PROTOCOL

PROTOCOL FOR "ACADEMIC INTERVENTIONS FOR CHILDREN AND STUDENTS WITH LOW SOCIOECONOMIC STATUS: A SYSTEMATIC REVIEW"

MARTIN BØG
JENS DIETRICHSON
TRINE FILGES
ANNE-MARIE KLINT JØRGENSEN

KØBENHAVN 2014

SFI DET NATIONALE FORSKNINGSCENTER FOR VELFÆRD

PROTOCOL FOR "ACADEMIC INTERVENTIONS FOR CHILDREN AND STUDENTS WITH LOW SOCIOECONOMIC STATUS: A SYSTEMATIC REVIEW
Afdelingsleder: Mette Deding
Afdelingen for kontrollerede forsøg

SFI – Det Nationale Forskningscenter for Velfærd
Herluf Trolles Gade 11
1052 København K
Tlf. 33 48 08 00
sfi@sfi.dk
www.sfi.dk

SFI-notater skal danne grundlag for en faglig diskussion. SFI-notater er foreløbige resultater, og læseren bør derfor være opmærksom på, at de endelige resultater og fortolkninger fra projektet vil kunne afvige fra notatet.

CONTENT

INTRODUCTION

This protocol describes the outline for a systematic review of interventions intended to improve the educational achievement of children and students from families that have low socioeconomic status (SES) in terms of for example parental income, parental education, and/or parental occupation. The review will focus on interventions performed by preschools, schools, and local stakeholders, where studies have used a treatment-control/comparison design and have measured effects on test scores, grade point averages, and/or take up of high school/upper secondary school. We will use a broad search strategy including interventions from preschool through grade 10, i.e. we will include interventions with target groups from about 1 to 16-17 year-olds. Depending on the comparability of interventions and the number of included studies, this may result in separate meta-analyses and/or separate reviews. For ease of exposition, we refer consequently to "the review" below.

We provide a short background and rationale for the review in sections 1.1 – 1.3. Section 2 describes the methods we will use, including a description of the types of studies we will include (2.1), search strategy (2.2), data extraction procedures (2.3), risk of bias assessment (2.4), and synthesis procedures (2.5). We also list the questions guiding the screening process in Appendix A, and describe the risk of bias tool in more detail in Appendix B.

1.1. BACKGROUND

SES is widely believed to be a major influence on educational achievement (e.g. Sirin, 2005; Currie, 2009; Kim & Quinn, 2013). The results

from the Programme for International Student Achievement (PISA) indicate for instance that most students who perform poorly in PISA are from socioeconomically disadvantaged backgrounds. The average test score difference between 15 year-old students in the top and bottom 15 percent of the PISA index of economic, social and cultural status is about 0.76 standard deviations in the OECD countries, roughly equivalent to two years' worth of schooling for that age group. At the same time, some low SES students manage to excel in PISA, and while low SES students do worse in all countries both the strength of the relationship between SES and test scores, and the share of resilient students differ markedly (OECD, 2010). These results indicate that overcoming a disadvantaged background is possible, and that the more interesting question is how low SES children and students can best be helped. This review will provide evidence in relation to this question by systematically examining which interventions preschools, schools, and local stakeholders can use to improve the educational achievements of children and students from disadvantaged socioeconomic backgrounds.

1.2 POTENTIAL EXPLANATIONS FOR SES STUDENTS' LOW EDUCATIONAL ACHIEVEMENT AND THE NEED FOR INTERVENTIONS

To provide a framework for the review, we present a very short overview of potential explanations for why low SES students perform less well academically and relate these explanations to the design of interventions.

A first possible explanation could be that low SES students have lower innate ability. The separation of hereditary from environmental effects is inherently difficult; especially when epigenetic effects – that is, heritable genetic changes that are not caused by changes in the DNA sequence but by environmental factors – might be present.[1] However, there is recent evidence that measures of mental ability differ neither significantly nor substantially between high and low SES children in the early ages. Tucker-Drob et al. (2011) found no significant differences on tests of infant mental ability at the age of 10 months between children in families with high and low SES. At age two however, children in high SES families scored significantly higher. One standard deviation higher SES was associated with about one third of a standard deviation higher mental ability. Genes accounted for nearly 50 percent of the variation in mental ability of children raised in high-SES homes, but only a negligible

1. For the presence of epigenetic effects, see e.g. Fraga et al. (2005) and Hackman and Farah (2009). For a nuanced discussion of the interplay between nature and nurture in child development, see Rutter (2006).

share of the variation in mental ability of children raised in low-SES homes. These results suggest that the environment is the constraining factor for low SES children. Similarly, the differences in educational achievement and test scores between black and white American children have been found to be about one standard deviation already at age 3. However, looking at differences between infants 8 to 12 months old, Fryer and Levitt (2013) found no significant differences between Hispanics, Asians, Blacks, and Whites. Lastly, early poverty is a better predictor of later cognitive achievement than poverty in middle or late childhood, which is hard to explain by differences in innate abilities (Hackman & Farah, 2009).

These results indicate that large and significant differences are present already well before children start school, which also provide partial evidence against another potential explanation: schools may be worse at helping low SES students. Heckman (2006) argue that schools are not the major source of inequality in student performance, as gaps in test scores across socioeconomic groups are stable from third grade and onwards. Later schooling and variations in schooling quality seem to neither reduce nor increase the gaps. Further evidence is provided by a literature documenting seasonality in the performance gaps. At least in the US, the gap between students with high and low SES widen during summer breaks; that is, when children do not have access to school resources (Alexander et al., 2001; Gershenson, 2013; Kim & Quinn, 2013). Moreover, schools with comparatively large shares of low SES students receive more resources in most OECD countries (OECD 2010).[2] However, even if schools do not seem to be the most important source of differences between high and low SES students, this does not imply that schools cannot be an important part of programs that can increase educational achievement for low SES students. Indeed programs targeted specifically towards the needs of low SES students hold potential to reduce or overcome the gap in educational achievement.

Given that differences in innate ability and school quality do not seem to explain most of the differences between high and low SES students, the early childhood environment is likely to be an important explanation. Currie (2009) surveys a large literature documenting that low SES children have worse health on a very broad range of measures, including fetal conditions, health at birth, incidence of chronic conditions, and mental health problems. The relationship seems to exist also in some countries with universal health care systems, such as Canada and the UK. Reviewing interventions that aim to increase only the health of low SES students will be outside the scope of this review, but the evidence do suggest that child health problems influence educational and labor mar-

2. The US is an exception, at least regarding the student/teacher ratio (OECD 2010).

ket outcomes. The effects seem to be smaller for educational outcomes compared to future earnings though (Currie, 2009).

Other explanations for the relatively poor performance of low SES students are likely to be found in their homes and neighborhoods. That is, low SES students lack family resources and live in an environment less conducive to high educational achievement (Jacob & Ludwig 2008). Family resources could include for example providing a rich language and literacy environment, and different parenting practices, but also money spent on early childhood education (Esping-Andersson et al., 2012). In the US at least, poor children are less likely to attend center-based care during preschool ages. More resources could also buy goods and services that may influence academic achievement, such as health care, nutrition, and enriching spare-time activities. Being poor is also likely to increase levels of stress, frustration and depression and may therefore also increase punitive and harsh parenting practices (Magnuson & Shager, 2010). Low SES parents also seem to have lower expectations for their children (something which may also be true for teachers) (Bradley & Corwyn, 2002).

The earlier literature thus point to several domains where low SES students lack resources or are disadvantaged. These domains are likely to correspond to the primary mechanisms of intervention effects: cognitive development, social adjustment (or prosocial behavior), family support, motivational support, increased expectations, and school and preschool support (Reynolds & Temple, 2008; Reynolds et al., 2010). Furthermore, if the differences between high and low SES students can be understood as a consequence of the lack of a *combination* of resources, remedial efforts may need to address several problems at once to be effective. Programs that combine certain components may therefore be more effective than others.[3] One aim of the review is to examine whether the domains interventions address, and whether combinations of components of interventions, influence the effect sizes of interventions.

1.3 RATIONALE FOR THE CURRENT STUDY

Given the importance of education for earnings, health, and well-being, finding interventions that effectively improve the educational achievement of disadvantaged children is of considerable importance, and a high priority for governments around the world (UNESCO, 1994). Many as-

3. For example, Cook et al. (2014) find large positive effects on test scores of a program for 9-10th graders that combined math tutoring with the development of social-cognitive skills. They argue that one reason for the lack of successful interventions to older disadvantaged students may be that they have only addressed one problem at a time, while this group is likely to face several problems simultaneously.

pects of interventions targeting educational outcomes for low SES children and students have accordingly been reviewed before. Below we focus on the most recent and related reviews.

