relevant
793 questions.[97].
794

METHODS

The systematic review protocol was preregistered 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 to age 8. Reviewers suggested extending this to age 12 hence the protocol was updated in September 2017.

Inclusion criteria

Publications were eligible if they involved non-cognitive abilities of children aged up to 12 years, 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 2 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 reporting original research were included. Publications involving non-cognitive characteristics in clinical subgroups (e.g., those already diagnosed with problems such as attention-deficit/hyperactivity disorder) were excluded because we were interested in 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. Hand searching 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 AS, LS, CC and TN). 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). Where 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 standardised 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 categorised as a population-based study the publication needed to report some intent and procedure to sample from a defined population base. Where studies 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. LS, JL, CC, AS, TG and TN extracted data from articles. ND independently (i.e., 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 standardised 'beta' coefficient or standardised effect size to have a unit free way of comparing effects across exposures and outcomes (i.e., the difference in SD units between intervention and control groups, or the effect of a 1SD increase in exposure on an outcome in observational studies). When unstandardised coefficients were reported, where possible we calculated standardised effect size to allow comparability of effects across the studies. When a standardised effect size could not be calculated (i.e., SDs for exposure and outcome were not reported) we reported the unstandardised effect sizes.

Screening to assess risk of bias

The authors JL, LS and AS reviewed all eligible studies and rated their evidence quality as 'better, weak, or poor' on the basis of study design and confounding adjustment (Table 1). 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 where 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 judgment 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. On the other hand, observational studies using fixed-effects regression (i.e., 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 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 some differences exist in how different studies calculated effect sizes in terms of how they included information on standard deviations of the outcome. We synthesised 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 p was reported as less than a specific value we assumed p equalled that value, e.g. if p was reported as $p<0.01$ we assumed $p=0.01$ for the purpose of calculating confidence intervals. Where p was reported as greater than a specific value, we followed the Cochrane Review Handbook which recommends removing any estimates where p is reported as greater than some value[91]. The main summary of results is shown in Figures 2a (EQIs) and 2b (observational studies). We show the meta-analysed average effect size (and its 95% confidence interval) in each sub-domain 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 there is no effect (i.e., the null is true), then 95% of all 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 (e.g. literacy) are presented in Supplementary Figures 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 the Supplementary Figures 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 inverse of study size separately for EQIs and observational studies (Supplementary Figures 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 length of follow up, we graphed each publication according to length of follow up, effect size and study size (Supplementary Figures 24-31). The size of the icon in Supplementary Figures 24-S31 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 Figure 32 specifically compares end of intervention (or as closely as we could approximate) and follow up effects for studies where it could be calculated.

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/).

References

1 Bowles, S. & Gintis, H. *Schooling in capitalist America: Educational reform and the contradictions of economic life*. (Basic Books, 1976).

2 Deming, D. J. The growing importance of social skills in the labor market. *Q J Econ* **132**, 1593-1640, doi:https://doi.org/10.1093/qje/qjx022 (2017).

3 OECD. *Skills for social progress: The power of social and emotional skills.* (OECD Publishing, 2015).

4 Institute of Education. *The impact of non-cognitive skills for young people.* (U.K. Cabinet Office, London, 2013).

5 Allen, G. *Early intervention: The next steps. An independent report to Her Majesty's government*. (UK Government, London, 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, doi:http://dx.doi.org/10.1016/j.labeco.2012.05.014 (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, doi:doi: 10.1257/app.3.1.101 (2011).

9 Cunha, F., Heckman, J. J. & Schennach, S. M. Estimating the technology of cognitive and non-cognitive skill formation. *Econometrica* **78**, 883-931, doi:10.3982/ECTA6551 (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, doi:10.1037/0012-1649.43.6.1428 (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, doi:http://dx.doi.org/10.1016/j.dr.2016.06.005 (2016).

13 Diamond, A., Barnett, W. S., Thomas, J. & Munro, S. Preschool program improves cognitive control. *Science* **318**, 1387-1388, doi:10.1126/science.1151148 (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, Cambridge, MA, 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, doi:http://dx.doi.org/10.1016/j.socscimed.2016.07.016 (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 Adol Psychiatry* **52**, 709-717 (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 Social 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 (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, doi:10.1111/j.1467-8624.2010.01561.x (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 Social Psychol* **86**, 130 (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 Psyc* **46**, 850-858, doi:10.1111/j.1469-7610.2005.01472.x (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, doi:10.1093/ije/dyh043 (2004).

