THE CAMPBELL COLLABORATION

Protocol for a Systematic Review: Targeted School-Based Interventions for Improving Reading and Mathematics for Students With or At-Risk of Academic Difficulties in Grade 7 to 12: A Systematic Review

Jens Dietrichson, Martin Bøg, Trine Filges, Anne-Marie Klint Jørgensen

Submitted to the Coordinating Group of:

☐	Crime and Justice
☒	Education
	☐ Disability
☐	International Development
	☐ Nutrition
☐	Social Welfare
☐	Other:

Plans to co-register:

☒	No		
☐	Yes	☐ Cochrane	☐ Other
☐	Maybe		

Date Submitted: 2014-08-15
Date Revision Submitted: 2015-11-24
Approval Date:
Publication Date: 2016-03-01

BACKGROUND

The problem

In member countries of the Organisation for Economic Co-operation and Development (OECD) countries, almost one in five of all youth between 25-34 years of age have not earned the equivalent of a high-school degree (or upper secondary education). Moreover, on average, 16% of 15-29 year-olds are neither employed, nor in education or training; this proportion increased substantially in 2009 and 2010 compared with pre-crisis levels (i.e., before 2008) (OECD, 2012). Entering adulthood with a low level of education is associated with reduced employment prospects as well as limited possibilities for financial progression in adult life (De Ridder, Pape, Johnsen, Westin, Holmen, & Bjørngaard, 2012; Johnson, Brett, & Deary, 2010; Scott & Bernhardt, 2000). Furthermore, low levels of education are also negatively correlated with numerous health related issues and risk behaviours, such as drug use and crime, which has serious implications for the individual as well as for society (Berridge, Brodie, Pitts, Porteous, & Tarling, 2001; Brook, Stimmel, Zhang, & Brook, 2008; Feinstein, Sabates, Anderson, Sorhaindo, & Hammond, 2006; Horwood et al., 2010; Sabates, Feinstein, & Shingal, 2013).

In many contexts, socioeconomic status (SES) is a major predictor of educational achievement (e.g. Björklund & Salvanes, 2011; Currie, 2009; Kim & Quinn, 2013; Sirin, 2005; White, 1982). For example, the results from Programme for International Student Achievement (PISA) point to the fact students from families with low SES tend to score much lower (OECD, 2010, 2013). Across OECD countries, students from a high SES backgrounds outperform students from an average background by about one year's worth of education in reading and mathematics, and outperform students with low SES by even more. While social disadvantage is strongly associated with lower school performance, results from PISA also show that some students with low SES excel in PISA, demonstrating that overcoming socio-economic barriers to academic achievement is indeed possible (OECD, 2010).

There is, for these reasons, a significant interest in information about effective interventions to increase academic achievement and enhance educational prospects for educationally disadvantaged youth. Interventions aimed at improving educational achievement described in the research literature are numerous and very diverse in terms of intervention focus, target group, and delivery mode. The review we plan to conduct will focus on targeted interventions performed in schools and provided to students with and at-risk of academic difficulties in grades 7-12 (ages range from 12-14 to 17-19, depending on country/state), where academic skill building and learning are the primary intervention aims. The outcome variables will be standardised tests of achievement in reading and mathematics. This relatively broad selection will identify a range of interventions, and will allow us to examine intervention effectiveness across settings and methods.

The intervention

We will make a broad review of interventions that aim to improve the academic achievement of students with or at-risk of academic difficulties in grades 7-12, performed in schools during the regular school year. We therefore expect to include a range of interventions, including literacy and mathematical interventions, tutoring and mentor programmes, and cognitive training and alternative teaching strategies interventions. Interventions may therefore include components that change the method of instruction – such as tutoring and cooperative learning interventions – or change the content of the instruction – as interventions emphasizing mathematical problem solving skills or vocabulary. Many interventions may change both method and content, and include several major components.

Our restriction to interventions that explicitly aim to improve the academic performance of students means that we will exclude interventions that may improve academic learning as a side-effect. Examples are interventions where behavioural or socioemotional problems are the primary intervention aim, like Classroom Management or the SCARE Program. However, interventions with behavioural and socio-emotional components may very well have academic achievement as one of their primary aims, and use standardized tests of reading and mathematics as one of their primary outcomes (e.g. some mindset and stereotype threat interventions). Such interventions will be included.

The intervention should be school-based, by which we mean performed in school, during the regular school year, and where schools are one of the stakeholders. This restriction excludes for example some after-school programmes, and summer camps and summer reading programmes. Such programmes appear to be qualitatively different from interventions performed in school (Dietrichson et al., 2015b).

Interventions should furthermore be targeted (or selected/indicated). That is, interventions should target certain students and/or student groups identified as having academic difficulties or being at-risk of such difficulties. This group includes for example youth with learning disabilities, students from families with low educational background, with a diverse ethnic/cultural background, or students with a low grade point average. Many targeted interventions are supplemental programmes delivered individually and are complementary to regular classes and school activities, such as the Reading Apprenticeship programme or individual computer-based training (e.g., CogMed). However, targeted interventions can be delivered in various settings, including in class (e.g., paired reading interventions or the Xtreme Reading programme), or in group sessions (e.g., the READ 180 programme), or individually.

Universal interventions applied to improve the quality of the common learning environment at school in order to raise academic performance of all students (including average and above average students) will be excluded. Whole-school reform strategy concepts such as Success for All, curriculum-based programmes like Elements of Mathematics (EMP), as well

as reduced class size interventions and general professional development interventions for principals and teachers will therefore be excluded.

How the Intervention Might Work

While all the included interventions strive to improve academic achievement for students with or at-risk of academic difficulties, they may do so with different approaches and with diverse strategies of how to create that improvement. This diversity reflects the varying reasons for why students are struggling or are at-risk. In turn, the theoretical background for the interventions varies accordingly. It is therefore not possible to specify one particular theory of change or one theoretical framework for this review. Instead, we first discuss possible reasons for the stratification in educational performance, and second, briefly review three theoretical perspectives that we believe are likely to be characteristic for the majority of the included interventions. Lastly, we discuss and exemplify how existing targeted interventions may address some of the reasons for academic difficulties, and how they fit into the theoretical perspectives.

Reasons for academic difficulties

Students may be struggling for a number of reasons. However, the latest PISA-tests suggest that students from families with low SES are overrepresented among low performing pupils (OECD 2010, 2013). The reasons for low achievement in general are thus likely connected to the challenges faced by low SES students, a group for which there is a relatively large research literature from different academic fields examining the reasons for why their educational achievement is lower. We discuss this literature below.[1]

Lower innate ability does not seem to be a major explanation of achievement differences. Recent evidence from the US indicates that measures of mental ability do not differ significantly 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 about one third of a standard deviation higher. Genes accounted for nearly 50 percent of the variation in mental ability of children raised in high SES homes, but only a negligible share of the variation in mental ability of children raised in low SES homes. Similar results were obtained in a follow-up measurement using tests of school readiness (Tucker-Drob et al., 2013). Similarly, the differences in test scores between black and white American children have been found to be about one standard deviation already at age 3.

[1] As we discuss in the section *The contribution of this review*, we will exclude interventions targeting students with physical learning disabilities (e.g. blind students), students with dyslexia/ dyscalculia, and interventions that are specifically directed towards students with a certain neuropsychiatric disorder (e.g. ADHD). The reasons for why these types of students are struggling seem different from the reasons discussed in this section, and interventions targeting these groups are probably also different from those targeting students with or at-risk of academic difficulties.

However, examining infants 8 to 12 months old, Fryer and Levitt (2013) found no significant differences between Hispanics, Asians, Blacks, and Whites. Furthermore, early childhood poverty has been shown to be 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). While hereditary factors cannot be completely ruled out as a determinant of differing educational achievement with current knowledge, these results suggest that the environment is the constraining factor for the achievement of low SES children (Burchinal et al., 2011; Nisbett et al., 2012).

Consistent with the differences between high and low SES children being present early on, the early childhood environment seems 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. Child health problems in turn influence both educational and labour market outcomes, but seem to be smaller for educational outcomes than for earnings (Currie, 2009).

Family resources and the home environment of low SES students also seem less conducive to high educational achievement (Jacob & Ludwig 2008). High SES families on average provide a richer language and literacy environment (Hart & Risley, 2003), use different parenting practices, and spend more money on early childhood education (Esping-Andersson et al., 2012). Low SES parents also seem to have lower academic expectations for their children (Bradley & Corwyn, 2002; Slates et al, 2012), and teachers have lower expectations for low SES students (e.g. Good et al., 2003; Timperley & Phillips, 2003).

The neighbourhoods students grow up in are another potential determinant of achievement. Regarding the relative importance of families and neighbourhoods, the review in Björklund & Salvanes (2011) indicate that family resources are the more important explanatory factor. Results from experiments where families are randomly given the opportunity to change neighbourhoods show mixed results (e.g. Chetty et al., 2015; Kling et al., 2007). But it seems likely that low SES students live in neighbourhoods that are less supportive of high educational achievement in terms of, for example, peer support and role models. To get by in a disadvantaged neighbourhood may also require a very different set of skills compared to what is needed to thrive in school, something which may increase the risk that pupils have trouble decoding the “correct” behaviour in educational environments (Heller, Shah, Guryan, Ludwig, Mullainathan & Pollack, 2015).

In sum, the evidence indicates that large and significant differences are present already well before children start school. Heckman (2006) furthermore argues 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. School interventions do however have the potential to significantly reduce the gap between high and low SES students (Björklund & Salvanes, 2011).

Theoretical perspectives

The reasons why students may be struggling laid out in the previous section are multifaceted, and the theoretical perspectives underlying interventions are therefore likely to be broad. Nevertheless, we anticipate that three superordinate components will be characteristic for the majority of the included interventions. These components can be abridged to:

- Adaptation of behaviour (social learning theory).
- Individual cognitive learning (cognitive developmental theory).
- Alteration of the social learning environment (pedagogical theory).

We emphasise that the following presentation of theoretical perspectives is not all-covering, and, though components are presented as demarcated, they contain some conceptual overlap.

Social learning theory has its origins in social and personality psychology, and was initially developed by psychologist Julian Rotter and further developed especially by Albert Bandura (1977; 1986). From the perspective of social learning theory, behaviour and skills are primarily learned by observing and imitating the actions of others, and behaviour is in turn regulated by the recognition of those actions by others (reinforcement), or discouraged by lack of recognition or sanctions (punishment). According to social learning theory, creating the right social context for the student can therefore stimulate more productive behaviour through social modelling and reinforcement of certain behaviours that can lead to higher educational achievement.

Cognitive developmental theory is not one particular theory, but rather a myriad of theories about human development that focus on how cognitive functions such as language skills, comprehension, memory and problem-solving skills enable students to think, act and learn in their social environment. Some theories emphasize a concept of intelligence where children gradually come to acquire, construct, and use cognitive functions as the child naturally matures with age (e.g. Piaget, 2001; Perry, 1999). Other theories hold a more socio-cultural view of cognitive development and use a more culturally distinct and individualized concept of intelligence that to a greater extent includes social interaction and individual experience as the basis for cognitive development. Examples include the theories of Robert Sternberg (2009) and Howard Gardner (1999).