High quality preschool programs, primarily directed to disadvantaged groups in the US, have been found to have positive effects on cognitive development including educational outcomes, as well as social skills (e.g. Blau & Currie, 2006; Camilli et al, 2010; Duncan & Magnuson, 2013, Yoshikawa et al., 2013).[4] Camilli et al. (2010) provide a meta-analysis of studies from the US and interventions performed before or during the year 2000. While they do not target low SES children in particular, a very large share of the studies include such children. They find the largest effect sizes for cognitive outcomes and that aspects of interventions such as teacher-directed instruction and small-group instruction are positively correlated with effect sizes. In a review and meta-analysis of studies from 1960 to 2007 (again only including US studies), Duncan and Magnuson (2013) describe a pattern where many early childhood programs appear to increase cognitive ability in the short run, whereas the effects seem to fade out during the first few years after the programs end.[5] However, beneficial effects on outcomes such as educational attainment, earnings, and crime rates reappear in the few studies of long-term effects that exist. The effects also differ substantially between preschool programs. Well-known and highly targeted programs such as the Perry Preschool and Abecedarian programs show larger effects. Newer evaluations generally find smaller effect sizes, something which the authors attribute to improved counterfactual conditions for children not attending a preschool program. Both Duncan and Magnuson (2013) and Camilli et al. (2010) recommended that future research should prioritize finding the connections between program components and particular child outcomes.

Chambers et al (2010) reviewed 27 early childhood programs that can be implemented in a preschool setting and aim to prepare chil-

4. Regarding evidence from other countries, expansions of universal preschool have also been shown to increase earnings in Norway (Havnes & Mogstad, 2011), in particular for low- to middle income groups (Havnes & Mogstad, 2014), decrease language gaps between immigrant and native children in Sweden (Fredriksson et al., 2010), increase educational attainment in Uruguay (Berlinski et al 2008), improve test scores and student's self-control in Argentina (Berlinski et al., 2009), and increase PISA scores in Spain (Felfe et al., 2012). However, a similar expansion in Quebec, Canada, affected short-run cognitive and non-cognitive child outcomes negatively (Baker et al., 2008). These types of expansions will not be included in the review as they were performed far back in time, and/or the mode of care is not clear for treatment and/or control groups.

5. Whether this pattern is particular to the US, or can be found also in other countries is currently not well-known. As the school system in the US differs from many other OECD countries (low SES students e.g. receive less school resources in some areas, see footnote 3), this may be a US phenomenon. For example, the effects of class size reductions in the Tennesee STAR experiment on test scores fade out in later grades, but increase e.g. college attendance (Chetty et al. 2011). The effects of class size reductions in Sweden have been shown to be sustained throughout the school years, and increase earnings (Fredriksson et al., 2013).

dren for success in primary school and beyond. The programs are directed to children between the ages 3-5, who are at risk of school failure due to poverty. All but one study comes from the US. Six programs were deemed to have strong evidence of effectiveness, and five to have moderate evidence of effectiveness. It is notable though that no program have been evaluated in more than three studies (one study), and most only once. The review does not examine which components of the programs that influence the effects.

Bridging the preschool and school areas, Reynolds & Temple (2008) and Reynolds et al. (2010) reviewed preschool to third grade programs and practices (again only including studies from the US). Both showed results in line with the reviews described above for preschool programs, but add that interventions such as small class sizes in the early grades also yield positive effects and that long run cost-benefit analyses indicate that many programs have positive, and quite sizeable returns.

Regarding interventions in school ages that include mostly or only low SES students, Zief et al. (2006) review after-school programs and find few studies and little evidence that these have a positive effect. Wilson et al. (2011) reviews school completion and dropout prevention programs, where a large share of the target population was low SES students, and find large positive effects in general. A major difference between their review and the current review is that they focused on outcome measures such as dropout and high school graduation rates. Our review will use test scores, grade point averages, and take up of high school/upper secondary school. Kim & Quinn (2013) review summer reading programs, and find positive effects for interventions that employed research-based reading instruction and included a majority of low-income children.

The reviews of Slavin & Lake (2008) (elementary mathematics programs), Slavin, Lake & Groff (2009) (middle and high school mathematics programs), and Slavin, Lake, Chambers, Cheung, & Davis (2009) (reading programs for elementary grades) do not target low SES children directly, but they found no indications that the overall positive effect sizes differ between low SES students and non-disadvantaged students. However, far from all studies report results for low SES students. The reviews do not contain information about whether the programs that in general show the largest effect sizes - instructional-process programs that e.g. include cooperative learning, classroom management and motivation programs, and supplemental tutoring programs - also have the largest effect sizes for disadvantaged students.

Low achieving students may to some degree overlap with our target population. Wanzek et al. (2006) reviewed reading programs directed to students in grades K-12 with learning disabilities, and Edmonds et al. (2009), Flynn et al. (2012), and Scammaca et al. (2013) reviewed

programs for struggling readers in grades 6-12, 5-9, and 4-12, respectively. These reviews reported positive effects in general but few reliable differences over types of interventions. Slavin et al. (2011), who also focused on programs directed to struggling readers, did find higher effect sizes for instructional process programs.

Most reviews have included similar types of treatment-control/comparison study designs as we will (see section 2 for more details), and they also included both randomized experiments, and quasi-experimental studies.[6] The question of what program components and combinations of components are important for low SES students is not settled in the reviews covered in this section. Our comparatively wider scope provides better possibilities to examine moderators that influence the effect sizes of intervention programs, and in turn to provide guidance about what components of interventions are effective.

Most of the previously mentioned reviews do not report the cost-effectiveness of programs either, something this review will assess to the extent that enough studies include this information. Narrative reviews have indicated that very few types of preschool or school-based interventions pass a cost-benefit test in the long run, with some high quality preschool programs, class size reductions in the early grades, and bonuses to attract and retain the highest-quality teachers being among the exceptions (Jacob & Ludwig, 2008; Reynolds & Temple, 2008; Reynolds et al., 2010). However, the long run cost-benefit ratio is hard to assess for many interventions, as many control groups are waitlist controls or receive other remedial treatments after the initial treatment, and small scale interventions may only require a small amount of attrition over time to make them difficult to analyze. Furthermore, many interventions are too recent in time to evaluate the effects on for example earnings. There is, to the best of our knowledge, no comprehensive review that provides guidance to policy makers on how scarce resources should be allocated between different types of interventions in the short run (Kim & Quinn, 2013; but see e.g. Cook et al., 2014 for a recent discussion of a few programs).

Lastly, as indicated, most reviews contain studies where a large majority is from the US. This may be because there are not many studies from other countries, but we hope to be able to find studies from a more diverse set of countries.

6. Wanzel et al. (2006) and Edmonds et al. (2009) also include single-group and single-subject pre- and post-test design, which we will not include.

CHAPTER 2

METHOD

This section describes the data collection process in the following steps: Section 2.1 discusses the definition of low SES. Section 2.2 presents the criteria we use for including and excluding studies in the review. Section 2.3 describes the search strategy, including an example of search terms. Section 2.4 discusses the data extraction process, including screening and coding practices. Section 2.5 describes how we will assess the risk of bias of included studies. Lastly, section 2.6 describes the synthesis procedures.

2.1 DEFINITION OF SOCIOECONOMIC STATUS

While there is no consensus on an exact definition, most researchers seem to agree on a tripartite nature of the concept of SES, which incorporates parental income, parental education, and parental occupation as its three main indicators (Sirin, 2005). We will broadly adhere to this definition, but the search will include many other terms used in the previous literature to capture populations with low SES (see section 2.3.2 below). Some researchers have suggested that a more narrow definition (e.g. low income) is feasible when coding studies for meta-analyses, as primary studies in some fields very seldom use composite measures of SES (e.g. Kim & Quinn, 2013). As we do not know whether this is the case for all types of interventions that we aim to find, we prefer to keep the search strategy broad. But depending on the studies we find, the actual coding of SES categories might use fewer categories than what the search strategy indicates.

2.2 CRITERIA FOR INCLUSION AND EXCLUSION OF STUDIES IN THE REVIEW

We will select studies based on the type of intervention, participants, outcome measures, study designs, and settings. See Appendix A for the questions that will guide the first and second level screening.

2.2.1. TYPES OF INTERVENTIONS

Interventions should explicitly aim to improve educational achievement, including the take up of youth education, school readiness, and/or specific academic skills. This does not mean that the intervention must consist of academic activities, but the aim should be to improve academic performance or skill levels in specific academic tasks. Programs that primarily aim to reduce for example criminal behavior or bullying, or to improve pro-social skills, and only have improved academic outcomes as secondary objectives, will be excluded.

To be included, the intervention should be implemented by individual schools or preschools, or by schools or preschools in cooperation with outside local stakeholders. We exclude studies that require changes to the entire school system, such as changes to the grade system, school start and leaving ages, the national/regional curriculum, the introduction of national standardized tests, the introduction or expansion of school choice, private schools, and the expansion of universal access to preschools and other reforms aimed to increase preschool attendance. We also exclude early childhood interventions performed outside of regular preschools. Outside local stakeholders could for example be local governments, NGOs, and researchers.