29 Whalley, L. J. & Deary, I. J. Longitudinal cohort study of childhood IQ and survival up to age 76. *BMJ* **322**, 819-822, doi:10.2307/25466668 (2001).

30 Schweinhart, L. J. ***et al.*** Lifetime effects: the High/Scope Perry Preschool study through age 40. (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, (Washington, D.C., 1969).

33 Schweinhart, L. J. *Significant benefits: The High/Scope Perry Preschool study through age 27.* Monographs of the High/Scope Educational Research Foundation, No. Ten. (ERIC, 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 Econ* **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 (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** (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 Psyc* **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 Psych* **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, doi:10.1007/s11121-014-0473-3 (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, doi:10.1093/aje/kwm165 (2007).

46 Franco, A., Malhotra, N. & Simonovits, G. Publication bias in the social sciences: Unlocking the file drawer. *Science* **345**, 1502-1505, doi:10.1126/science.1255484 (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, doi:10.1037/a0037493 (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, doi:http://dx.doi.org/10.1016/j.ecresq.2008.03.001 (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 Psych* **77**, 302 (2009).

52 Shelleby, E. C. *et al.* Behavioral control in at-risk toddlers: The influence of the family check-up. *J Clin Child Adolesc* **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, doi:10.1080/00221320903300353 (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, doi:10.1037/a0012954 (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. *Global Advances In Health and Medicine* **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, doi:10.1348/000712609x406762 (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**, 76, doi:10.3310/hta7010 (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, doi:http://dx.doi.org/10.1016/S0140-6736(13)62229-1.
65 Ioannidis, J. P. *et al.* Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* **383**, 166-175, doi:10.1016/s0140-6736(13)62227-8 (2014).
66 Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, doi:10.1126/science.aac4716 (2015).
67 Munafò, M. R. *et al.* A manifesto for reproducible science. *Nat Hum Behav* **1**, 0021, doi:10.1038/s41562-016-0021 (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, doi:10.1038/s41562-018-0399-z (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, doi:http://dx.doi.org/10.1016/j.jrp.2011.02.004 (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, 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* doi:10.1037/pspp0000102 (2016).
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, doi:10.1111/j.1467-8624.2012.01768.x (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, doi:http://dx.doi.org/10.1016/j.childyouth.2016.09.018 (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. *New Engl J Med* **361**, 325-328 (2009).

78 Fiore, L. D. & Lavori, P. W. Integrating Randomized Comparative Effectiveness Research with Patient Care. *New Engl J Med* **374**, 2152-2158, doi:doi:10.1056/NEJMra1510057 (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, doi:10.1371/journal.pone.0112393 (2014).
80 Knol, M. J. & VanderWeele, T. J. Recommendations for presenting analyses of effect modification and interaction. *Int J Epidemiol* **41**, 514-520, doi:10.1093/ije/dyr218 (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 variaion in program effects. *J Policy Anal Manag* **33**, 778-808, doi:10.1002/pam.21760 (2014).
83 Higgins, J. P. T. *et al.* The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* **343**, doi:10.1136/bmj.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, doi:10.1371/journal.pone.0132382 (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. *Roy Soc Open Sci* **3**, doi:10.1098/rsos.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, doi:10.1093/ije/dyw314 (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. & (editors). (The Cochrane Collaboration, 2011).
92 Cohen, J. *Statistical power analysis for the behavioral sciences*. Second edn, (Lawrence Erlbaum Associates, 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.
95 Cheung, A. C. K. & Slavin, R. E. How methodological features affect effect sizes in education. *Educ Researcher* **45**, 283-292, doi:10.3102/0013189X16656615 (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, Washington DC, 2012).
97 Watts, D. Should social science be more solution-oriented? *Nat Hum Behav* **1**, doi:10.1038/s41562-41016-40015 (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, doi:10.1146/annurev.publhealth.21.1.121 (2000).
102 Altman, D. G. & Bland, J. M. How to obtain the confidence interval from a P value. *BMJ Br Med J* **343** (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 American Psychological Association. *APA Concise Dictionary of Psychology* (ed VandenBos, G.R.) (APA, Washington, DC, 2009).
106 Corsini, R. The Dictionary of Psychology. (Ann Arbor, MI, 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, Canada, 2012). Available at: www.child-encyclopedia.com (Accessed 30 January 2017).
108 Nock, M., Wedig, M., Holmberg, E. & Hooley, J. The emotion reactivity scale: development, evalution and relation to self-injurious thoughts and behaviours. *Behavior 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 would like to thank Janet Grant, Tamara Nuske and Tom Goodwin for their research assistance in collecting, and initially screening eligibility, and in the preparation of Tables and Figures. JL is funded by a National Health and Medical Research Council of Australia Partnership Project Grant (1056888) and Centre of Research Excellence (1099422). NMD 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

LGS, AS, CC, GDS and JL conceived the study. LGS, AS, CC, ND and JL screened the literature and extracted data. LGS, AS, CC and ND analysed the data. JL 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.