Pedagogical theory draws on the different disciplines in psychology and social theory such as cognitivism, social-interactional theory and socio-cultural theory of learning and development. There is not one uniform pedagogical model, but examples of contemporary models in mainstream pedagogy are concepts such as Scaffolding (Bruner, 2006) and the Zone of Proximal Development (Vygotsky, 1978), which origins in developmental and educational psychology. These notions hold that learning and development emerge through practical activity and interaction. Acquisition of new knowledge is therefore considered to be dependent on social experience and previous learning, as well as the availability and type of

instruction. Accordingly, school interventions require educators to interact and organise the learning environment for the student in certain ways to fit the individual student's needs and potentials for development.

Interventions in practice

In general, school interventions affect academic achievement by changing the methods by which instruction is given (instructional methods), or by changing the content of what is taught (the content domain), and many combine several intervention components as well as theoretical perspectives. Previous reviews (e.g. Dietrichson et al. 2015a) indicate that we will for example find interventions using the following categories of instructional methods: tutoring, coaching/mentoring, cooperative learning/peer-assisted learning, computer-assisted instruction, feedback and progress monitoring, behavioural/psychological interventions, and incentive programs. Reading interventions directed to older students often target content domains such comprehension, fluency, word study, and vocabulary (e.g. Scammaca et al. 2015). Slavin et al. (2009) compared curricula for middle and high school mathematics that differed over for example how much they emphasised domains such as problem solving and conceptual understanding. Gersten et al. (2009) used the following domains to divide mathematics interventions into categories: (a) operations (e.g. addition, subtraction, and multiplication), (b) word problems, (c) fractions, (d) algebra, and (e) general math proficiency (or multiple components).

Earlier research has shown that very different types of academic interventions can improve academic performance, both across methods, delivery mode, age group and duration (e.g. Cheung & Slavin, 2012; Dietrichson et al. 2015a). Both reading strategy instruction and peer-mediated learning programmes such as paired reading have been shown to be effective in improving literacy skills of struggling secondary school readers. These are two types of programmes that clearly have different components and delivery modes (Edmonds, Vaughn, Wexler, Reutebuch, Cable, Klingler Tackett & Wick Schnakenberg, 2009). In another example, Good, Aronson, and Inzlicht (2003) show that changing expectations of seventh grade students at risk for stereotype-based underperformance (minority and low-income students in general, and girls regarding mathematics) can improve standardised test scores.

On the other hand, while some interventions which rely on a specific approach may prove effective, other interventions relying on a similar approach may not. As an example, computer-assisted instruction programmes range from strong effects to no effects at all on mathematical achievement (Kulik, 2003; Chambers, 2003; Cheung & Slavin, 2013), and while computer-based instruction programmes overall show some effect on math skills, they seem to have smaller impact on reading skills (Kulik, 2003; Cheung & Slavin, 2012).

There are indications that one-on-one tutoring and small group tutoring have some of the largest effects on academic outcomes across conditions in both reading and mathematics. However, the evidence base varies across interventions, and in general there have been more studies examining reading interventions than math interventions (e.g. Cohen, Kulik, & Kulik,

1982, Dietrichson et al., 2015a; Flynn, Marquis, Paquet, Peeke, & Aubry, 2012; Forsman & Vinnerljung, 2012; Reisner, Petry, & Armitage, 1989, 1990; Robinson et al., 2005). Furthermore, recent research also demonstrates that peer-mediated interventions such as collaborative learning interventions and peer-tutoring in general have promising effects for disadvantaged and low performing secondary school students (McMaster & Fuchs, 2002; Bowman-Perrott et al., 2013).

This outlined research points to direct and individual instruction, small or one-on-one defined settings, and mediation from adults or more competent peers as being especially important for struggling learners. Furthermore, interventions such as tutoring and structured peer-mediated interventions often have in common that they comprise an eclectic theoretical model that combines components from all three perspectives on learning presented in the previous section. They are comprehensive interventions that relies on a complex of mechanisms such as increased feedback and tailor-made instruction (pedagogical theory), regulation of behaviour by for example rewards or interaction with role models (social learning theory), and development of cognitive functions such as learning how to learn (cognitive developmental theory).

Another way of viewing these and other types of interventions is that they address the differential family and neighbourhood resources of high and low SES students described in the previous section. Students from high SES families are likely to have access to "tutors" all year round, as parents, siblings and other family members help out with homework and schoolwork. Interventions to change mindsets, increase expectations, and mitigate stereotype threat also substitute for high SES families and teachers already having such expectations or teaching their children such a mindset. Different types of extrinsic rewards may be a way to bolster motivation, which may be especially important for students whose families place less weight on educational achievement.

Furthermore, if, as indicated in the previous section, the differences between high and low SES students and students with academic difficulties can be understood as a consequence of differential access to 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. To exemplify, both programmes deemed to be backed by strong evidence of effectiveness in improving middle and high school mathematics in Slavin et al. (2009) include several components. The first, Student Teams-Achievement Divisions, includes learning in small teams, individual assessments and accountability, as well as rewards based on team performance. The second, IMPROVE, combines cooperative learning, metacognitive instruction, and mastery learning. A further example highlights that it does not have to be just academic problems that affect school achievement. Two recent studies examine the programme Becoming A Man, which includes features from cognitive behavioural therapy and the development of social-cognitive skills such as generating new solutions to problems, learning new ways of behaving, considering another's perspective, thinking ahead, and evaluating consequences ahead of time. The program significantly

reduced instances of violent-crime and decreased dropout rates, but did not increase test scores in a randomised field experiment including 2,740 male youth in grades 7-10 from high-crime and high-poverty Chicago neighbourhoods (Heller, Pollack, Ander & Ludwig, 2013). However, combined with a math tutoring intervention, the programme also significantly increased standardized test scores in an experiment with a population of 106 males from similar neighbourhoods (Cook, Dodge, Farkas, Fryer, Guryan, Ludwig, Mayer, Pollack & Steinberg, 2014).

Another reason why it is interesting to examine combinations of components relates to an often suggested explanation for missing impacts: lack of motivation among participants (e.g. Fuchs, Fuchs & Kazdan, 1999; Edmonds et al. 2009). It is therefore possible that interventions will be more effective if they also include some form of rewards for participating students and implementing teachers, along with other components providing for instance specific pedagogical support. At the same time, just providing motivation or incentives may not be enough. For example, in a large scale randomised experiment (in total the experiment involved around 27,000 students) second graders were paid to read books, fourth and seventh grade students were paid for performance on a series of assessments, and ninth graders were paid for grades. None of these treatments yielded significant effects on the aggregate treatment level (Fryer, 2011).

For struggling students in grades 7-12, who are likely to have a history of low achievement, finding the right combination of intervention components may be especially pertinent (e.g. Fuchs et al. 1999; Edmonds et al. 2009). Some researchers have recommended, based on the perceived low relative cost-effectiveness of interventions directed to adolescents, that resources should disproportionally be used for early interventions (e.g. Esping-Andersen, 2004, Heckman, 2006), or that secondary schools should primarily be providing technical and vocational training for disadvantaged teenagers (Cullen, Levitt, Robertson, & Sadoff, 2013). However, Cook et al. (2014) argued that the low relative cost-effectiveness may be a premature conjecture, as previous interventions for youths have often not combined the fostering of academic skills with other important factors for academic success, such as social-cognitive (or non-cognitive) skills. As for example social information processing programmes (Wilson & Lipsey, 2006a; 2006b), and programmes based on cognitive behavioural therapy (e.g. Lipsey, Landenberger & Wilson, 2007) have been found to effectively reduce problematic behaviour and promote social-cognitive skills, combinations with more academically oriented interventions look promising.

Why it is Important to do the Review

In this section we first discuss earlier related reviews, and then the contributions of this review in relation to the earlier literature.

Prior reviews

In some regards, this review shares common ground with existing Campbell reviews and reviews in progress such as "Impacts of After-School Programs on Student Outcomes: A Systematic Review" (Zief, Lauver, & Maynard, 2006), "Dropout Prevention and Intervention Programs: Effects on School Completion and Dropout among School-aged Children and Youth" (Wilson, Tanner-Smith, Lipsey, Steinka-Fry, & Morrison, 2011), and "Effects of College Access Programs on College Readiness and Enrollment" (Harvill, Maynard, Nguyen, Robertson-Kraft, Tognatta, & Fester, 2012).[2]

Nevertheless, this review differs in substantial ways from these existing Campbell reviews. First, with the exception of Zief et al. (2006), the listed reviews do not explicitly target an educationally disadvantaged or low performing student population. Zief et al. (2006) on the other hand excluded interventions performed outside North America, and three of the five studies included were of programmes primarily designed to reduce negative behaviours such as delinquency and drug use; i.e. the programmes did not target academic achievement as their primary outcome. Wilson et al. (2011) did not explicitly target students with or at-risk of academic difficulties, many of the studies in their review of dropout prevention and interventions programmes of course included at-risk groups. Except their review, existing Campbell reviews all focus on one specific type of intervention or setting. A major difference between their review and the current proposal is that they focused on programmes of school completion and dropout prevention, and outcome measures as dropout and graduation rates. This review will only include studies that report results on standardised tests in reading and mathematics. There is some overlap between the types of interventions included but also clear differences, as many of the interventions we will include do not target dropout and interventions such as for example paid employment for students, community service programs, and vocational training will not feature in our review.

In addition to these Campbell reviews and reviews in progress, there are other related reviews with a similar broad scope and a target group overlapping ours to some degree.[3] Slavin et al. (2009) reviewed programmes in middle and high school mathematics, whereas Slavin, Cheung, Groff, & Lake (2008) reviewed reading programmes for middle and high schools. However, these reviews focused on all kinds of programmes, not programmes for

---

[2] Thematically, and to some extent in the age groups included, the Campbell review of volunteer tutoring programmes in grades K-8 by Ritter, Albin, Barnett, Blankenship, & Denny (2006) also overlaps with this review. However, their review contains only two studies, from the same dissertation, of students in the same age as our target group (in grade 7), none of which targets low achieving or at-risk students.

[3] The following reviews are also related, but focus on more general populations and/or have a more narrow scope (topic and target population in parentheses): McMaster & Fuchs (2002, cooperative learning for students with learning disabilities), Alfieri et al. (2011, discovery-based instruction for general student populations), Dexter & Hughes (2011, graphic organizers for students with learning disabilities), Cheung & Slavin (2012, technology applications for general student populations), Kyndt et al. (2013, cooperative learning for general student populations), de Boer et al. (2014, attributes of interventions for general student populations), and Reljic et al. (2015, bilingual programs to European students). We will use these reviews to snowball references.

at-risk or low-performing students. Furthermore, Wanzek, Vaughn, Wexler, Swanson, Edmonds & Kim (2006) reviewed reading interventions directed to students in grades K-12 with learning disabilities, and Edmonds et al. (2009), Flynn, Zheng & Swanson (2012), and Scammaca et al. (2015) reviewed interventions for struggling readers in grades 6-12, 5-9, and 4-12, respectively.[4] These reviews thus covered low achieving students, but neither at-risk students nor areas other than reading. Gersten et al. (2009) examined four types of components of mathematics instruction for students with learning disabilities, but did not include studies for students at-risk (or more general reasons for low performance than learning disabilities). Dietrichson et al. (2015a) on the other hand included studies in both reading and mathematics and based inclusion on the share of students with low SES, but did not consider whether students had academic difficulties or not.