2.2.2 TYPES OF PARTICIPANTS

To be included, interventions should, at least partly, be aimed at students identified in the study under consideration on the basis of having low SES, measured as e.g. family income, education, and occupational (including on social welfare) or minority status. This does not mean that studies including other students should be excluded by default, but if the intervention includes other students as well, results for students with low SES should be reported separately. Some studies may only report school or district statistics on SES, instead of individual information on participants (Kim & Quinn, 2013). We will use the aggregate information to code these samples, i.e. assume that the school or district statistics is reflected in the study population, but we will check the sensitivity of our results to the inclusion of such studies.

We will include students and children in preschool to grade 10 that attend regular private, public, and boarding schools and preschools. Interventions in high school/upper secondary school will be excluded

(grade 10 is not included in high school/upper secondary school in some countries). Interventions targeting students receiving special education services within these school settings will also be included. Interventions for students attending special education schools outside a regular school setting will be excluded.

2.2.3 TYPES OF OUTCOME MEASURES

To be included, studies should use one or more of the following primary outcome variables: standardized academic tests (e.g. Iowa Test of Basic Skills, Stanford Achievement Test), specific measures designed to measure preschoolers' school readiness (e.g. Metropolitan Readiness Test, Peabody Individual Achievement Tests), grade point averages, or take up rates of upper secondary education/high school.

We restrict our attention to standardized tests mainly for two reasons: first, earlier reviews of academic interventions have pointed out that effect sizes tend to be significantly lower for standardized tests compared to researcher-developed tests (e.g. Flynn et al., 2012; Scammaca et al., 2013). Second, Scammaca et al. (2013) also found that while publication year was a significant predictor of effect size when all measures were considered – earlier publications tend to report larger effect sizes – this was not so for standardized test measures.

If reported, we will also record the costs per participant of each intervention. If enough studies report costs, we will also perform an analysis of the cost-effectiveness of interventions. Studies will not be excluded on account of not reporting the costs of an intervention.

2.2.4 TYPES OF STUDY DESIGNS

We will limit ourselves to study designs that employ a treatment-control or a treatment-comparison group design. A control group is defined as a non-treatment condition, which includes waitlist controls. A comparison group receives an alternative treatment. We will code treatment-comparison group designs separately, and they will be treated in a meta-analysis separate from treatment-control study designs. We will include both randomized controlled trials (RCT), quasi-randomized controlled trials (QRCT), i.e., where participants are located by means such as alternate allocation, person's birth date, the date of the week or month, case number, or alphabetical order; and quasi-experimental studies (QES). QES can include e.g. difference-in-differences designs, matching or statistical controls; that is, QES use some form of non-experimental technique to mitigate selection bias. Studies using instrumental variables (IV) to estimate a local average treatment effect (LATE) (Angrist & Pischke, 2009) will be included, but may be subject to a separate analysis depending on the comparability between the LATE's and the effects from other

studies. We will in any case check the sensitivity of our results to the inclusion of IV studies.

A fair amount of studies within educational research use single group pre-post comparisons (e.g. Wanzek et al., 2006; Edmonds et al., 2009); such studies will not be included. We will also include only primary research, reviews will be excluded.[7]

2.2.5 TYPES OF SETTINGS

Only studies of interventions carried out in regular schools and preschools in the OECD and EU countries will be included. This selection is conducted to ensure a certain degree of comparability between settings to align treatment as usual conditions in included studies. Due to language constraints, we will have to restrict ourselves to studies written in English, German, Danish, Norwegian, and Swedish.

2.2.6 SCOPE OF INTERVENTION YEAR AND DURATION OF FOLLOW-UP

Our starting point is to include interventions performed in or after the year 2000. This choice of starting year is due to our expected resource constraints. Should we have the resources, we will consider extending the period restriction backwards in time.

We will not apply any restriction on the duration of follow-up measurements, apart from those that follow from our choice of outcome measures. That is, we will not be able to analyze the effects of interventions on earnings or take up of college or university education.

2.3 SEARCH STRATEGY

2.3.1 ELECTRONIC SEARCHES

Relevant studies will be identified through electronic searches of bibliographic databases, government and policy databanks. The following bibliographic databases will be searched:

- Campbell Library
- Centre for Reviews and Dissemination Databases
- Cochrane Library
- EconLit
- Education Research Complete
- ERIC - Education Resource Information Center
- PsycINFO

7. Reviews will be retained and manually searched for references, but they will not be included in the analysis.

- SocIndex
- Social Care Online
- Forskningsdatabasen.dk (Danmark)
- Diva-portal.org, Libris (Sverige)
- Cristin (Norge), Current Research Information System In Norway

2.3.2 SEARCH TERMS

An example of the search strategy for ERIC searched through the EBSCO platform is listed below. This strategy will be modified for the different databases. We will report details of the modifications used for other databases in the completed review.

S1	DE "Preschool Children" OR (child* n3 preschool) OR (Child* n3 daycare) or (child*n3 day-care) or (Child n3 (day n1 care)) or (child n2 prekindergarten) or (child n3 prekindergarten) or (child n3 pre n1 kindergarten) or (child n3 nursery school*)
S2	((Primary N1 School) n3 (Student* or pupil*)) or ((Elementary N1 School) n3 (Student* or pupil*)) or (DE "Elementary School Students") or ((Secondary N1 school) or (high N2 school) or (middle N1 School) N3 (student* or pupil*))
S3	(at-risk or at N1 risk)) N1 (student* or pupil*) or ((high-risk or high N1 risk) N1 (student* or pupil*)) or ((Special N1 Need*) N1 (Student* or pupil*)) or ((Low N1 income) N1 (student* or pupil*))
S4	inner city schools or Urban districts
S5	non-employed
S6	un employed
S7	unemployed
S8	"Vulnerable population*"
S9	"Adverse social characteristics*"
S10	"Adverse social background*"
S11	"Adverse social environment"
S12	"Low social capital"
S13	"Low socioeconomic status"
S14	"Low socioeconomic background"
S15	"at risk"
S16	"reduced lunch" AND eligib*
S17	"free lunch" AND eligib*
S18	free lunch
S19	Eligible for free lunch
S20	On federal benefits OR On federal payments
S21	On state benefits OR On state payments

S22	"Ethnic group*"
S23	"Minority status"
S24	"Minority group*"
S25	Minority group*
S26	"low earning*"
S27	"low wage*"
S28	"low resources"
S29	"low status"
S30	"low education"
S31	low education
S32	high poverty
S33	low income
S34	underserved
S35	needy
S36	Underprivileged
S37	Disparities
S38	Disadvantaged
S39	Impoverished
S40	Poor
S41	deprived
S42	DE "Economically Disadvantaged"
S43	DE "Educationally Disadvantaged"
S44	DE "Disadvantaged Youth"
S45	DE "Disadvantaged"
S46	DE "Low Income"
S47	DE "Minority Group Children"
S48	DE "Minority Group Students"
S49	DE "Ethnic Groups"
S50	DE "American Indian Students"
S51	DE "Welfare Recipients"
S52	DE "Socioeconomic Background"
S53	DE "Socioeconomic Status"
S54	DE "Social Capital"
S55	DE "Unemployment"
S56	S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10 OR S11 OR S12 OR S13 OR S14 OR
S15	OR S16 OR S17 OR S18 OR S19 OR S20 OR S21 OR S22 OR S23 OR S24 OR S25 OR S26 OR S27 OR S28 OR S29 OR S30 OR S31 OR S32 OR S33 OR S34 OR S35 OR S36 OR S38 OR S39 OR S40 OR S41 OR S42 OR S43 OR S44 OR S47 OR S48 OR S52 OR S53 OR S54 OR S55
S57	((S1 or S2) AND S56) OR S3
S58	transfer* N2 effect

S59	(DE “School Readiness” OR (school N1 (ready OR readiness))
S60	Write or writing or DE "Writing Ability" or DE "Writing Achievement"
S61	Numeracy or Mathematic* or Math
S62	DE "Mathematics" or DE "Numeracy"
S63	Reading or Literacy
S64	DE "Reading" or DE "Literacy"
S65	S95 OR S96 OR S97 or s98
S66	Program* or intervent*
S67	School N1 (performan* or achiev*)
S68	Academic* N2 (performance* or achiev* or abilit* or outcome*)
S69	DE "Academic Achievement" or DE "Academic Ability"
S70	(S58 or S59 or S60 or S61 or S62 or S63 OR S64 OR S65 OR S66 OR S67 OR
S68	OR S69
S71	TI ((control N5 case) or (control N5 subject*) or (control N5 group*) or (control N5 patient*) or (control N5 intervention)) OR AB ((control N5 case) or (control N5 subject*) or (control N5 group*) or (control N5 patient*) or (control N5 intervention))
S72	TI ((treatment N5 case) or (treatment N5 subject*) or (treatment N5 group*) or (treatment N5 patient*) or (treatment N5 intervention)) OR AB ((treatment N5 case) or (treatment N5 subject*) or (treatment N5 group*) or (treatment N5 patient*) or (treatment N5 intervention))
S73	TI ((experiment* N5 case) or (experiment* N5 subject*) or (experiment* N5 group*) or (experiment* N5 patient*) or (experiment* N5 intervention)) OR AB ((experiment* N5 case) or (experiment* N5 subject*) or (experiment* N5 group*) or (experiment* N5 patient*) or (experiment* N5 intervention))
S74	TI ((intervention N5 case) or (intervention N5 subject*) or (intervention N5 group*) or (intervention N5 patient*)) OR AB ((intervention N5 case) or (intervention N5 subject*) or (intervention N5 group*) or (intervention N5 patient*))
S75	((assign* N5 case) or (assign* N5 subject*) or (assign* N5 group*) or (assign* N5 patient*) or (assign* N5 intervention)) OR AB ((assign* N5 case) or (assign* N5 subject*) or (assign* N5 group*) or (assign* N5 patient*) or (assign* N5 intervention))