Table 1: Distribution of publications (n=554) by outcome domain, study type and quality*

		Outcome Domains			
	Number of publications (%)	*Academic achievement*	*Psycho-social*	*Cognitive and language*	*Physical health*
1. '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
2. 'Weak' evidence	119/554 (21%)				
Observational longitudinal	73/119 (61%)	16	49	5	13
Observational cross-sectional	46/119 (39%)	20	28	1	3
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

* Individual publications generated multiple outcomes. For example, there were 222 publications considered as 'Better' evidence that examined 293 outcomes.

1239 **Table 2:** 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 (a) in response to a wide array of stimuli (emotion sensitivity) (b) strongly or intensely (emotion intensity), and (c) for a prolonged period of time before returning to a baseline level of arousal (emotion persistence)[108].
Emotion 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 in order 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 organise and order behaviour, such as judgement, abstraction and concept formation, logic and reasoning, problem solving, planning and sequencing of actions[105].
Impulsivity	Behaviour characterised by little or no forethought, reflection or consideration of the consequences[105].
Inhibitory control	The ability to suppress a pre-potent response, interrupt and 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 behavior through the use of self-monitoring (keeping a record of behavior), 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 such characteristics 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 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.

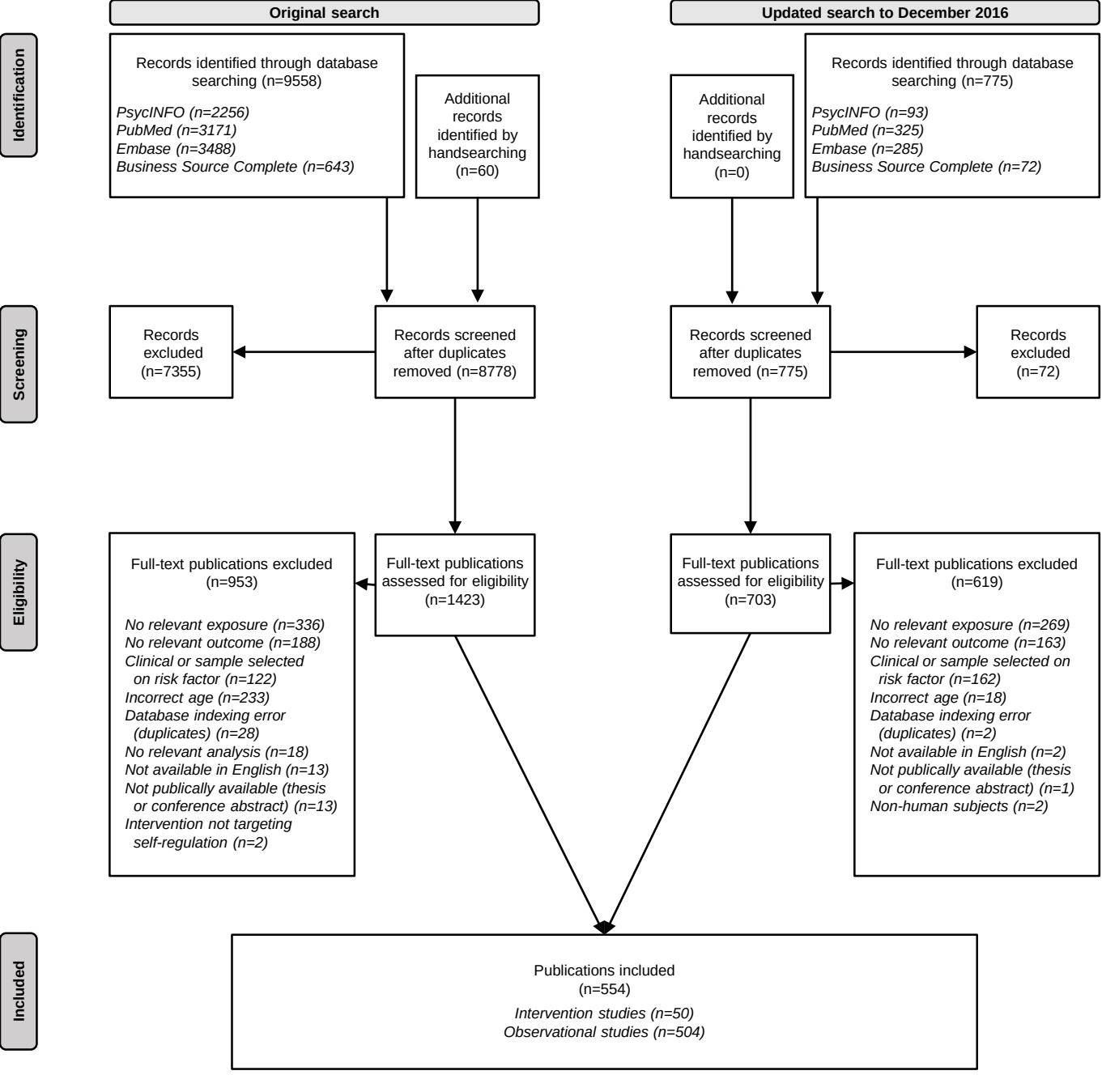
FIGURE LEGENDS

Figure 1

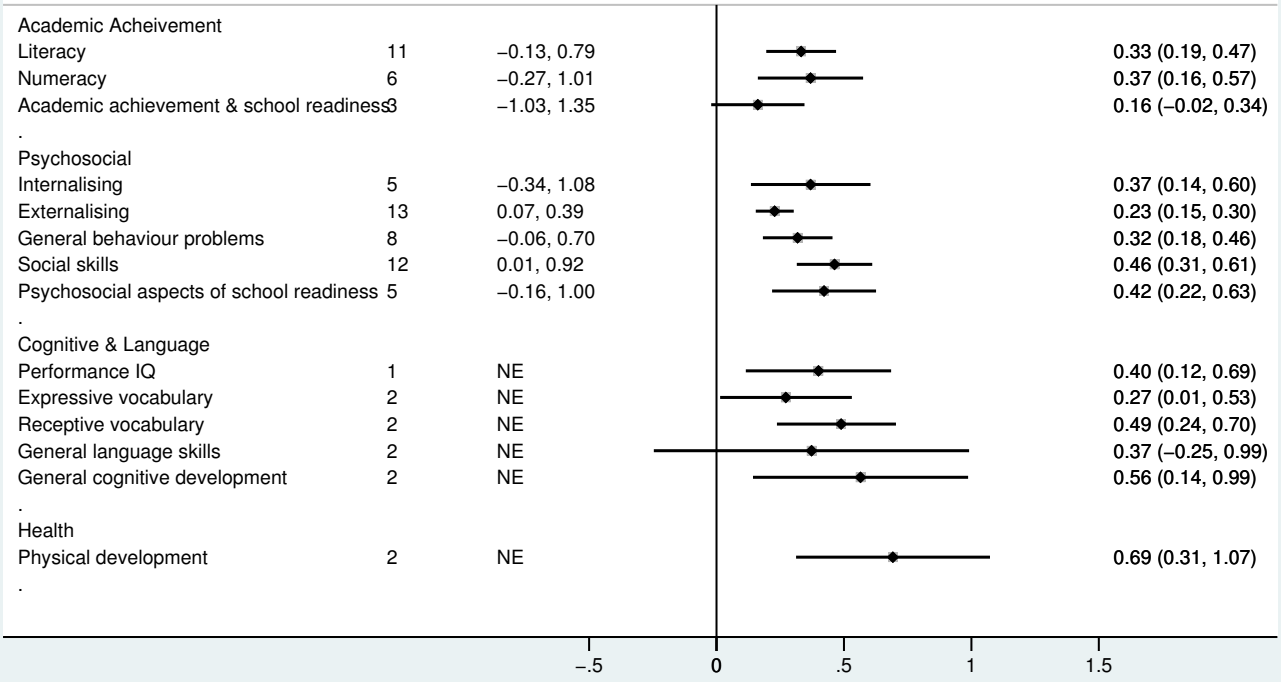
Title: Flow of publications through different stages of the systematic review

Figure 2

Title: Effect sizes from studies presenting "better quality" evidence according to outcome. a, Experimental and quasi-experimental studies. b, Observational studies. NE, not estimable; Effect sizes were calculated from random effects meta-analysis with inverse variance weighting.