In terms of findings related to this review's primary outcome measures, the reviews that have focused on the effects of academic interventions on reading test scores all showed positive overall effect sizes, although there was a rather large variation between interventions in all reviews (Edmonds et al., 2009; Flynn et al., 2012; Scammaca et al., 2015; Slavin et al., 2008; Slavin et al., 2009; Wanzek et al., 2006). The four reviews of reading interventions directed to struggling readers reported positive effects in general but few reliable differences over types of interventions (Edmonds et al., 2009; Flynn et al., 2012; Scammaca et al., 2015; Wanzek et al., 2006). An exception is that reading comprehension interventions were associated with significantly higher effect sizes than fluency interventions in Scammaca et al. (2015), but this difference disappears when only standardised measures were considered.

Gersten et al. (2009) examined four components of mathematics instruction for students with learning disabilities, and found most support for approaches to instruction (e.g. explicit instruction, use of heuristics) and/or curriculum design, and providing formative assessment data and feedback to teachers. Dietrichson et al. (2015a) examined interventions that have used standardised tests in reading and mathematics and categorise 14 intervention components mainly delimited by the instructional methods used. Tutoring, feedback and progress monitoring, and cooperative learning have the largest and most robust average effect sizes.

The best evidence syntheses by Slavin et al. (2008) and Slavin et al. (2009) both point to instructional-process programmes, especially programmes that incorporate cooperative learning, as having larger effects than curricula based interventions, and computer assisted instruction programmes. Slavin et al. (2009) found no indications that effect sizes differ between socioeconomically disadvantaged students and non-disadvantaged students. However, only a relatively small subset of studies reported results differentiated by SES, and

[4] Wanzek et al. (2006) and Flynn et al. (2012) contain only a few studies of interventions directed to students in our target group students though. Note also that all studies in Wanzek, Vaughn, Scammacca, Metz, Murray, Roberts, & Danielson (2013), a review of extensive interventions for struggling readers covering grades 3-12, are included in Scammaca et al. (2015).

the review does not contain information about whether the programmes that in general show the largest effect sizes also has the largest effect sizes for disadvantaged students.

Slavin et al. (2009) and Edmonds et al. (2009) reported that some programmes, which have been shown to be effective for younger students, may have smaller or no effects for older students. Effect sizes were smaller for older students also in Scammaca et al. (2015), although not significantly different. As discussed in the previous section, there are also other indications that earlier interventions are more cost-effective, but, as argued in Cook et al. (2014), this may be because programmes directed to older target groups often have lacked components that are especially important for older students. Neither the question of whether interventions are less effective for older students, nor whether combinations of components are important is settled in the reviews covered in this section.

The contribution of this review

Academic difficulties and lack of educational attainment are significant societal problems, and special education is challenging and costly, not least because research on ability grouping indicates that grouping students based on prior displayed abilities or subjective expectations about their abilities might have the unintended consequence of reproducing social inequalities in educational attainment (Condron, 2008; Gamoran 2004; Hattie 2002; Justice, Petscher, Schatschneider, & Mashburn, 2011; Kerckhoff 1993; Lubbers, Snijders, & Van Der Werf, 2011; Schofield 2010; Van de Werfhorst & Mijs 2010). Moreover, as shown by the Salamanca declaration from 1994 (UNESCO, 1994), there has for decades been a great interest among policy makers to improve the inclusion of students with academic difficulties in mainstream schooling, and a desire to increase the number of empirically supported interventions for these student groups.

The main objective of this review is to provide policy makers and educational decision-makers at all levels – from governments to teachers – with evidence of the effectiveness of interventions aimed to improve the results of students with or at-risk of academic difficulties. To achieve this objective we will compare the effects of interventions that differ in terms of their components regarding both instructional methods and the content taught. To be specific, we are interested in providing evidence on whether for example tutoring improves educational achievement. However, we would also like to examine whether tutoring interventions improve educational achievement more than, say, cooperative learning interventions, and if interventions work better in mathematics than in reading, or when they emphasize vocabulary rather than fluency. Furthermore, it is presently not known whether interventions that combine components, for example cooperative learning combined with a component that gives teachers and students frequent feedback on student progress, or tutoring combined with socio-emotional training, are more effective than interventions that use only a single component.

To this end, we have chosen a broad scope in terms of the target group and the types of interventions we include. We will also include interventions where the effects are measured

by standardised tests in reading and mathematics. The reason is that many interventions are not directed specifically to either subject and outcomes are therefore often measured in both (Dietrichson et al. 2015a). Earlier reviews of interventions to reasonably similar target groups (e.g. Gersten et al. 2009, Slavin et al. 2011, Dietrichson et al. 2015a) provide tentative evidence that similar types of interventions are effective for both struggling and low SES students, but more knowledge about whether this is so would be welcome. That this knowledge is not complete is a reason to keep both the types of interventions we include and the target group relatively broad. Including both students with and at-risk of academic difficulties in the target group should also decrease the risk of biasing the results due to omission of studies where information about either academic difficulties or at-risk status is available, but not both. Furthermore, making comparisons over intervention components such as instructional methods and content domains within one review, rather than across reviews, should increase the possibilities of a fair comparison. For instance, controlling that effect sizes are calculated in the same way, that the definitions of intervention components are consistent, and that moderators are coded in the same way, is easier within the scope of one review.

In isolation, this last argument suggests that all interventions aiming to improve educational achievement for our target population should be included. However, we also want to explore why certain interventions work better than others. The results in the reviews of for example Slavin et al. (2008, 2009) and Dietrichson et al. (2015a) point to substantial variation in effect sizes aimed to improve test scores in reading and mathematics. Importantly, this variation is also found within types of interventions. For the exploration of variation in effect sizes, a broad scope may turn into a disadvantage, as information about moderators that are important in order to explain variation for some types of interventions are not relevant for others. We have therefore delimited the included interventions to those that are targeted, rather than universal, and performed in a regular school situation during the regular school year. This delimitation increases the probability that potentially important moderators, such as dosage are reported in a comparable way.

Hopefully, the review should therefore be able to provide guidance about what components of interventions, and combinations of components, that are effective. Earlier reviews with a comparable focus have either not included intervention components together with other moderators in a meta-regression, or only included broad categories of interventions. For example, reviews have coded interventions over contrasts between treatment and control groups regarding the instructional methods used, or regarding the type of content taught, but not both (e.g. Dietrichson et al., 2015a; Gersten et al., 2009; Scammaca et al., 2015). Thus, the first risks confounding the effects of intervention components with for example participant characteristics, and the second risks confounding methods with content.

OBJECTIVES

The objective of this review is to assess the effectiveness of interventions aimed at students with or at-risk of academic difficulties in grades 7 to 12 for increasing academic abilities and enhancing educational outcomes, as measured by standardised tests in reading and mathematics.

The analysis will centre on the comparative effectiveness of different types of interventions in an attempt to identify those intervention components that have the largest and most reliable effects on academic outcomes as measured by standardised test scores. In addition, evidence of differential effects for students with different characteristics will be explored, e.g., in relation to age or grade, gender, and socioeconomic status. We will also examine moderators related to study design, measurement of effect sizes, and the dosage and delivery of interventions.

METHODS

Characteristics of the studies relevant to the objectives of the review

We will include three types of study designs in the review: randomised controlled trials (RCT), quasi-randomised controlled trials (QRCT), and quasi-experimental studies (QES). A fair amount of studies within educational research use single group pre-post comparisons (e.g. Edmonds et al., 2009; Wanzek et al., 2006); such studies will however not be included. See the next section "Criteria for inclusion and exclusion of studies in the review" for more details about when the different study designs will be included.

We expect that a certain amount of studies are conducted without randomisation of participants (24 percent are QES in Dietrichson et al., 2015a). The main reason for including QRCTs and QESs is that we want the review to be as comprehensive as possible and we expect that there will be information that is contained in QRCTs and QESs that are of relevance to this review. For example, in some circumstances it may be difficult to conduct blind RCTs in educational research. This may for instance imply that control groups, their teachers, and/or their parents know that the control group students did not receive the treatment. Such knowledge may alter behaviour and imply that the control group is affected by the intervention. RCTs do not necessarily provide more credible measures of intervention effects in such situations. Furthermore, RCTs and QRCTs require providers to prescribe treatment based on lotteries or other means of semi-randomisation instead of professional assessment. Therefore, randomisation designs may also raise issues concerning the self-perceived professional integrity of the providers and institutions taking part in experimental research, and thereby complicate study feasibility. We will include study design as a potential moderator in the meta-analysis.

One example of a QES likely to be included is Fuchs et al. (1999), who study the effects of a peer-assisted learning strategies (PALS) programme on reading comprehension and fluency for struggling readers in high school. A PALS session comprises three activities: partner reading, paragraph shrinking, and prediction relay. A total of 102 students (52 treated) with low levels of reading proficiency were included. Researchers assign treatment to nine teachers and control group status to nine other teachers. Statistical tests showed small pre-treatment differences between treatment and control groups on important confounders such as grade, age, prior reading level, gender, free/reduced lunch status, race, type of reading class, and disability status. Treatment consisted of teachers supplementing their reading instruction with PALS sessions five times every two weeks for the duration of 16 weeks, while the control condition had teachers providing instruction using their conventional programme (which had no peer-mediated learning activities). The study reported means and standard deviations.

An RCT likely to be included is Allinder, Dunse, Brunken & Obermiller-Krolikowski (2001). They randomise the instruction of how to use oral reading strategies among 50 grade 7 students in three remedial reading classes in a suburban middle school. Randomisation was made on the individual level, and the control group received the intervention after the current study was completed. Means and standard deviations were reported.

Criteria for inclusion and exclusion of studies in the review

Types of interventions

For intervention studies to be included in the review it must be clear that the intervention is structured so that it works to improve academic achievement or specific academic skills. This does not mean that the intervention must consist of academic activities, but rather that the explicit expectation must be that the intervention, regardless of the nature of the intervention content, will result in improved academic performance or a higher skill level in a specific academic task. Furthermore, an explicit academic aim of the intervention does not per se exclude interventions that also include non-academic objectives and outcomes.

Interventions without academic outcomes or interventions having academic learning as a possible secondary goal (such as interventions where behavioural or socioemotional problems is the primary intervention aim, like Classroom Management or Families and Schools Together) will be excluded. However, interventions with behavioural and socio-emotional components may very well have academic achievement as one of their primary aims (e.g. some mindset and stereotype threat interventions). Such interventions will be included if this aim is made explicit in the study (and the outcomes are measured by standardised tests in reading or mathematics, see below section Types of outcome measures for more details).