S76	TI (quasi-experiment* or quasiexperiment* OR Propensity score* or (compar* N1 group*) or (match* N1 control*) OR (match* N1 group*) OR (match* N1 compar*) OR experiment* trial* OR experiment* design* OR experiment* method* OR experiment* stud* OR experiment* evaluation* OR experiment* test* OR experiment* assessment* OR assessment only OR (comparison n1 samp*) OR propensity match* or (Between N1 group*) or longitud*) OR AB (quasi-experiment* or quasiexperiment* OR Propensity score* or (compar* N1 group*) or (match* N1 control*) OR (match* N1 group*) OR (match* N1 compar*) OR experiment* trial* OR experiment* design* OR experiment* method* OR experiment* stud* OR experiment* evaluation* OR experiment* test* OR experiment* assessment* OR assessment only OR (comparison n1 samp*) OR propensity match* or (Between N1 group*) or longitud*)
S77	TI ((random* N2 trial*) or RCT) OR AB ((random* N2 trial*) or RCT)
S78	TI Non-random* or nonrandom* or (non N1 random*) OR AB Non-random* or nonrandom* or (non N1 random*)
S79	TI (Propensity score* or (match* N1 control*) or (match* N1 compar*) or assessment only or comparison samp* or propensity match*) OR AB (Propensity score* or (match* N1 control*) or (match* N1 compar*) or assessment only or comparison samp* or propensity match*)
S80	TI assign* N3 (subject* or patient*) or AB assign* N3 (subject* or patient*)
S81	Ti (quasi-experiment* or quasiexperiment* or experiment*) OR AB (quasi-experiment* or quasiexperiment* or experiment*)
S82	TI Intervention* N1 Stud* OR AB Intervention* N1 Stud*
S83	TI retrospective OR AB retrospective
S84	TI (prospective n2 study) or AB (prospective n2 study)
S85	TI observational OR AB observational
S86	TI longitudinal or AB longitudinal
S87	Ti ((follow up or followup) N2 study) OR AB ((follow up or followup) N2 study)
S88	TI (epidemiol* N2 stud*) or AB (epidemiol* N2 stud*)
S89	TI cross sectional or AB cross sectional
S90	TI cohort OR AB cohort

S91	TI (case control) or AB (case control)
S92	DE "Case Studies"
S93	DE "COHORT analysis"
S94	TI groups or AB groups
S95	TI trial or AB trial
S96	TI randomly or AB randomly
S97	TI placebo or AB placebo
S98	ti randomi?ed or AB randomi?ed
S99	TI (regression N1 discontinuity OR difference-in-difference* OR event N1 stud* OR interrupted time serie* OR instrumental variable* OR waitlist control*) OR AB (regression N1 discontinuity OR difference-in-difference* OR event N1 stud* OR interrupted time serie* OR instrumental variable* OR waitlist control*)
S100	S71 OR S72 OR S73 OR S74 OR S75 OR S76 OR S77 OR S78 OR S79 OR S80 OR S81 OR S82 OR S83 OR S84 OR S85 OR S86 OR S87 OR S88 OR S89 OR S90 OR S91 OR S92 OR S93 OR S94 OR S95 OR S96 OR S97 OR S98 OR S99
S101	S57 AND S70 AND S100
S102	(S101) Limiters - Date Published: 20000101-20141231

2.3.3 SNOWBALLING

The reference lists of relevant articles will be searched in order to possibly identify more relevant studies.

2.3.4 HANDSEARCH

The most recent year of the following journals will be handsearched:

- American Educational Research Journal
- Journal of Educational Research
- Learning and Instruction
- Journal of Educational Psychology

2.3.5 GREY LITERATURE

OpenGrey (http://www.opengrey.eu/) will be searched for European grey literature. The following websites will be manually searched:

- What Works Clearinghouse - U.S. Department of Education, http://www.whatworks.ed.gov
- Dansk Clearinghouse for Uddannelsesforskning, http://edu.au.dk/clearinghouse/
- European Educational Research Association (EERA), http://www.eera-ecer.eu/

- American Educational Research Association (AERA), http://www.aera.net
- Deutsche Gesellschaft für Erziehungswissenschaft (DGfE), German Educational Research Association (GERA), http://www.dgfe.de/
- Skolverket, http://Skolporten.com (Sweden)
- Forskning.no (Norway)

2.4 DATA EXTRACTION

Under the supervision of review authors, review team assistants will first independently screen titles and abstracts to exclude studies that are clearly irrelevant. Studies considered eligible will be retrieved in full text. The full texts will then be screened by review team assistants under the supervision of the review authors. Any uncertainty of eligibility will be resolved by the review authors. The study inclusion criteria (see Appendix A) will be piloted by the review authors and the review team assistants together. The overall search and screening process will be illustrated in a flow-diagram.

The review authors will code and extract data from included studies. A coding sheet will be piloted on several studies and revised as necessary. Data will be extracted on the characteristics of participants (e.g. age, gender), characteristics of the intervention and control/comparison conditions, research design, sample size, outcomes, and results. Extracted data will be stored electronically.

We will code included articles using the following rough outline of categories and variables: report characteristics (year, language, publishing status), study characteristics (objectives, intervention year, study location, study design, participant characteristics), intervention characteristics (name, domain, type, site, delivery, duration, frequency, intensity, provider), sample size (group size, sample size for outcome measurements), outcome measurement (timing, tools, assessment periods), and outcomes (effect sizes, costs). Coding categories and variables may be subject to change after we have piloted a scheme on sample of included articles.

2.5 RISK OF BIAS

We will assess the methodological quality of studies using a risk of bias model developed by Prof. Barnaby Reeves in association with the Cochrane Non-Randomized Studies Method group. This model, an extension of the Cochrane Collaboration's risk of bias tool, covers risk of bias both in RCTs and in non-randomized studies that have a well-

defined control or comparison group. The extended model is organized as, and follows the same steps as, the risk of bias model described in the Cochrane Handbook, chapter 8 (Higgins & Green, 2011). The model is extended as follows:

1. The existing Cochrane risk of bias tool needs elaboration when assessing non-randomized studies because particular attention must be paid to selection bias and risk of confounding. The extended model therefore specifically incorporates a formalized and structured approach for the assessment of selection bias in non-randomized studies by adding an explicit item that focuses on confounding. This is based on a list of confounders considered important and defined in the protocol for the review. The assessment of confounding is made using a worksheet which is marked for each confounder according to whether it was considered by the researchers, the precision with which it was measured, the imbalance between groups, and the care with which adjustment was carried out (see Appendix B). This assessment will inform the final risk of bias score for confounding.
2. RCTs should have a protocol that is defined prior to commencing recruitment, whereas non-randomized studies usually does not have such a protocol. This makes non-randomized studies at greater risk of bias compared to RCTs. The item concerning selective reporting therefore also requires assessment of the extent to which analyses (and potentially other choices) could have been manipulated to bias the findings reported (for example, by the choice of method of model fitting, and by the potential confounders that are considered). In addition, the model include two separate yes/no items asking review authors whether they judge the study investigators to have had a pre-specified protocol and analysis plan.
3. The risk of bias assessment is refined, making it possible to discriminate between studies with varying degrees of risk. This refinement is achieved by the use of a 5-point scale for certain items (see the following section Risk of bias judgment items for details). The refined assessment is pertinent when considering data synthesis as it operationalizes the identification of those studies with a very high risk of bas (especially in relation to non-randomized studies). This refinement increases transparency in assessment judgments and provides justification for excluding a study with a very high risk of bias from the meta-analysis.