Original search
Updated search to December 2016
Identification
Screening
Eligibility
Included
Records identified through database searching (n=9558)
PsycINFO (n=2256)
PubMed (n=3171)
Embase (n=3488)
Business Source Complete (n=643)
Additional records identified by handsearching (n=60)
Additional records identified by handsearching (n=0)
Records identified through database searching (n=775)
PsycINFO (n=93)
PubMed (n=325)
Embase (n=285)
Business Source Complete (n=72)
Records excluded (n=7355)
Records screened after duplicates removed (n=8778)
Records screened after duplicates removed (n=775)
Records excluded (n=72)
Full-text publications excluded (n=953)
No relevant exposure (n=336)
No relevant outcome (n=188)
Clinical or sample selected on risk factor (n=122)
Incorrect age (n=233)
Database indexing error (duplicates) (n=28)
No relevant analysis (n=18)
Not available in English (n=13)
Not publically available (thesis or conference abstract) (n=13)
Intervention not targeting self-regulation (n=2)
Full-text publications assessed for eligibility (n=1423)
Full-text publications assessed for eligibility (n=703)
Full-text publications excluded (n=619)
No relevant exposure (n=269)
No relevant outcome (n=163)
Clinical or sample selected on risk factor (n=162)
Incorrect age (n=18)
Database indexing error (duplicates) (n=2)
Not available in English (n=2)
Not publically available (thesis or conference abstract) (n=1)
Non-human subjects (n=2)
Publications included (n=554)
Intervention studies (n=50)
Observational studies (n=504)

Type of outcome	Number of studies	95% prediction interval	ES (95% CI)
Academic Acheivement			
Literacy	11	−0.13, 0.79	0.33 (0.19, 0.47)
Numeracy	6	−0.27, 1.01	0.37 (0.16, 0.57)
Academic achievement & school readiness	3	−1.03, 1.35	0.16 (−0.02, 0.34)
Psychosocial			
Internalising	5	−0.34, 1.08	0.37 (0.14, 0.60)
Externalising	13	0.07, 0.39	0.23 (0.15, 0.30)
General behaviour problems	8	−0.06, 0.70	0.32 (0.18, 0.46)
Social skills	12	0.01, 0.92	0.46 (0.31, 0.61)
Psychosocial aspects of school readiness	5	−0.16, 1.00	0.42 (0.22, 0.63)
Cognitive & Language			
Performance IQ	1	NE	0.40 (0.12, 0.69)
Expressive vocabulary	2	NE	0.27 (0.01, 0.53)
Receptive vocabulary	2	NE	0.49 (0.24, 0.70)
General language skills	2	NE	0.37 (−0.25, 0.99)
General cognitive development	2	NE	0.56 (0.14, 0.99)
Health			
Physical development	2	NE	0.69 (0.31, 1.07)



Type of outcome	Number of studies	95% prediction interval	ES (95% CI)
Academic Acheivement			
Literacy	30	−0.01, 0.33	0.16 (0.12, 0.20)
Numeracy	31	−0.04, 0.37	0.17 (0.12, 0.21)
Academic achievement & school readiness	15	−0.07, 0.52	0.22 (0.14, 0.31)
Psychosocial			
Internalising	17	−0.07, 0.37	0.15 (0.09, 0.21)
Externalising	28	−0.08, 0.51	0.21 (0.15, 0.28)
General behaviour problems	10	0.00, 0.39	0.19 (0.12, 0.27)
Social skills	14	−0.06, 0.31	0.13 (0.07, 0.18)
Psychosocial aspects of school readiness	4	−0.09, 0.36	0.14 (0.06, 0.21)
Cognitive & Language			
Total IQ	2	NE	0.20 (0.11, 0.30)
Verbal IQ	1	NE	0.09 (−0.02, 0.20)
Expressive vocabulary	1	NE	0.09 (0.04, 0.14)
Receptive vocabulary	9	−0.17, 0.50	0.16 (0.05, 0.27)
General language skills	6	−0.12, 0.33	0.11 (0.04, 0.18)
General cognitive development	1	NE	0.08 (−0.01, 0.17)
Health			
Physical development	1	NE	0.14 (0.05, 0.24)
BMI	5	−0.25, 0.37	0.06 (−0.03, 0.16)