Furthermore, we will only include school-based interventions; that is, interventions performed in schools during the regular school year, and schools are one of the stakeholders.

Judging by the results in the related review of Dietrichson et al. (2015a), this restriction excludes summer reading programs and some after-school programs (which may, but need not, be performed outside of school by other actors). Both of these types of interventions appear to be using qualitatively different components compared to interventions performed in school. They are often also different in terms of for example who deliver them, how different aspects of intervention dosage are measured, and whether and how implementation is assessed. In addition, there is a very recent review of summer reading programs (Kim & Quinn, 2013), and one earlier Campbell review of after-school programs (Zief et al., 2006). Our criteria would also exclude for example parent tutoring programmes and other programmes delivered in the home of students. If interventions are mainly delivered in school during the school year, but also include a component delivered outside of school, they will be included.

Besides having as their explicit primary expectation that the intervention will improve the academic performance of the student, eligible interventions for review must also be targeted (or selected/indicated). That is, interventions which, in contrast to universal interventions, are aimed at certain students and/or student groups identified as having academic difficulties, or being at-risk of such difficulties (see below for a detailed description of the types of participants we will include).

Universal interventions, applied to improve the quality of the common learning environment at the school level in order to raise academic achievement of all students (including average and above average students), will be excluded. Interventions such as the one described in Fryer (2014) where a bundle of best practices are implemented at the school level in low achieving schools, where most or possibly all students are struggling or at risk, will therefore be excluded. This criteria also excludes whole-school reform strategy concepts such as Success for All, curriculum-based programmes like Elements of Mathematics (EMP), as well as reduced class size interventions. It also excludes interventions where teachers or principals receive professional development training in order to improve general teaching or management skills. Interventions targeting students with or at-risk of academic difficulties may on the other hand include a professional development component, for example when a reading programme includes providing teachers with reading coaches. Such interventions will be included.

Types of participants

The population samples eligible for the review include students attending regular schools in grades 7-12, who are having academic difficulties, or are at-risk of such difficulties. Students attending regular private, public, and boarding schools are included, and students receiving special education services within these school settings are also included. Grades 7-12 corresponds roughly to secondary school, defined as the second step in a three-tier educational system consisting of primary education, secondary education and tertiary or higher education. The number of years a child attend secondary schooling varies across the

OECD countries, though most often secondary schooling is grades 7-12 or 10-12. The former is the case for instance in France, Spain, Japan, UK, and most parts of Australia, and the second is the case for school systems in countries such as Italy, Turkey, Sweden and Denmark. We will include studies with a student population younger than 7-12 as long as the majority of the students are in grades 7-12. The age range included will also differ between countries, and sometimes between states within countries. Typically, ages will range from 12-14 to 17-19.

The eligible student population includes both students identified in the studies by their observed academic achievement (e.g., low academic test results, low grade point average or students with specific academic difficulties such as learning disabilities), and students that have been identified primarily on the basis of their educational, psychological, or social background (e.g., students from families with low socioeconomic status, students placed in care, students from diverse ethnic/cultural backgrounds, and second language learners). We will however exclude interventions targeting students with physical learning disabilities (e.g. blind students), students with dyslexia/ dyscalculia, and interventions that are specifically directed towards students with a certain neuropsychiatric disorder (e.g. autism, ADHD), as these interventions are probably very different from interventions targeting the general struggling or at-risk student population.

We believe it is important to include students that for other reasons are struggling together with groups that are deemed at-risk, or are considered educationally disadvantaged. There is substantial overlap between these groups in the studies we have found in a previous review (Dietrichson et al. 2015a). A motivating example comes from studies that target a high poverty area, and then randomly select a number of students with test scores below a certain level in each school that receive the intervention. These students are thus likely to be low SES, but information about SES is not always included. That is, shares of low SES students are only reported on the school or district level, and sometimes not at all. A second example would be studies that target low performing schools, and then perform an intervention for the sub-group of low SES students. In this case, low SES students are likely to be struggling, although this information is not always included.

Thus, choosing to include only studies that examine either students with academic difficulties or low SES students may exclude studies that in all likelihood target the same student population. We think that the risk of biasing our results by such a choice is larger than the possible comparison problems arising from including both students with academic difficulties and low SES students. A similar case can be made for other at-risk groups, for example students from diverse ethnic/cultural backgrounds, which in many cases overlap with low SES students.

Finally, there are also good reasons to suspect a substantial overlap of the reasons for why these groups need interventions. While the earlier literature has not fully converged on a ranking of these reasons, the differential access to family resources is a major contributor to

these groups' educational disadvantage; something which school-based interventions may compensate for. The reasons for low performance are thus likely connected to the challenges faced by at-risk students.

Some interventions may include other students, who are neither with nor at-risk of academic difficulties. An example may be a cooperative learning intervention where high performing students are paired with struggling students. Studies of such interventions will be included if the total sample (treatment and control group) include at least 50% students that are either having academic difficulties or are at-risk of developing such difficulties.

Types of outcome measures

As the overall purpose of the review is to evaluate evidence on effects of educational interventions on academic achievement, we will include outcomes that cover two main areas of fundamental academic skills:

- Standardised tests in reading
- Standardised tests in mathematics

Studies will only be included if they consider one or more of the primary outcomes. As standardised tests, we will consider norm-referenced tests (e.g. Gates-MacGinitie Reading Tests and Star Math), state-wide tests (e.g. Iowa Test of Basic Skills), and national tests (e.g. National Assessment of Educational Progress). If it is not clear from the description of outcome measures in the studies, we will use electronic sources to determine whether a test is standardised or not. For example, if a commercial test has been normed, this is typically mentioned on the publisher's homepage. If there is no such mention, we will consider the test as being not standardised.

We restrict our attention to standardised tests in part to increase the comparability between effect sizes. Earlier related reviews of academic interventions have pointed out that effect sizes tend to be significantly lower for standardised tests compared to researcher-developed tests (e.g. Flynn et al., 2012; Gersten et al., 2009; Scammaca et al., 2015). Scammaca et al. (2015) furthermore reported that whereas mean effect sizes differed significantly between the periods 1980-2004 and 2005-2011 for other types of tests, mean effect sizes were not significantly different for standardised tests. As researcher developed tests are usually less comprehensive and more likely to measure aspects of content inherent to treatment but not control group instruction (Slavin & Madden, 2011), standardised tests should provide a more reliable measure of lasting differences between treatment and control groups. For this reason, we will not consider tests where researchers have picked a subset of questions from a norm-referenced test as being standardised. In sum, while researcher developed tests may be highly useful for certain purposes (e.g. testing specific intervention mechanisms), we believe they would be less useful for the purposes of this review.

We will include tests of specific domains (e.g. vocabulary, fractions) as well as more general tests, which test several domains of a subject. Tests of subdomains have significantly larger effect sizes compared to more general tests in Dietrichson et al. (2015a). This result may indicate that interventions often target certain domains and not general performance skills, or that it may be easier to improve scores on tests of subdomains than on tests of more general skills, or that tests of subdomains may be more likely to be inherent to treatment (see Slavin & Madden, 2011 for a discussion of the latter). At the same time it seems reasonable that interventions that target subdomains of reading and mathematics be tested on whether they affect these subdomains. Therefore, we do not want to exclude either type of test, but will code the type of test, as well as the content domain of the intervention and use the type of test as a variable in the moderator analyses.

Based on findings in Dietrichson et al. (2015), we expect that a large majority of studies only have reported outcomes of tests performed within 3 months after the end of intervention. We will consider longer run outcomes as well, if they are available (see section Multiple time points below).

There are many other important outcome measures that we do not include (e.g. grades, dropout, and uptake of secondary/tertiary education). We make this choice to streamline the review, and to increase comparability across contexts. Grade setting and the presence of certain education options (e.g. vocational training tracks) are likely to differ more across school systems and countries than standardised tests.

Types of study designs

Types of studies included are studies that use a treatment-control group design or a comparison group design, and adequately address the subject of effectiveness of interventions to improve the students' academic achievement: RCTs, including cluster-RCTs; QRCTs, i.e., where participants are allocated by means such as alternate allocation, person's birth date, the date of the week or month, case number, or alphabetical order; and QES. To be included, QES must credibly demonstrate that outcome differences between treatment and control groups is the effect of the intervention and not the result of systematic baseline differences between groups. That is, selection bias should not be driving the results. This assessment is included as a part of the risk of bias tool, which we elaborate on in section *Risk of bias*, and in Appendix C.

A control group is defined as a non-treatment condition; a comparison group is defined as an alternative treatment condition. Eligible types of control groups include waitlist controls and no-treatment controls. However, in this review the waitlist controls and no-treatment controls only differ in the time frame in which researchers can follow the differences between groups because students in both waitlist and no-treatment controls are offered regular schooling by default.

Comparison designs compare alternative treatments against each other. Comparison designs will be analysed separately from treatment-control designs. We elaborate in section Synthesis procedures and statistical analysis on how we will use comparison designs. Studies using single group pre-post comparison will not be included. Effect sizes from such studies are not comparable to effect sizes from treatment-control designs if, for example, there is progression in students' knowledge over time, which is typically the case.

Duration of interventions

There will be no initial criteria for duration of interventions, but the duration of included interventions will be coded for the review.

Types of settings

Only studies carried out in OECD countries will be included. This selection is conducted to ensure a certain degree of comparability between school settings to align treatment as usual conditions in included studies. For similar reasons we will only include studies published in or after 1980. Due to language restrictions, we will only include studies written in English, German, Danish, Norwegian, and Swedish.

Search strategy for finding eligible studies

This section describes the search strategy for finding potentially relevant studies. We will use EPPI software to track the search and screening process.

Electronic databases

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

- Academic Search Premier
- Australian Education Index
- British Education Index
- CBCA Education
- Centre for Reviews and Dissemination Databases
- Cochrane Library
- Cristin
- DIVA
- Education Research Complete
- Embase
- ERIC
- Forskningsdatabasen.dk
- FRANCIS
- Medline
- PsycINFO

- ProQuest dissertation & theses A&I
- Social Science Citation Abstract
- Science Citation Abstract
- Socindex
- Social Care Online
- Theses Canada

Search terms

An example of the search strategy for ERIC searched through the Ovid 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. The strategy contains also terms on primary school, since the search also will contribute to a review about this younger age group. There may be overlap in the literature among the age groups, and in order to rationalize and accelerate the screening process, we have decided upon performing one extensive search.