2.5.1 RISK OF BIAS JUDGMENT ITEMS

The risk of bias model used in this review is based on 9 items (see Appendix B). The 9 items refer to:

- Sequence generation (judged on low risk/high risk/unclear scale).
- Allocation concealment (judged on low risk/high risk/unclear scale).
- Confounders (judged on a 5-point/unclear scale).
- Blinding (judged on a 5-point/unclear scale).
- Incomplete outcome data (judged on a 5-point/unclear scale).
- Selective outcome reporting (judged on a 5-point/unclear scale).
- Other potential threats to validity (judged on a 5-point/unclear scale).
- A priori protocol (judged on a yes/no/unclear scale).
- A priori analysis plan (judged on a yes/no/unclear scale).

2.5.2 CONFOUNDING

An important part of the risk of bias assessment of non-randomized studies is how the studies deal with confounding factors (see Appendix B). Selection bias is understood as systematic baseline differences between groups and can therefore compromise comparability between groups. Baseline differences can be observable (e.g. age and gender) and unobservable to the researcher (e.g. ability). There is no single non-randomized study design that always deals adequately with the selection problem. Different designs represent different approaches to dealing with selection problems under different assumptions and require different types of data. There can be particularly great variations in how different designs deal with selection on unobservables. The "adequate" method depends on the model generating participation, i.e. assumptions about the nature of the process by which participants are selected into a program.

For this review, we have identified the following observable confounding factors to be most relevant: grade level (or age), performance at baseline, gender, and socioeconomic background.[8] In each study, we will assess whether these confounding factors have been considered, and in addition we will assess other confounding factors considered in the individual studies. Furthermore, we will assess how each study deals with unobservables.

2.6 SYNTHESIS PROCEDURES AND STATISTICAL ANALYSIS

2.6.1 EFFECT SIZE CALCULATIONS

For dichotomous outcomes we will use the natural logarithm of odds ratios (LOR) or risk ratios (LRR) in the calculations, together with 95%

8. Although the review will focus on interventions directed towards children and students with low SES, this concept may be of different magnitudes and types for participants. It may therefore still be an important confounder.

confidence intervals and p-values, and then convert the results back to the original odds and risk ratios once the meta-analysis is performed. The LOR and its approximate standard deviation are calculated as (Lipsey & Wilson, 2001:53-54):

$$LOR = log\left(\frac{ad}{bc}\right), SE_{LOR} = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

where *a* is the frequency of "good" outcomes in the treatment group (e.g. the frequency of students passing a test), *b* is the frequency of "bad" outcomes in the treatment group (the frequency of students not passing), and *c* and *d* are the frequencies of good and bad outcomes in the control group, respectively. The risk ratio and its approximate standard deviation are calculated as (Borenstein, Hedges, Higgins & Rothstein, 2009:34):

$$LRR = log\left(\frac{a/n_1}{c/n_2}\right), SE_{LRR} = \sqrt{\frac{1}{a}-\frac{1}{n_1}+\frac{1}{c}-\frac{1}{n_2}}$$

where n_1 and n_2 are the sample sizes in the treatment and control group, and all other terms are defined as before.

For continuous data, standardized mean differences (SMDs) will be calculated with 95% confidence intervals when means and standard deviations are available. We will use Hedges' *g* to estimate SMDs where scales have been used to measure the same outcomes in different ways. Hedges' *g* and its standard error are calculated as (Lipsey & Wilson, 2001:47-49):

$$g = \left(1-\frac{3}{4N-9}\right)\times\left(\frac{\bar{X}_1-\bar{X}_2}{s_p}\right), SE_g = \sqrt{\frac{N}{n_1 n_2}+\frac{g^2}{2N}}$$

where $N = n_1 + n_2$ is the total sample size, $\bar{X}$ is the mean in each group, and s_p is the pooled standard deviation defined as

$$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1-1)+(n_2-1)}}.$$

Here, s_1 and s_2 denotes the standard deviation of the treatment and control group. If there is a mix of studies with some reporting change scores and others reporting final values, we will contact the trial investigators and request the final values. If these are unobtainable, we will analyze change scores and final values separately.

We will conduct the analysis separately for dichotomous dependent variables if there are enough such studies, and transform dichotomous effect sizes to SMD otherwise. If transformation is necessary, we will use the methods suggested by Sáchez-Meca et al. (2003) to allow dichotomous and continuous data to be pooled together.

We will use covariate adjusted SMDs and odds ratios whenever available.

2.6.2 OUTLIERS

We will examine the distributions of effect sizes for each outcome category for the presence of outliers. If outliers are found and deemed unrepresentative, then, depending on the nature of the outlier studies, we will examine the sensitivity of the results by methods suggested by Lipsey & Wilson (2001): trimming the distribution by dropping the outliers and/or by Windsorizing the outliers to the nearest non-outlier value.

2.6.3 DEALING WITH MISSING DATA

Missing data and attrition rates in the individual studies will be assessed using the risk of bias tool. Furthermore, for RCTs and QRCTs we will record whether intention to treat analysis (ITT) was conducted. Sensitivity analysis will be performed to examine the impact of excluding trials in which adequate ITT analysis was not used.

Studies must permit calculation of a numeric effect size for the outcomes to be eligible for inclusion in the meta-analysis. Where studies have missing summary data, such as missing standard deviations, we will derive these where possible from e.g. F-ratios, t-values, chi-squared values and correlation coefficients using the methods suggested by Lipsey & Wilson (2001). If these statistics are also missing, the review authors will request information from the study investigators. If missing summary data cannot be retrieved within two weeks, the study results will be reported in as much detail as possible, i.e. the study will be included in the review but excluded from the meta-analysis.

2.6.4 CLUSTER RANDOMIZED TRIALS

Errors in statistical analysis can occur when the unit of allocation differs from the unit of analysis. In cluster randomized trials, participants are randomized to groups in clusters, either when data from multiple participants in a setting are included (creating a cluster within the school or community setting), or when participants are randomized by treatment locality or school. In such studies, standard errors may be biased if the unit-of-analysis is the individual. When an appropriate cluster analysis has been used (e.g. cluster summary statistics, robust standard errors), effect estimates and their standard errors will be meta-analysed (Higgins & Green, 2011). In cases where trial investigators have not applied ap-

propriate statistical methods to control for clustering, we will attempt to estimate the intra-cluster correlation coefficient (ICC) and correct standard error (Donner, Piaggio, & Villar, 2001). If the ICC is not available in the study affected, we will use external estimates obtained from the literature (e.g. from similar studies).

2.6.5 MULTIPLE INTERVENTION GROUPS AND MULTIPLE INTERVENTIONS PER INDIVIDUAL

Studies with multiple intervention groups with different individuals will be included in this review. To avoid problems with dependence between effect sizes we will apply robust standard errors (Hedges, Tipton, & Johnson, 2010). However, simulation studies show that this method needs around 20-40 studies included in the data synthesis (Hedges et al., 2010). If this number cannot be reached we will conduct a data synthesis where we use a synthetic effect size (the average) in order to avoid dependence between effect sizes. Studies including multiple interventions per individuals will be included. Only one intervention group (control group) will be coded and compared to the control group (intervention group) to avoid overlapping samples.

2.6.6 MULTIPLE STUDIES USING THE SAME SAMPLE OF DATA

In some cases, several studies may have used the same sample of data, e.g. studies using the same administrative data. We will review all such studies, but will only include in the meta-analysis one estimate of the intervention effect from each sample of data to avoid dependencies between the estimates of the intervention effect. The choice of which estimate to include will be based on our quality assessment of the studies. We will choose the estimate from the study that we judge to have the least risk of bias using the risk of bias tool (see Appendix B).

2.6.7 DATA SYNTHESIS

The overall data synthesis in this review will be conducted where effect sizes are available or can be calculated, and where studies are similar in terms of the outcome measured. Random effects inverse variance weighted mean effect sizes will be used and we will report the 95% confidence intervals and provide a graphical display (forest plot) of effect sizes. Given the results in Stanley & Doucouliagos (2013) that weighted least squares often has better small-sample properties than both random and fixed effects, we will also consider weighted least squares estimation; especially if the sample of studies turns out to be small.

Studies that have been coded with a very high risk of bias (score of 5 in any item judged on a 5-point scale) will not be included in the data synthesis. Additionally, a moderator analysis will attempt to identify the characteristics of study methods, interventions, and participants that

are associated with smaller and larger effects on the various outcomes. The following moderators will be examined:

- Gender
- Grade level of sample (or age)
- Measure of socioeconomic status (such as ethnicity, family income, family status, parents' level of education)
- Treatment modality (including e.g. components, type and/or theory of treatment and treatment duration)
- Implementation quality
- Study design

Control group and comparison group designs will be analyzed separately. If the number of included studies is sufficient (at least 10 degrees of freedom) and there is sufficient variation in the covariates, we will perform meta-regression using the mixed-model to explore how observed variables are related to heterogeneity. Weighted least squares will be considered also in the moderator analysis. We will report 95% confidence intervals for regression parameters. If we do not find a sufficient number of studies, single factor subgroup analysis will be performed. The assessment of any difference between subgroups will be based on 95% confidence intervals. To avoid problems with dependence between effect sizes we will apply robust standard errors (Hedges et al., 2010).