1. (Underachiev* or Under n1 achiev* or lowachiev* or low n1 achiev* or Low N1 perform* or lowperform* or (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*))
2. ((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*))
3. Child* N2 placed n1 care or (DE "Foster Care") AND child*
4. (Student* or pupil*) N3 (Learn* N2 (disab* or Problem*))
5. S1 or S2 or S3 or S4
6. DE "Academic Achievement" or DE "Academic Ability" or DE "Learning Problems" or (DE "Learning Disabilities")
7. Learn* N2 (disab* or Problem*)
8. Academic* N2 (performance* or achiev* or abilit* or outcome*)
9. School N1 (performan* or achiev*)
10. DE "Intellectual Development"
11. Intellect* N2 develop*
12. S6 or S7 or S8 or S9 or S10 or S11
13. DE "Reading" or DE "Literacy"
14. Reading or Literacy
15. DE "Mathematics" or DE "Numeracy"
16. Numeracy or Mathematic* or Math
17. transfer* N2 effect
18. S13 or S14 or S15 or S16 or S17
19. S5 and S12 and S18
20. AB randomized or AB placebo or AB randomly or trial or AB groups

21. DE "Cohort analysis" or DE "Case Studies"
22. TI ((case control) or AB (case control)) or TI cohort or AB cohort
23. TI cross sectional or AB cross sectional
24. (TI (epidemiologic N2 study) or AB (epidemiologic N2 study)) or (Ti (follow up or followup) N2 study) or AB ((follow up or followup) N2 study))
25. (TI longitudinal or AB longitudinal) or (TI observational or AB observational)
26. TI ((prospective n2 study) or AB (prospective n2 study)) or (TI retrospective or AB retrospective)
27. TI Intervention* N1 Stud* or AB Intervention* N1 Stud*
28. TI (quasi-experiment* or quasiexperiment* or experiment*) or AB (quasi-experiment* or quasiexperiment* or experiment*)
29. TI assign* N3 (subject* or patient*) or AB assign* N3 (subject* or patient*)
30. 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*))
31. TI Non-random* or nonradom* or (non N1 random*) or AB Non-random* or Nonrandom* or (non N1 random*)
32. TI ((random* N2 trial*) or RCT) OR AB ((random* N2 trial*) or RCT)
33. 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 AB (quasi-experiment* or quasiexperiment* or Propensity score* or (compar* N1 group*) or (match* N1 control*) or (match* N1 group*) or (match* N1compar*) 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 n1samp*) or propensity match* or (Between N1 group*))
34. ((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))
35. 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*))
36. 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))

37. 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))
38. 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))
39. 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*)
40. S20-S39/or
41. S19 and S39

Searching other resources

The review authors will check reference lists of other relevant reviews and included primary studies for new leads. Citation searching in the Web of Science will also be considered. We will contact international experts to identify unpublished and ongoing studies, and provide them with the inclusion criteria for the review along with the list of included studies, asking for any other published, unpublished or ongoing studies relevant for the review. We will primarily contact corresponding authors of the related reviews mentioned in the section Prior reviews, but extend the contacts to others if we find references to or mentions of ongoing studies in screened publications. We will also search two trial registries: The Institute for Education Sciences' Registry of Randomized Controlled Trials (http://ies.ed.gov/ncee/wwc/references/registries/index.aspx), and American Economic Association's RCT Registry (https://www.socialscienceregistry.org).

Handsearch

The following international journals will be hand searched for relevant studies:

- American Educational Research Journal
- Journal of Educational Research
- Journal of Educational Psychology
- Journal of Learning Disabilities
- Journal of Research on Educational Effectiveness
- Journal of Education for Students Placed at Risk

The search will be performed on editions from 2015 to review submission of the journals mentioned, in order to capture any relevant studies recently published and therefore not captured in the systematic search.

Grey literature

Additional searches will be made by means of Google and Google Scholar and we will check the first 150 hits. OpenGrey (http://www.opengrey.eu/) will also be used to search for European grey literature. Copies of relevant documents will be made and we will record the exact URL and date of access for each relevant document. In addition we will search the following sites:

- What Works Clearinghouse - U.S. Department of Education, http://www.whatworks.ed.gov.
- Dansk Clearinghouse for Uddannelsesforskning, edu.au.dk/clearinghouse.
- European Educational Research Association (EERA), http://www.eera-ecer.eu.
- American Educational Research Association (AERA), www.aera.net
- Deutsche Gesellschaft für Erziehungswissenschaft (DGfE), German Educational Research Association (GERA), http://www.dgfe.de/
- NBER working paper series, http://www.nber.org
- Best Evidence Encyclopedia, http://www.bestevidence.org/

Data extraction and study coding practices

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

Two members of the review team will independently code and extract data from included studies. A coding sheet will be piloted on several studies and revised as necessary (see Appendix B). Any disagreements will be resolved by discussion. Data will be extracted on the characteristics of participants (e.g. age, gender, at-risk status), characteristics of the intervention and control/comparison conditions, research design, sample size, outcomes, and results. Extracted data will be stored electronically, and we will use EPPI, Microsoft Excel, and Stata as primary software tools.

Risk of bias

We will assess the risk of bias of effect estimates using a risk of bias model developed by Prof. Barnaby Reeves in association with the Cochrane Non-Randomised Studies Methods Group. This model is an extension of the Cochrane Collaboration's risk of bias tool and covers risk of bias in non-randomised studies that have a well-defined control group. The extended model is organised and follows the same steps as the risk of bias model according

to the 2008-version of the Cochrane Hand book, chapter 8 (Higgins & Green, 2008). The extension to the model is explained in the three following points:

1) The extended model specifically incorporates a formalised and structured approach for the assessment of selection bias in non-randomised studies by adding an explicit item about confounding. This is based on a list of confounders considered to be important and defined in the protocol for the review. The assessment of confounding is made using a worksheet where, for each confounder, it is marked whether the confounder 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 C). This assessment will inform the final risk of bias score for confounding.

2) Another feature of effect estimates in non-randomised studies that make them at high risk of bias is that they need not have a protocol in advance of starting the recruitment process (this is however also true for many RCTs in education). 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, e.g., choice of method of model fitting, potential confounders considered/included. In addition, the model includes two separate yes/no items asking reviewers whether they think the researchers had a pre-specified protocol and analysis plan.

3) Finally, the risk of bias assessment is refined, making it possible to discriminate between effect estimates with varying degrees of risk. This refinement is achieved with the addition of a 5-point scale for certain items (see the next section and Appendix C for details).

The refined assessment is pertinent when thinking of data synthesis as it operationalizes the identification of studies (especially in relation to non-randomised studies) with a very high risk of bias. The refinement increases transparency in assessment judgements and provides justification for not including a study with a very high risk of bias in the meta-analysis.

Risk of bias judgement items

The risk of bias model used in this review is based on nine items (see Appendix C for a fuller description). The nine items refer to: Sequence generation, allocation concealment, blinding, incomplete outcome data, selective outcome reporting, other potential threats to validity, a priori protocol, a priori analysis plan, and confounders (for non-randomised studies).

Confounding

An important part of the risk of bias assessment of effect estimates in non-randomised studies is how studies deal with confounding factors. 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. motivation). Included studies use for example matching,

difference-in-differences, and statistical controls to mitigate selection bias, or demonstrate evidence of pre-treatment equivalence on key risk variables and participant characteristics. In each study, we will assess whether confounding factors have been considered. Furthermore, we will assess how each study deals with unobservables.

There is no single non-randomised 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. For example, differences in pre-treatment test score levels do not have to be a problem in a difference-in-differences design, where the main identifying assumption is that the trends of the outcome variable in the treatment and control group would not have differed, had the intervention not occurred. Similar differences in levels would, in general, be more problematic in a matching design as they indicate that the matching technique has not been able to balance the sample even on observable variables. For this reason, we will not specify thresholds in terms of pre-treatment differences (in say, effect sizes) for when a study has too high risk of bias on confounding. Each QES will be assessed in terms of the risk that the effect of the intervention is being confounded with observed and unobserved variables.

Importance of pre-specified confounding factors

The motivation for focusing on age and grade level, performance at baseline, gender, socioeconomic background and local education spending is given below.

Development of cognitive functions relating to school performance and learning are age dependent, and furthermore systematic differences in performance level often refer to systematic differences in preconditions for further development and learning of both cognitive and social character (Piaget, 2001; Vygotsky, 1978). Therefore, to be sure that an effect estimate is a result from a comparison of groups with no systematic baseline differences it is important to control for the students' grade level (or age).

Performance at baseline is generally a very strong predictor of post-test scores (Hedges & Hedberg, 2007), and controlling for this confounder is therefore highly important.

With respect to gender it is well-known that there exist gender differences in school performance (Holmlund & Sund, 2005). Girls outperform boys with respect to reading and boys outperform boys with respect to mathematics (Stoet & Geary, 2013), although part of the literature finds that these gender differences vanished over time (Hyde, Fennema, & Lamon, 1990; Hyde & Linn, 1988). As there is no consensus around the disappearance of gender differences, we find it important to include this potential confounder.

Students from more advantaged socioeconomic backgrounds on average begin school better prepared to learn and receive greater support from their parents during their schooling years (Ehrenberg et al., 2001, Fryer & Levitt, 2013). As outlined in the background section,

students with socio-economically disadvantaged backgrounds have lower test scores on international tests (OECD, 2010, 2013). Therefore, the accuracy of the estimated effects of an intervention may depend on how well socioeconomic background is controlled for. Socioeconomic background factors are, e.g. parents' educational level, family income, ethnic/cultural background, etc.

Bias assessment in practice

At least two, review authors will independently assess the risk of bias for each included study. Disagreements will be resolved by discussion. We will report the risk of bias assessment in risk of bias tables for each included study.

In accordance with Cochrane and Campbell methods we will not aggregate the 5-point scale across items. Effect sizes given a rating of 5 on any item will not be included in the meta-analysis (the items with a three-point scale do not warrant exclusion). We will only give 5 points for an item to denote a very high risk of bias. A stark example would be a study with 100% attrition in the comparison group and no follow up data. Effect sizes from this study would receive 5 points on incomplete outcome data. This study would not be included in the meta-analysis as it in effect has become a single group study with pre- and post- measures for the experimental group only. Further examples may be QES, which have not controlled for any confounders or not reported the balance on any pre-treatment tests, or studies that completely confound treatment with other effects. An example of the latter is when treatment is assigned on school level and there is one treated school and one control school. Treatment is then completely confounded with school effects. For studies with a lower than 5-point rating, we will use the ratings of the major items in sensitivity analyses.

Synthesis procedures and statistical analysis

Analysis of absolute effects will involve comparing an intervention to no treatment conditions and to untreated waitlist controls. Analysis of studies comparing different interventions (comparison designs) will be conducted separately. We will also conduct separate analyses for short- and long-term outcomes. The analysis plan laid out below applies to both types of outcomes.

Effect sizes using continuous data

For continuous data, standardized mean differences (SMDs) will be calculated 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):

$$(1) \quad g = \left(1 - \frac{3}{4N - 9}\right) \times \left(\frac{\bar{X}_1 - \bar{X}_2}{s_p}\right)$$

$$(2) \quad SE_g = \sqrt{\frac{N}{n_1 n_2} + \frac{g^2}{2N}}$$

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

$$(3) \quad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}}.$$

Here, s_1 and s_2 denotes the standard deviation of the treatment and control group. We will use covariate adjusted means whenever available, and the unadjusted standard deviation. We will use intention-to-treat (ITT) estimates of the mean difference whenever possible, and test whether results are sensitive to the inclusion of treatment-on-the-treated (TOT) or local average treatment effects (LATE). 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 analyse change scores and final values separately.

Effect sizes using discrete data

Based on findings in Dietrichson et al. (2015a), where only two out of 101 included studies exclusively reported discrete outcome measures, we expect that almost all studies in this literature use continuous outcome measures. We therefore expect to use the methods described in Sánchez-Meca, Marín-Martines, & Chacón-Moscoso (2003) to transform any dichotomous outcomes into SMDs.

Should we find a large enough number of studies using dichotomous outcomes, we will test whether our results are sensitive to combining dichotomous and continuous outcome measures. If this is the case, we will also perform a sensitivity analysis using only dichotomous measures, and the following procedure to calculate effect sizes: We will use the natural logarithm of odds ratios (LOR) in the calculations, together with 95% confidence intervals and p-values, and then convert the results back to the original odds ratios once the meta-analysis is performed. The LOR and its approximate standard deviation are calculated as (Lipsey & Wilson, 2001:53-54):

$$(4) \quad LOR = log\left(\frac{ad}{bc}\right)$$

$$(5) \quad SE_{LOR} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

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.

Outliers

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

Dealing with missing data

Missing data and attrition rates in the individual studies will be assessed using the risk of bias tool. 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.[5]

If missing summary data necessary for the calculation of effect sizes cannot be derived or retrieved, 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. If data is missing regarding moderators, we will use methods for multiple imputation in order to not bias our results by excluding these studies (see e.g. Rubin (1996) and Pigott (2009) for why leaving out studies/effect sizes with missing values normally yields biased estimates). We will use the Stata command *mi impute* with sequential imputation using chained equations to generate values for missing observations. All variables without missing observations will be used in the estimation to impute values for variables with missing observations.

Clustered assignment of treatment

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 treatment and control 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. QES may also include clustered assignment of treatment. Effect sizes and standard errors from such studies may be biased if the unit-of-analysis is the individual and an appropriate cluster adjustment is not used (Higgins & Green, 2011).

If possible, we will adjust effect sizes individually using the methods suggested by Hedges (2007) and information about the intra-cluster correlation coefficient (ICC), realized cluster

[5] We will inform about a deadline when sending our inquiry. If the trial investigators notify us before this deadline that they will be able to provide the information in a short time, we will include it even though provided after the deadline. All inquiries and answers will be stored electronically.

sizes, and/or estimates of the within and between variances of clusters. If it is not possible to obtain this information, we will adjust effect sizes using estimates from the literature of the ICC (e.g. Hedges & Hedberg, 2007), and assume equal cluster sizes. To calculate an average cluster size, we will divide the total sample size in a study by the number of clusters (typically the number of classrooms or schools).

Multiple intervention groups and multiple interventions per individual

Studies with multiple intervention groups with different individuals, and studies using multiple tests for the same intervention groups, will be included in the review. To avoid problems with dependence between effect sizes we will apply the robust variance estimation methods developed by Hedges, Tipton, & Johnson (2010). If we do not find enough studies in order for this method to consistently estimate the standard errors (Tanner-Smith & Tipton, 2014; Tipton, 2014) we will conduct a data synthesis where we use a synthetic effect size (the average) in order to avoid dependence between effect sizes. See below for more details about how we plan the data synthesis.

Studies including multiple interventions per individual may also be included, but only one intervention group (control or comparison group) will be coded and compared to the control or comparison group (intervention group) to avoid overlapping samples. We will choose the estimate from the intervention that we judge to have the least risk of bias.

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 effect from each sample of data to avoid dependencies. The choice of which estimate to include will be based on our risk of bias assessment. We will choose the estimate from the study that we judge to have the least risk of bias.

Multiple time points

Outcomes will, if possible, be considered for the following intervals:

- Short-term effects (less than 3 months after the end of intervention).
- Medium- to long-term effects (3 months or more after the end of intervention).

We realize that 3 months is not a particularly long-term period. However, we expect having to use these definitions of short- and long-term effects based on the findings in Dietrichson et al. (2015a), where very few studies where found that reported outcome measurements more than 3 months after the end of intervention. Even fewer reported results after more than 6 months. If there are more studies reporting longer term effects found for this review, we will consider changing our definition of medium- to long term effects.

Data synthesis

The overall data synthesis in this review will be conducted where effect sizes are available. 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.

Random effects inverse variance weighted mean effect sizes will be used for all parts of the analysis and we will report 95% confidence intervals. The weighting function will be:

$$(4) \quad w_i = \frac{1}{SE_i^2 + \tau^2}$$

where w_i is the weight assigned to effect size i, SE_i^2 is the variance of i as defined by equation (2), and τ^2 is the random effects variance component estimated for each analysis with a method of moments or maximum likelihood estimator.

The analysis will be conducted in the following steps: Summary and descriptive statistics of the study-level contextual characteristics, methodological quality characteristics, group and subject level characteristics, as well as outcome characteristics will be used to describe the included studies. We will also include a correlation matrix with all moderators. Main effects analysis will be conducted first. Heterogeneity will be assessed with Chi-squared (Q) test, and the I-squared, and τ-squared statistics (Higgins, Thompson, Deeks, & Altman, 2003).

If there is heterogeneity in effect sizes, we will perform a moderator analysis to attempt to identify the characteristics of study methods, interventions, and participants that are associated with smaller and larger effects on the various outcomes. We will use a mixed-model meta-regression to minimize the risk of misleading results due to correlated independent variables. We will start by pooling all effect sizes from studies with a treatment-control design (see below for a description of the analysis of comparison designs) and include the following types of moderators (variables in parentheses):

- Subject (math or reading test score)
- Study design (RCT, QRCT, or QES)
- Effect size measurement (type of test)
- Participant characteristics (share of girls, grade level of sample or age, share of target group, subgroup of target group (e.g. low SES)
- Treatment modality (type of instructional method(s) and content domain)
- Dosage (duration, frequency, intensity)
- Implementation quality

The exact definition of the moderators may be subject to change during the data extraction process, but see Appendix B for preliminary version of the code book including more details on some of the moderators.

We will report 95% confidence intervals for regression parameters. To avoid problems with dependence between effect sizes we will apply robust standard errors (Hedges et al., 2010),

using the Stata command *robumeta*. If there is significant heterogeneity also in the moderator analysis, this will warrant further examination of sub-groups. Sub-group examination could take the form of using interaction variables. However, we do not expect to find enough studies in order to run a meta-regression model where all relevant interactions are included. For example, interacting all instructional components and relevant combinations of these with all content domains while at the same time including other moderators would require a very large number of studies in order to not run into problems with degrees of freedom. On the other hand, it seems highly unlikely that the number of included studies will be so small that robust variance estimation becomes completely infeasible. The review of Dietrichson et al. (2015a) contains 25 studies of interventions targeting low SES students in middle school performed during 2000-2014. The target group for this review is broader and we will include studies of interventions further back in time. Therefore, we expect to include much more than 25 studies. We will use the simulation results reported in Tipton (2014) and Tanner-Smith & Tipton (2014) to assess how many moderators that can be included in each meta-regression.

If all moderators listed above cannot be included in the same regression due to limited degrees of freedom, we will proceed in two ways. First, we will exclude highly correlated variables, starting with moderators that have a higher correlation than 0.7, and then move down to 0.5, if necessary. Second, we will try factor analysis.

Subgroup analyses will be the next step. The primary objective of the review is to provide educational decision-makers with evidence of the effectiveness of interventions aimed to improve the results of educationally disadvantaged students. We will therefore focus the subgroup analysis on instructional methods and content domains. These are substantive features of interventions that for example teachers and school managers can affect, in contrast to other moderators (e.g. participant characteristics may be more difficult to affect for a school). The final categories of instructional methods and content domains will be developed during coding, but see section Interventions in practice for a description of what methods and domains that have been found in related reviews.

We will, if the number of studies allows it, use mixed-model meta-regressions and robust variance estimation in all sub-group analyses. The exact specification will depend on the outcome of the meta-regressions on the full sample of effect sizes. We will proceed in one of the following ways: 1) If there are enough studies on each instructional method/content domain so that an indicator for each component can be included in the meta-regressions using the full sample, we will get an indication of the comparative effectiveness of instructional methods/content domains from these regressions. That is, we can for example test whether one intervention component (or a combination of components) has a larger effect size than another, conditional on other moderators, with a t-test for the regression coefficients. 2) If it is not possible to test differences between instructional methods/content domains via t-tests of regression coefficients, we will evaluate the comparative effectiveness using similar methods as Wilson et al. (2011).

In 1) the subgroup analysis will aim to explain variation of effect sizes within the group of studies using the same instructional methods (if such variation exists). Alternatively, depending on the number of content domains in relation to the number of instructional methods, we will focus on explaining variation of effect sizes between content domains. In this type of sub-group analysis, we are not likely to be able to include all moderators in the same regression. We will then use a similar two-step procedure as described above: first exclude highly correlated variables, and then try factor analysis. If we do not have enough studies, and there is, as in Dietrichson et al. (2015a), small and insignificant differences between effect sizes in math and reading, we will pool studies using both math and reading interventions in the sub-group analysis of instructional methods (which are often the same across the two subjects) using a similar procedure with meta-regressions as described in the previous paragraphs.

In 2) meta-regressions of this kind are not possible. We will evaluate the comparative effectiveness using similar methods as Wilson et al. (2011); that is, adjust effect sizes using variables that do not share variance with the instructional methods/content domains. The procedure artificially makes every study equal on all the variables from the regression model. We will also show the unadjusted average effect sizes per instructional method/content domain for comparison.

For methodological quality, we will consider sensitivity analysis for each major component of the risk of bias tool. Statistical analyses will be conducted using Stata, as well as other software programs if needed.

Comparison designs

We will use comparison designs in the analysis only in cases where they may shed light on an issue, which could not be fully analysed using the sample of treatment-control studies. A concrete example may be that we, as in Dietrichson et al. (2015a), find relatively large but insignificant differences between tutoring interventions that are performed one-to-one and in small groups. Looking at comparison design studies that focus specifically on the issue of whether small-group tutoring can produce similar results as one-to-one tutoring, may then be useful to explore the variation in effect sizes. We will use meta-analytic techniques, including network analysis techniques (e.g. Higgins et al., 2012; Lumley, 2002; White, Barrett, Jackson & Higgins, 2012), to examine such questions if possible. If comparison design effect sizes cannot be pooled, study-level effects will be reported narratively.

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. We will also use Egger's test, and test whether published studies have different effect sizes compared to unpublished studies.

Treatment of qualitative research

We do not plan to include qualitative research in the review.

REFERENCES

Allinder, R. M., Dunse, L., Brunken, C. D., Obermiller-Krolikowski, H. J., (2001). Improving fluency in at-risk readers and students with learning disabilities. *Remedial and Special Education,* 22(1), 48-54.