Sensitivity analysis will be used to evaluate whether the pooled effect sizes are robust across study designs and components of methodological quality. Furthermore, sensitivity analysis will be used to examine the strength of conclusions in relation to the quality of the data, and in relation to the use of ITT analysis in the included studies.

2.6.8 ASSESSMENT OF HETEROGENEITY

Heterogeneity among primary outcome studies will be assessed with the Chi-squared (Q), and the I-squared test, and τ-squared statistics (Higgins, Thompson, Deeks, & Altman, 2003). Any interpretation of the Chi-squared test will be made cautiously on account of its low statistical power.

2.6.9 ASSESSMENT OF REPORTING BIAS

Reporting bias refers to both publication bias and selective reporting of outcome data and results. Bias from selective reporting of outcome data and results is one of the main items in the risk of bias tool. We will use funnel plots for information about possible publication bias (Higgins & Green, 2011). However, asymmetric funnel plots are not necessarily caused by publication bias (and publication bias does not necessarily

cause asymmetry in a funnel plot). If asymmetry is present, we will consider possible reasons for this.

APPENDIX

APPENDIX A: SCREENING CRITERIA

First level screening is made on the basis of titles and abstracts. Second level screening is made on the basis of full texts. A study will be excluded in the first level screening if one or more of the answers to question 1-3 are 'No'. If the answers to question 1-3 are 'Yes' or 'Uncertain', then the full text of the study will be retrieved for second level screening. All unanswered questions need to be posed again on the basis of the full text. If not enough information is available in the full text study, the author of the study will be contacted.

FIRST LEVEL SCREENING BASED ON TITLE AND ABSTRACT:

1. Is the study about an intervention with the primary purpose of improving educational achievement and/or school readiness?
   Yes – include
   Uncertain – include
   No – stop here and exclude

Question guidance: Interventions should explicitly aim to improve educational achievement, including the take up of youth education ("ungdomsuddannelse"), school readiness, or specific academic skills. This does not mean that the intervention must consist of academic activities, but rather that the expectation must be that the intervention will primarily result in improved academic performance, and/or a higher skill level in a specific academic task. Programs that primarily aim to reduce for example criminal behavior or bullying, or to improve pro-social skills, and only have improved academic outcomes as secondary objectives, should be excluded. Note that programs that aim to improve cognitive skills may have an aim to improve educational achievement, as cognitive skills are sometimes measured by standardized tests in e.g. mathematics.

2. Are the participants in the intervention program students in a regular pre-, elementary, or middle school (i.e. preschool to grade 10)?
   Yes – include
   Uncertain – include
   No – stop here and exclude

Question guidance: A regular setting implies that studies of students attending special education schools should be excluded, but studies of students in remedial and special education classes in regular schools should be included. Studies of interventions in high school/upper secondary school, and interventions in tertiary education, such as universities, colleges, technical training institutes, community colleges, nursing schools, research laboratories, centers of excellence, and distance learning centers

should be excluded. Interventions in early childhood that are not connected to a preschool should also be excluded. Note that pre-kindergarten, daycare, and childcare can be synonyms to preschool. If the intervention targets participants who are only partly within the correct grade span, e.g. in grades 10 and 11, then the study should be included.

3. Did the intervention take place in an OECD and/or EU country?
   Yes – include
   Uncertain – include
   No – stop here and exclude

Question guidance: The OECD countries are (OECD, 2014): Australia, Austria, Belgium, Canada, Chile, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, United Kingdom, and United States.

The EU countries not in the OECD are: Bulgaria, Croatia, Cyprus, Latvia, Lithuania, Malta, and Romania.

SECOND LEVEL SCREENING BASED ON FULL TEXT:
Repeat if necessary, questions 1 – 3 based on full text. Exclude the study if the answer is 'No' to one or more of these questions; otherwise continue with questions 4 – 7 below. Exclude the study if the answer to one or more of these three questions is 'No'. Any remaining uncertainty or disagreement of eligibility will be resolved by the review authors.

4. Did the intervention take place in or after the year 2000?
   Yes – include
   Uncertain – include
   No – stop here and exclude

Question guidance: We exclude interventions performed before year 2000, i.e. all studies of interventions taking place in 1999 or before should be excluded. If an intervention is performed in the school year 1999/2000, it should be included. Studies excluded because of an early intervention should be retained for use if our resources permit an extension of the study period.

5. Is the study a primary impact study using a treatment-control design?
   Yes – include
   Uncertain – include

No – stop here and exclude

Question guidance: We include only primary research; reviews will be excluded but retained and manually searched for references to validate the search strategy. The study should also use a design that compares outcomes over one or more treatment groups to one or more control or comparison groups. A control group is defined as a non-treatment condition, which includes waitlist controls. A comparison group receives an alternative treatment. We exclude studies that compare outcomes for a single group, or a single student, pre- and post-intervention, such as case studies. Note though that studies using a case-control design include a control group, and should be included.

6. Does the study report quantitative outcomes for children and students with low socioeconomic status?
   Yes – include
   Uncertain – include
   No – stop here and exclude

Question guidance: To be included, interventions should be, at least partly, be aimed at students identified in the study under consideration on the basis of having low socioeconomic status, measured as e.g. family income, education, and occupational (including on social welfare) or minority status. This does not mean that studies including other students should by default be excluded, but if the intervention includes other students as well, results for students with low socioeconomic status must be reported separately, or the share of low SES students be reported. The study should report quantitative outcomes, we exclude qualitative studies.

7. Is the intervention implemented by individual schools or preschools, or by schools or preschools in cooperation with outside local stakeholders?
   Yes – include
   Uncertain – include
   No – stop here and exclude

Question guidance: We exclude studies that require changes to a whole school system, such as changes to the grade system, school start and leaving ages, the national/regional curriculum, the introduction of national standardized tests, and the introduction or expansion of school choice, private schools, and the expansion of universal access to preschools and other reforms aimed to increase preschool attendance. Outside local stakeholders could for example be local governments, NGOs, and researchers.

APPENDIX B: RISK OF BIAS TOOL

APPENDIX TABLE B1

Risk of bias table.

Item	Judgement[a]	Description (quote from paper, or describe key information)
1. Sequence generation		
2. Allocation concealment		
3. Confounding[b,c]		
4. Blinding?[b]		
5. Incomplete outcome data addressed?[b]		
6. Free of selective reporting?[b]		
7. Free of other bias?		
8. A priori protocol?[d]		
9. A priori analysis plan?[e]		

Ann.: Some items on low/high risk/unclear scale (double-line border), some on 5 point scale/unclear (single line border), some on yes/no/unclear scale (dashed border). For all items, record "unclear" if inadequate reporting prevents a judgement being made.

[b] For each outcome in the study.

[c] This item is only used for QESs. It is based on a list of confounders considered as important at the outset and defined in the protocol for the review (*assessment against worksheet*).

[d] Did the researchers write a protocol defining the study population, intervention and comparator, primary and other outcomes, data collection methods, etc. in advance of starting the study?

[e] Did the researchers have an analysis plan defining the primary and other outcomes, statistical methods, subgroup analyses, etc. in advance of starting the study?

RISK OF BIAS TOOL

STUDIES FOR WHICH ROB TOOL IS INTENDED

The risk of bias model is developed by Prof. Barnaby Reeves in association with the Cochrane Non-Randomised Studies Methods Group. This model, an extension of the Cochrane Collaboration's risk of bias tool, covers both risk of bias in randomised controlled trials (RCTs and QRCTs), but also risk of bias in non-randomised studies (QESs).

The point of departure for the risk of bias model is the Cochrane Handbook for Systematic Reviews of interventions (Higgins & Green, 2011). The existing Cochrane risk of bias tool needs elaboration when assessing non-randomised studies because, for non-randomised studies, particular attention should be paid to selection bias / risk of confounding. Additional items on confounding are used only for non-randomised studies (QESs) and are not used for randomised controlled trials (RCTs and QRCTs).

ASSESSMENT OF RISK OF BIAS

Issues when using modified RoB tool to assess included non-randomised studies:

- Use existing principle: score judgement and provide information (preferably direct quote) to support judgement.

- Additional items on confounding used only for non-randomised studies (QESs).
- 5-point scale for some items (distinguish "unclear" from intermediate risk of bias).
- Keep in mind the general philosophy – assessment is not about whether researchers could have done better but about risk of bias; the assessment tool must be used in a standard way irrespective of the difficulty / circumstances of investigating the research question of interest or the study design used.
- Anchors: "1/No/low risk" of bias should correspond to a high quality RCT. "5/high risk" of bias should correspond to a risk of bias that means the findings should not be considered (too risky, too much bias, more likely to mislead than inform).

1. Sequence generation
    - Low/high/unclear RoB item.
    - Always high RoB (not random) for a non-randomised study.
    - Might argue that this item is redundant for QES since it is always high – but it is important to include it in an RoB table ('level playing field' argument).

2. Allocation concealment
    - Low/high/unclear RoB item.
    - Potentially low RoB for a non-randomised study, e.g. quasi-randomised (too high RoB to sequence generation) but concealed (reviewer judges that the people making decisions about including participants didn't know how allocation was being done, e.g. odd/even date of birth/hospital number).

3. RoB from confounding (additional item for QES; assess for each outcome)
    - Assumes a pre-specified list of potential confounders defined in the protocol
    - Low(1) / 2 / 3 / 4 / high(5) / unclear RoB item
    - Judgement needs to factor in:
        - proportion of confounders (from pre-specified list) that were considered
        - whether most important confounders (from pre-specified list) were considered
        - resolution/precision with which confounders were measured
        - extent of imbalance between groups at baseline
        - care with which adjustment was done (typically a judgement about the statistical modeling carried out by authors)

- Low RoB requires that all important confounders are balanced at baseline (not primarily/not only a statistical judgement OR measured 'well' and 'carefully' controlled for in the analysis.

Assess against pre-specified worksheet. Reviewers will make an RoB judgement about each factor first and then 'eyeball' these for the judgement RoB table.

4. RoB from lack of blinding (assess for each outcome)
   - Low(1) / 2 / 3 / 4 / high(5) / unclear RoB item
   - Judgement needs to factor in:
     - nature of outcome (subjective / objective; source of information)
     - who was / was not blinded and the risk that those who were not blinded could introduce performance or detection bias see Ch.8.

5. RoB from incomplete outcome data (assess for each outcome)
   - Low(1) / 2 / 3 / 4 / high(5) / unclear RoB item
   - Judgement needs to factor in:
     - reasons for missing data
     - whether amount of missing data balanced across groups, with similar reasons
     - whether censoring is less than or equal to 25% and has been taken into account
     - see Ch.8

6. RoB from selective reporting (assess for each outcome)
   - Low(1) / 2 / 3 / 4 / high(5) /unclear RoB item
   - Judgement needs to factor in:
     - existing RoB guidance on selective outcome reporting (see Ch.8)
     - also, extent to which analyses (and potentially other choices) could have been manipulated to bias the findings reported, e.g. choice of method of model fitting, potential confounders considered / included
     - look for evidence that there was a protocol in advance of doing any. analysis / obtaining the data (difficult unless explicitly reported); QES very different from RCTs. RCTs must have a protocol in advance of starting to recruit (for REC/IRB/other regulatory approval); QES need not (especially older studies).

- hence, separate yes/no items asking reviewers whether they think the researchers had a pre-specified protocol and analysis plan.

7. RoB from other bias
   - Low(1) / 2 / 3 / 4 / high(5) /unclear RoB item
   - Judgement needs to factor in:
     - existing RoB guidance on other potential threats to validity (see Ch.8)
     - also, assess whether suitable cluster analysis is used (e.g. cluster summary statistics, robust standard errors, the use of the design effect to adjust standard errors, multi-level models and mixture models), if assignment of units to treatment is clustered.

APPENDIX TABLE B.2

Confounding Worksheet.

Assessment of how researchers dealt with confounding		
Method for *identifying* relevant confounders described by researchers:	yes	☐
	no	☐
If yes, describe the method used:		
Relevant confounders described:	yes	☐
	no	☐
List confounders described on next page		
Method used for controlling for confounding		
At design stage (e.g. matching, regression discontinuity, instrument variable):		
.....		
.....		
.....		
At analysis stage (e.g. stratification, regression, difference-indifference):		
.....		
.....		
.....		
Describe confounders controlled for below		

Confounders described by researchers.

Tick (yes[0]/no[1] judgment) if confounder considered by the researchers [Considered]. Score (1 [good precision] to 5 [poor precision]) precision with which confounder measured. Score (1 [balanced] to 5[major imbalance]) imbalance between groups. Score (1 [very careful] to 5 [not

at all careful]) care with which adjustment for confounder was carried out.

Confounder	Considered	Precision	Imbalance	Adjustment
Gender	☐	☐	☐	☐
Age	☐	☐	☐	☐
Grade level	☐	☐	☐	☐
Socioeconomic background	☐	☐	☐	☐
Performance at baseline	☐	☐	☐	☐
Unobservables[9]	☐	Irrelevant	☐	☐
Other:	☐	☐	☐	☐

USER GUIDE FOR UNOBSERVABLES

Selection bias is understood as systematic baseline differences between groups and can therefore compromise comparability between groups. Baseline differences can be observable (e.g. age and gender) and unobservable (to the researcher; e.g. 'appearance'). There is no single non-randomized study design that always solves the selection problem. Different designs solve the selection problem under different assumptions and require different types of data. There can be particularly great variations in how different designs deal with selection on unobservables. The "right" method depends on the model generating participation, i.e. assumptions about the nature of the process by which participants are selected into a program.

As there is no universally correct way to construct counterfactuals, we will assess the extent to which the identifying assumptions (the assumption that makes it possible to identify the counterfactual) are explained and discussed (preferably by the authors in an effort to justify their choice of method). We will look for evidence of authors using the following examples (this is NOT an exhaustive list):

NATURAL EXPERIMENTS

Discuss whether they face a truly random allocation of participants and that there is no change of behavior in anticipation of, e.g. policy rules.

INSTRUMENT VARIABLE (IV)

Explain and discuss the assumption that the instrument variable does not affect outcomes other than through their effect on participation.

9. See User guide for unobservables.

MATCHING (INCLUDING PROPENSITY SCORES)

Explain and discuss the assumption that there is no selection on unobservables, only selection on observables.

(MULTIVARIATE, MULTIPLE) REGRESSION

Explain and discuss the assumption that there is no selection on unobservables, only selection on observables. Further discuss the extent to which they compare comparable people.

REGRESSION DISCONTINUITY (RD)

Explain and discuss the assumption that there is a (strict!) RD treatment rule. It must not be changeable by the agent in an effort to obtain or avoid treatment. Continuity in the expected impact at the discontinuity point is required.

DIFFERENCE-IN-DIFFERENCE (TREATMENT-CONTROL-BEFORE-AFTER)

Explain and discuss the assumption that outcomes of participants and nonparticipants evolve over time in the same way.

REFERENCES

Alexander, K.L., D.R. Entwisle & L.S. Olson (2001): "Schools, Achievement, and Inequality: A Seasonal Perspective." *Education Evaluation and Policy Analysis* 23(2), p. 171-191.

Angrist, J.D. J-S. & Pischke (2009): *Mostly Harmless Econometrics – An Empiricist's Companion*, Princeton: Princeton University Press.

Baker, M., J. Gruber & K. Milligan (2008): "Universal Child Care, Maternal Labor Supply and Family Well-being." *Journal of Political Economy*, 116(4), p. 709-745.

Berlinski, S., S. Galiani & M. Managorda (2008): "Giving Children a Better Start: Pre-school Attendance and School-age Profiles." *Journal of Public Economics*, 92(5-6), p. 1416-1440.

Berlinski, S., S. Galiani & P. Gertler (2009): "The Effect of Pre-primary Education on Primary School Performance." *Journal of Public Economics*, 93(1), p. 219-234.

Blau, D. & J. Currie (2006): "Who is Minding the Kids?" In: F. Welch & E. Hanushek (eds) *Handbook of Education Economics*. New York: North Holland.

Borenstein, M., L.V. Hedges, J.P.T. Higgins & H.R. Rothstein (2009): *Introduction to Meta-Analysis*. Chichester: John Wiley & Sons Ltd.

Bradley, R.H. & R.F. Corwyn (2002): "Socioeconomic Status and Child Development." *Annual Review of Psychology*, 53, p. 371-399.

Camilli, G., S. Vargas, S. Ryan & W.S. Barnett (2010): "Meta-analysis of the Effects of Early Education Interventions on Cognitive and Social Development." *Teachers College. Record,* 112(3), p. 579-620.

Chambers, B., A. Cheung, R.E. Slavin, D. Smith & M. Laurenzano (2010): "Effective Early Childhood Education Programmes: A

Systematic Review." *Best Evidence Encyclopedia*, www.bestevidence.org.uk.

Chetty, R., J.N. Friedman, N. Hilger, E. Saez, D. Whitmore Schanzenback & D. Yagan (2011): "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *Quarterly Journal of Economics*, 126, p. 1593-1660.

Cook, P.J., K. Dodge, G. Farkas, R.J. Fryer, J. Guryan, J. Ludwig, S. Mayer, H. Pollack & L. Steinberg (2014): *The (Surprising) Efficacy of Academic and Behavioral Intervention with Disadvantaged Youth: Results From a Randomized Experiment in Chicago*. NBER Working Paper, no. 19862.

Currie, J. (2009): "Healthy, Wealthy, and Wise: Socioeconomic Status, Poor Health in Childhood and Human Capital Development." *Journal of Economic Literature*, 47(1), p. 87-122.