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: *Prentice Hall.*

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: *Prentice-Hall.*

Berridge, D., Brodie, I., Pitts, J., Porteous, D., & Tarling, R. (2001). *The independent effects of permanent exclusion from school on the offending careers of young people* (Occasional Paper No 71). Research, Development and Statistics Directorate, UK. Retrieved from http://troublesofyouth.pbworks.com/f/occ71-exclusion.pdf

Björklund, A. & Salvanes, K. (2011). Education and family background. In Hanushek, E. A., Machin, S. & Woessmann, L. (eds.), *Handbook of the Economics of Education*, Volume 3, 201-247.

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

Bowman-Perrott, L., Davis, H., Vannest, K., Williams, L., Greenwood, C., & Parker, R. (2013). Academic benefits of peer tutoring: A meta-analytic review of single-case research. *School Psychology Review*, *42*(1), 39-55.

Bradley, R. H. & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology*, 53, 371-399.

Brook, J. S., Stimmel, M. A., Zhang, C., & Brook, D. W. (2008). The association between earlier marijuana use and subsequent academic achievement and health problems: A longitudinal study. *The American Journal on Addictions*, *17*, 155–160.

Bruner, J. S. (2006). *In search of pedagogy, Volume I: The selected works of Jerome Bruner, 1957-1978 (Vol. 1)*. New York NY: Routledge.

Burchinal, M., Steinberg, L., Friedman, S. L., Pianta, R., McCartney, K., Crosnoe, R., & McLoyd, V. (2011). Examining the black-white achievement gap among low-income children using the NICHD study of early child care and youth development. *Child Development*, 82(5), 1404-1420.

Chambers, E. A. (2003). *Efficacy of educational technology in elementary and secondary classrooms: A meta-analysis of the research literature from 1992-2002*. Doctoral dissertation, Southern Illinois University at Carbondale.

Chetty, R., Hendren, N. & Katz, L. (2015). *The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment*, NBER Working Paper no. 21156.

Cheung, A., & Slavin, R. (2012). How features of educational technology applications affect student reading outcomes: a meta-analysis. *Educational Research Review*. Published online May 2012. Retrieved from http://dx.doi.org/10.1016/j.edurev.2012.05.002

Cheung, A., & Slavin, R. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review*, *9*, 88-113.

Cohen, P. A., Kulik, J. A., & Kulik, C. L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, *19*(2), 237-248.

Condron, D. J. (2008). An early start: Skill grouping and unequal reading gains in the elementary years. *The Sociological Quarterly*, *49*, 363-394.

Cook, P. J., Dodge, K., Farkas, G., Fryer, R. J. Guryan, J., Ludwig, J., Mayer, S., Pollack, H., & Steinberg, L. (2014). *The (surprising) efficacy of academic and behavioral intervention with disadvantaged Youth: Results from a randomized experiment in Chicago*. NBER Working Paper no. 19862.

Cullen, J. B., Levitt, S. D., Robertson, E., & Sadoff, S. (2013). What Can Be Done To Improve Struggling High Schools? *The Journal of Economic Perspectives*, 27(2), 133-152.

Currie, J. (2009). Healthy, wealthy, and wise: Socioeconomic status, poor health in childhood and human capital development. *Journal of Economic Literature*, 47(1), 87-122.

De Ridder, K. A. A., Pape, K., Johnsen, R., Westin, S., Holmen, T. L., & Bjørngaard, J. H. (2012). School dropout: A major public health challenge: A 10-year prospective study on medical and non-medical social insurance benefits in young adulthood, the Young-

HUNT 1 Study (Norway). *Journal of Epidemiology & Community Health*. *66*(11), 995-1000.

Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A-M. (2015a). *Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis*. Unpublished manuscript.

Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A-M. (2015b). *Skolerettede indsatser for elever med svagsocioøkonomisk baggrund – En systematisk forskningskortlægning og syntese*. SFI rapport 15:07, København: SFI – Det Nationale Forskningscenter for Velfærd.

Dunn, S. R., & Griggs, A. S. (2007). *Synthesis of the Dunn and Dunn learning-style model research : Who, what, when, where, and so what?* St. John's University.

Edmonds, M. S., Vaughn, S., Wexler, J., Reutebuch, C., Cable, A., Klingler Tackett, K., & Wick Schnakenberg, J. (2009). A synthesis of reading interventions and effects on reading comprehension outcomes for older struggling readers, *Review of Educational Research*, 79(1), 262-300.

Eiberg, M., Lindstrøm, M., Klint Jørgensen, A-M., (2013). Title registration for a systematic review: Academic intervention programmes for improving school outcomes for educationally disadvantaged children and low achieving students in grade K to 6: a systematic review, *Campbell Systematic Reviews*. Retrieved from http://www.campbellcollaboration.org/lib/project/293/.

Esping-Andersen, G. (2004). Untying the Gordian knot of social inheritance. In *Inequality: Structures, Dynamics, and Mechanisms. Essays in Honor of Aage B. Sørensen, Research in Social Stratification and Mobility*, 21, 115-138.

Esping-Andersson, G., Garfinkel, I., Han, W.-J., Magnuson, K., Wagner, S., & Waldfogel, J. (2012). Child care and school performance in Denmark and the United States. *Children and Youth Services Review*, 34, 576-589.

Feinstein, L., Sabates, R., Anderson, T. M., Sorhaindo, A., & Hammond, C. (2006). *Measuring the effects of education on health and civic engagement*. Proceedings of the Copenhagen Symposium. OECD. Retrieved from http://www.oecd.org/education/country-studies/37437718.pdf

Feuerstein, R. S., Falik, L., & Rand, Y. (2006). Creating and Enhancing Cognitive Modafiability: The Feuerstein Instrumental Enrichment Program. *ICELP Publications*.

Flynn, L. J., Zheng, X., & Swanson, H. L., (2012). Instructing struggling older readers: A selective meta-analysis of intervention research. *Learning Disabilities Research & Practice*, 27(1): 21-32.

Flynn, R. J., Marquis, R. A., Paquet, M. P., Peeke, L. M., & Aubry, T. D. (2012). Effects of individual direct-instruction tutoring on foster children's academic kills: A randomized trial. *Children and Youth Services Review*, *34*, 1183-1189.

Fuchs, L. S., Fuchs, D., & Kazdan, S. (1999). Effects of peer-assisted learning strategies on high school students with serious reading problems, *Remedial and Special Education*, *20(5)*, 309-318.

Fryer, R. G. (2011). Financial incentives and student achievement: Evidence from randomized trials. *Quarterly Journal of Economics*, 2011, 126(4), 1755-1798.

Fryer, R. G. (2014). Injecting charter school best practices into traditional public schools: Evidence from field experiments. *Quarterly Journal of Economics*, 129(3), 1355-1407.

Fryer, R. G. & Levitt, S. D. (2013). Testing for racial differences in the mental ability of young children. *American Economic Review*, 103(2), 981-1005.

Forsman, H., & Vinnerljung, B. (2012). Interventions aiming to improve school achievements of children in out-of-home care: A scoping review. *Children and Youth Services Review*, *34*, 1084-1091.

Gamoran, A. (2004). Classroom organization and instructional quality. In H. Walberg & M.Wang (eds.), *Can unlike students learn together? Grade retention, tracking, and grouping*. Greenwich, CT: *Information Age Publishing*, 141-155.

Gardner, H. (1999). *Intelligence reframed: Multiple intelligences for the twenty-first century*. New York, NY: Basic Books.

Gersten, R., Chard, D. J., Jayanti, M., Baker, S. K., Morphy, P., & Flojo, P. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research*, 79(3), 1202-1242.

Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Applied Developmental Psychology*, 24(6), 645-662.

Hackman, D.A. & Farah, M. J. (2009). Socioeconomic status and the developing brain. *Trends in Cognitive Science*, 13(2), 65-73.

Hart, B. & Risley, T. (2003). The early catastrophe – The 30 million word gap by age 3. *American Educator*, spring 2003.

Harvill, E. L., Maynard, R. A., Nguyen, H. T. H., Robertson-Kraft, C., Tognatta, N., & Fester, R. (2012). Protocol: Effects of college access programs on college readiness and

enrolment. *Campbell Systematic Reviews*. Retrieved from http://www.campbellcollaboration.org/lib/project/160/.

Hattie, J. (2002). Classroom composition and peer effects. *International Journal of Educational Research*, *37* (5), 449-481.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, 312, 1900-1902.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, 32(4), 341-370.

Hedges, L.V. & Hedberg, E. C. (2007). Intraclass correlation values for planning group randomized trials in education. *Education Evaluation and Policy Analysis*, 29(1), 60-87.

Hedges, L .V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*, 39-65.

Heller, S., Pollack, H. A., Ander, R., & Ludwig, J., 2013. *Preventing youth violence and dropout: A randomized field experiment*. NBER Working Paper no. 19014.

Heller, S. B., Shah, A. K., Guryan, J., Ludwig, J., Mullainathan, S., & Pollack, H. A. (2015). *Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago*. NBER working paper no. 21178.

Higgins, J. P. T., & Green, S. (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. T., Jackson, D., Barrett, J. K., Lu, G., Ades, A. E., & White, I. R. (2012). Consistency and inconsistency in network meta-analysis: Concepts and models for multi-arm studies, *Research Synthesis Methods*, 3(2), 98-110.

Higgins, J.P., Thompson, S.G., Deeks, J.J., & Altman, D.G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*(7414), 557-60.

Holmlund, H., & Sund, K. (2005). *Is the gender gap in school performance affected by the sex of the teacher?* Swedish Institute for Social Research (SOFI), Stockholm University, Working paper 5/2005.

Horwood, J. L., Fergusson, D. M., Hayatbakhsh, M. R., Najman, J. M., Coffey, C., Patton, G. C., Silins, E., & Hutchinson, D. M. (2010): Cannabis use and educational achievement: Findings from three Australasian cohort studies. *Drug and Alcohol Dependence, 110,* 247–253.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, *107*(2), 139-155.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*(1), 53-69.

Jacob, B. & Ludwig, J. (2008). *Improving outcomes for poor children*. NBER Working Paper, no. 14550.

Johnson, J., Brett, E. B., & Deary, I. J. (2010). The pivotal role of education in the association between ability and social class attainment: A look across three generations. *Intelligence,* 38, 55–65.

Justice, L. M., Petscher, Y., Schatschneider, C., & Mashburn, A. (2011). Peer effects in preschool classrooms: Is children's language growth associated with their classmates' skills? *Child Development*, *82*(6), 1768-1777.

Kerckhoff, A. (1993). *Diverging pathways: Social structure and career deflections*. Cambridge: Cambridge University Press.

Kim, J. S., & Quinn, D. M. (2013). The effects of summer reading on low-income children's literacy achievement from kindergarten to grade 8: A meta-analysis of classroom and home interventions. *Review of Educational Research*, 83(3), 386-431.

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, 75(1), 83-119.

Kulik, J. A. (2003). *Effects of using instructional technology in elementary and secondary schools: What controlled evaluation studies say*. SRI Project Number P10446.001. Arlington, VA: SRI International.