Donner, A., G. Piaggio & J. Villar (2001): "Statistical Methods for the Meta-analysis of Cluster Randomized Trials." *Statistical Methods in Medical Research*, 10(5), p. 325-338.

Duncan, G. & K. Magnuson (2013): "Investing in Preschool Programs." *Journal of Economic Perspectives*, vol. 27, no 2, p. 109-132.

Edmonds, M.S., S. Vaughn, J. Wexler, C. Reutebuch, A. Cable, K. Klingler Tackett & J. Wick Schnakenberg (2009): "A Synthesis of Reading Interventions and Effects on Reading Comprehension Outcomes for Older Struggling Readers." *Review of Educational Research*, 79(1), p. 262-300.

Esping-Andersson, G., I. Garfinkel, W-J.Han, K. Magnuson, S. Wagner & J. Waldfogel (2012): "Child Care and School Performance in Denmark and the United States." *Children and Youth Services Review*, 34, p. 576-589.

Felfe, C., N. Nollenberger & N. Rodriguez-Planas (2012): *Can't Buy Mommy's Love? Universal Childcare and Children's Long-term Cognitive Development*. IZA. Discussion Paper no. 7053.

Flynn, L.J., X. Zheng & H.L. Swanson (2012): "Instructing Struggling Older Readers: A Selective Meta-analysis of Intervention Research." *Learning Disabilities Research & Practice*, 27(1), p. 21-32.

Fraga, M.F., E. Ballestar, M.F. Paz, S. Ropero, F. Setien, M.L. Ballestar, D. Heine-Suner, J.C. Cigudosa, M. Urioste, J. Benitez, M. Boix-Chornet, A. Sanchez-Aguilera, C. Ling, E. Carlsson, P. Poulsen, A. Vaag, Z. Stephan, T.D. Spector, Y-Z. Wu, C. Plass & M. Esteller (2005): "Epigenetic Differences Arise During the Lifetime of Monozygotic Twins." *Proceedings of the National Academies of Sciences*, 102(30), p. 10604-10609.

Fredriksson, P., E-A. Johansson, C. Hall & P. Johansson (2010): "Do Pre-school Interventions Further the Integration of Immigrants?"

In: E-A. Johansson *Essays on Schooling, Gender, and Parental Leave.* IFAU Dissertation series 2010:1.

Fredriksson, P., B. Öckert & H. Oosterbeek (2013): "Long-term Effects of Class Size." *Quarterly Journal of Economics*, 128(1), p. 249-285.

Fryer, RG. Jr. & S.D. Levitt (2013): "Testing for Racial Differences in the Mental Ability of Young Children. *American Economic Review*, 103(2), p. 981-1005.

Gershenson, S. (2013): "Do Summer Time-use Gaps Vary by Socioeconomic Status?" *American Educational Research Journal*, 50(6), p. 1219-1248.

Hackman, D.A. & M.J. Farah (2009): "Socioeconomic Status and the Developing Brain." *Trends in Cognitive Science* 13(2), p. 65-73.

Havnes, T. & M. Mogstad (2011): "No Child Left Behind: Subsidized Child Care and Children's Long-run Outcomes." *American Economic Journal: Economic Policy,* 3(2), p. 97-129.

Havnes, T. & M. Mogstad (2014): "Is Universal Child Care Leveling the Playing Field?" *Forthcoming in Journal of Public Economics.*

Heckman, J.J. (2006): "Skill Formation and the Economics of Investing in Disadvantaged Children." *Science*, 312, p. 1900-1902.

Hedges, L.V., E. Tipton & M.C. Johnson (2010): "Robust Variance Estimation in Meta-regression with Dependent Effect Size Estimates." *Research Synthesis Methods*, 1, p. 39-65.

Higgins, J.P.T. & S. Green (eds.) (2011): *Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0* [updated March 2011]. Wiley-Blackwell. The Cochrane Collaboration. Available from www.cochrane-handbook.org.

Higgins, J.P., S.G. Thompson, J.J. Deeks & D.G. Altman(2003): "Measuring Inconsistency in Meta-analyses." *British Medical Journal*, 327 (7414), p. 557-60.

Jacob, B. & J. Ludwig (2008): *Improving Outcomes for Poor Children.* NBER working paper, no. 14550.

Kim, J.S. & D.M. Quinn (2013): "The Effects of Summer Reading on Low-income Children's Literacy Achievement from Kindergarten to grade 8: A Meta-analysis of Classroom and Mome Interventions." *Review of Educational Research*, 83(3), p. 386–431.

Lipsey, M.W. & D.B. Wilson (2001): "Practical Meta-analysis." *Applied Social Research Methods Series*, v. 49.

Magnuson, K. & H. Shager (2010): "Early Education: Progress and Promises for Children from Low-income Families." *Children and Youth Services Review*, 32, p. 1186-1198.

OECD (2010). *PISA 2009 Results: Overcoming Social Background – Equity in Learning Opportunities and Outcomes (Volume II).* Retrieved from http://dx.doi.org/10.1787/9789264091504-en.

Reynolds, A.J., K.A. Magnuson & S-R. Ou (2010): "Preschool-to-third Grade Programs and Practices: A Review of Research." *Child and Youth Services Review,* 32(8), p.1121-1131.

Reynolds, A.J. & J.A. Temple (2008): "Cost-effective Early Childhood Development Programs from Preschool to Third Grade." *Annual Review of Clinical Psychology*, 4, p.109-139.

Rutter, M. (2006): *Genes and Behavior: Nature-nurture Interplay Explained.* Malden: Blackwell Publishing.

Sánchez-Meca, J., F. Marín-Martínez & S. Chacón-Moscoso (2003): "Effect-size Indices for Dichotomized Outcomes in Meta-analysis." *Psychological Methods*, 8(4), p. 448-467.

Scammaca, N.K., G. Roberts, S. Vaughn & K.K. Stuebing (2013): "A Meta-analysis of Interventions for Struggling Readers in Grades 4-12: 1980-2011." *Journal of Learning Disabilities*.

Sirin, S.R. (2005): "Socioeconomic Status and Academic Achievement: A Meta-analytic Review of Research." *Review of Educational Research*, 75(3), p. 417–453.

Slavin, R.E. & C. Lake (2008): "Effective Programs in Elementary Mathematics: A Best-evidence Synthesis." *Review of Educational Research*, 78(3), p. 427-515.

Slavin, R.E., C. Lake & C. Groff (2009): "Effective programs in Middle and High School Mathematics: A Best-evidence Synthesis." *Review of Educational Research*, 79(2), p. 839-911.

Slavin, R.E., C. Lake, B. Chambers, A. Cheung & S. Davis (2009): "Effective Reading Programs for the Elementary Grades: A Best-evidence Synthesis." *Review of Educational Research*, 79(4), p. 1391-1466.

Slavin, R., C. Lake, S. Davis & N. Madden (2011): "Effective Programs for Struggling Readers: A Best-evidence Synthesis." *Educational Research Review*, 6, p. 1–26.

Stanley, T.D. & H. Doucouliagos (2013): *Neither Fixed nor Random: Weighted Least Squares Meta-analysis.* Working paper SWP 2013/1, Deakin University.

Tucker-Drob, E.M., M. Rhemtulla, K.P. Harden, E. Turkheimer & D. Fask (2011): "Emergence of a Gene X Socioeconomic Status Interaction on Infant Mental Ability Between 10 months and 2 Years." *Psychological Science*, 22(1), p. 125-133.

UNESCO (1994). *The Salamanca Statement and Framework for Action on Special Needs Education.* Salamanca, Spain.

Wanzek, J., S. Vaughn, J. Wexler, E.A. Swanson, M. Edmonds & A-H. Kim (2006): "A Synthesis of Spelling and Reading Interventions and their Effects on the Spelling Outcomes of Students with LD." *Journal of Learning Disabilities*, 39(2), p. 528-543.

Wilson, S., E.E. Tanner-Smith, M.W. Lipsey, K. Steinka-Fry & J. Morrison (2011): "Dropout Prevention and Intervention programs: Effects on School Completion and Dropout among School-aged Children and Youth." *Campbell Systematic Reviews*. retrieved from http://www.campbellcollaboration.org/lib/project/158/.

Yoshikawa, H., C. Weiland, J. Brooks-Gunn, M.R. Burchinal, L.M. Espinosa, W.T. Gormley, J. Ludwig, K.A. Magnuson, D. Phillips & M.J. Zaslow (2013): *Investing in our Future: The Evidence Base on Preschool Education*. Foundation for Child Development & Society for Research in Child Development, http://fcd-us.org/resources/evidence-base-preschool.

Zief, S.G., S. Lauver & R.A. Maynard (2006): "Impacts of After-school Programs on Student Outcomes: A Systematic Review." *Campbell Systematic Reviews*. retrieved from http://campbellcollaboration.org/lib/project/12/