Lipsey, M. W., Landenberger, N. A., & Wilson, S. J. (2007). Effects of cognitive-behavioral programs for criminal offenders. *Campbell Systematic Reviews*. Retrieved from http://www.campbellcollaboration.org/lib/project/29/.

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

Lubbers, M. J., Snijders, T. A. B., & Van Der Werf, M. P. C. (2011). Dynamics of peer relationships across the first two years of junior high as a function of gender and changes in classroom composition. *Journal of Research on Adolescence*, *21*(2), 488-504.

Lumley, T. (2002). Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine*, 21(16), 2313-2324.

McMaster, N. K., & Fuchs, D. (2002). Effects of cooperative learning on the academic achievement of students with learning disabilities: An update of Tateyama-Sniezek's review. *Learning Disabilities Research & Practice*, *17*(2), 107–117.

Nisbett, R.E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence – New findings and theoretical developments. *American Psychologist*, 67(2), 130-159.

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.

OECD (2012). *Education at a Glance 2012: Highlights*. OECD Publishing. Retrieved from http://www.oecd-ilibrary.org/education/education-at-a-glance-2012_eag_highlights-2012-en

OECD (2013). *PISA 2012 Results: Excellence Through Equity – Giving Every Student the Chance to Succeed (Volume II)*. PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264201132-en.

OECD (2014). *Members and partners*. Retrieved from http://www.oecd.org/about/membersandpartners/

Perry Jr, W. G. (1999). *Forms of intellectual and ethical development in the college years: A scheme*. San Francisco, CA: Jossey-Bass Publishers.

Piaget, J. (2001). *The psychology of intelligence*. New York, NY: Routledge.

Pigott, T.D. (2009). Handling missing data. In Cooper, H. & Hedges, L. V. & Valentine, J. C. (eds.), *The Handbook of Research Synthesis and Meta-Analysis*, 399-416. New York: Russell Sage Foundation.

Pressley, M., Brown, R., El-Dinary, P. B., & Allferbach, P. (1995). The comprehension instruction that students need: Instruction fostering constructively responsive reading. *Learning Disabilities Research & Practice, 10*(4), 215-224.

Reisner, E., Petry, C., & Armitage, M. (1989*). A review of programs involving college students as tutors or mentors in grades K-12. Volume I*. Prepared for the U.S. Department of Education. Retrieved from http://www.policystudies.com/studies/?id=44

Reisner, E., Petry, C., & Armitage, M. (1990). *A review of programs involving college students as tutors or mentors in grades K-12. Volume II*. Prepared for the U.S. Department of Education. Retrieved from http://www.policystudies.com/studies/?id=44

Ritter, G., Albin, G., Barnett, J., Blankenship, V., & Denny, G. (2006). The effectiveness of volunteer tutoring programs: A systematic review. *Campbell Systematic Reviews*, 7. DOI: 10.4073/csr.2006.7. Retrieved from http://campbellcollaboration.org/lib/project/16/.

Robinson, D. R., Schofield, J. W., & Steers-Wentzell, K. L. (2005). Peer and cross-age tutoring in math: outcomes and their design implications. *Educational Psychology Review*, *17*(4), 327-362.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434), 473-489.

Sabates, R., Feinstein, L., & Shingal, A. (2013). Inequality in academic performance and juvenile convictions: An area-based analysis. *British Journal of Educational Studies*, *59*(2), 143-158.

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods, 8*(4), 448-467.

Scammaca, N. K., Roberts, G., Vaughn, S., & Stuebing, K. K. (2015). A meta-analysis of interventions for struggling readers in grades 4-12: 1980-2011. *Journal of Learning Disabilities*, 48(4), 369-390.

Schofield, J. (2010). International evidence on ability grouping with curriculum differentiation and the achievement gap in secondary schools. *The Teachers College Record*, *112*(5), 1492-1528.

Scott, M. A., & Bernhardt, A. (2000). *Pathways to educational attainment: Their effect on early career development* (IEE Brief, n28). Institute on Education and the Economy, Teacher's College, Columbia University. Retrieved from http://nrccte.education.louisville.edu/sites/default/files/publication-files/pathways_to_ed_attainment.pdf

Sirin, S.R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, 75(3), 417–453.

Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (2008). Effective reading programs for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly, 43*(3), 290-322.

Slavin, R. E., Lake, C., & Groff, C. (2009). Effective programs in middle and high school mathematics: A best-evidence synthesis. *Review of Educational Research*, 79(2), 839-911.

Slavin, R.E. & Madden, N. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness*, 4, 370-380.

Sternberg, R. J. (2009). *The essential Sternberg: Essays on intelligence, psychology, and education*. Springer Publishing Company.

Stoet, G., & Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within- and across-nation assessment of 10 years of PISA data. *PLoS ONE*, *8*(3).

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods*, 5(1), 13-30.

Timperley, H. S., & Phillips, G. (2003). Changing and sustaining teachers' expectations through professional development in literacy. *Teaching and Teacher Education*, 19, 627-641.

Tipton, E. (2014). Small sample adjustments for robust variance estimation with meta-regression. Forthcoming in *Psychological Methods*.

Tucker-Drob, E. M., Rhemtulla, M., Harden, K. P., Turkheimer, E. & Fask, D. (2011). Emergence of a gene x socioeconomic status interaction on infant mental ability between 10 months and 2 years. *Psychological Science*, 22(1), 125-133.

Tucker-Drob, E.M., Briley, D.A. & Harden, K.P. (2013). Genetic and environmental influences on cognition across development and context. *Current Directions in Psychological Science*, 22(5), 349-355.

UNESCO (1994). *The Salamanca statement and framework for action on special needs education*. Salamanca, Spain.

Van de Werfhorst, H. G., & Mijs, J. J. (2010). Achievement inequality and the institutional structure of educational systems: A comparative perspective. *Annual Review of Sociology*, *36*(1), 407-428.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard university press.

Wanzek, J., Vaughn, S., Wexler, J., Swanson, E. A., Edmonds, M., & Kim, A-H., 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), 528-543.

Wanzek, J., Vaughn, S., Scammacca, N., Metz, K., Murray, C., Roberts, G., & Danielson, L. (2013). Extensive reading interventions for students with reading difficulties after Grade 3. *Review of Educational Research*, 83, 163–195.

White, K.R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin*, 91(3), 461-481.

White, I. R., Barrett, J. K., Jackson, D., & Higgins, J. P. T. (2012). Consistency and inconsistency in network meta-analysis: Model estimation using multivariate meta-regression, *Research Synthesis Methods*, 3(2), 111-125.

Wilson, S. J., & Lipsey, M. W. (2006a). The effects of school-based social information processing interventions on aggressive behavior, part I: Universal programs. *Campbell Systematic Reviews*. Retrieved from http://www.campbellcollaboration.org/lib/project/14/.

Wilson, S. J. & Lipsey, M. W. (2006b). The effects of school-based social information processing interventions on aggressive behavior, part II: Selected/indicated pull-out programs. *Campbell Systematic Reviews*. Retrieved from http://www.campbellcollaboration.org/lib/project/15/.

Wilson, S., Lipsey, M. W., Tanner-Smith, E. E., Huang, C., & Steinka-Fry, K., (2010). Protocol: Dropout Prevention and Intervention Programs: Effects on School Completion and Dropout Among School-aged Children and Youth. Retrieved from http://www.campbellcollaboration.org/lib/project/158/.

Wilson, S., Tanner-Smith, E. E., Lipsey, M. W., Steinka-Fry, K., & Morrison, J. (2011). Dropout prevention and intervention programs: Effects on school completion and dropout among school-aged children and youth. *Campbell Systematic Reviews,* 7 Retrieved from http://www.campbellcollaboration.org/lib/project/158/.

Zief, S. G., Lauver, S., & Maynard, R. A. (2006). Impacts of after-school programs on student outcomes: A systematic review. *Campbell Systematic Reviews*. Retrieved from http://campbellcollaboration.org/lib/project/12/

SOURCES OF SUPPORT

SFI Campbell– The Danish National Centre for Social Research (SFI).

DECLARATIONS OF INTEREST

The authors have no vested interest in the results of this review.

REVIEW AUTHORS

Lead review author:

Name:	**Jens Dietrichson**
Title:	Ph. D., Economics
Affiliation:	SFI – The Danish National Centre for Social Research
Address:	Herluf Trolles Gade 11
City, State, Province or County:	Copenhagen K
Postal Code:	1052
Country:	Denmark
Phone:	+45 3369 7797
Email:	jsd@sfi.dk

Co-authors:

Name:	**Martin Bøg**
Title:	Ph.D., Economics
Affiliation:	SFI – The Danish National Centre for Social Research
Address:	Herluf Trolles Gade 11
City, State, Province or County:	Copenhagen K
Postal Code:	1052
Country:	Denmark
Phone:	+45 33697741
Email:	**mbg@sfi.dk**
Name:	**Trine Filges**
Title:	Ph.D., Economics
Affiliation:	SFI – The Danish National Centre for Social Research
Address:	Herluf Trolles Gade 11
City, State, Province or County:	Copenhagen K
Postal Code:	1052

Country:	Denmark
Phone:	+45 3348 0926
Email:	tif@sfi.dk

Name:	**Anne-Marie Klint Jørgensen**
Title:	Information specialist
Affiliation:	SFI – The Danish National Centre for Social Research
Address:	Herluf Trolles Gade 11
City, State, Province or County:	Copenhagen K
Postal Code:	1052
Country:	Denmark
Phone:	+45 3348 0868
Email:	amk@sfi.dk

ROLES AND RESPONSIBLITIES

- Content: Martin Bøg, Jens Dietrichson
- Systematic review methods: Martin Bøg, Trine Filges
- Statistical analysis: Martin Bøg, Jens Dietrichson, Trine Filges
- Information retrieval: Anne-Marie Klint Jørgensen

Jens Dietrichson holds a Ph.D. in economics from Lund University. Jens has research related experience of program evaluation and work experience of developing intervention programmes to improve the academic achievement of children-at-risk.

Martin Bøg holds a Ph.D. in economics from University College London. Martin has experience with systematic reviews and meta-analyses, and has contributed to several Campbell reviews.

Trine Filges holds a Ph.D. in economics from University of Copenhagen. Trine has extensive experience in systematic review methods and meta-analysis and has been a co-author of several other Campbell systematic reviews.

Anne-Marie Klint Jørgensen is educated at The Royal School of Library and Information Science. Anne-Marie is experienced in doing searches relating to Campbell systematic reviews and has broad competence and knowledge of literature searching and management across many subjects and many platforms.

ACKNOWLEDGEMENTS

Several additional persons have made very valuable inputs to the production of this protocol, including Anne-Sofie Due Knudsen, Misja Eiberg, Ulla Højmark Jensen, Majken Mosegaard Svendsen, and Christoffer Sonne-Schmidt.

EXPECTED TIMEFRAME

We plan to submit the first draft of this review during the second half of 2015, with the following approximate benchmarks:

- Training and pilot testing the inclusion criteria: winter 2015.
- Searches for eligible studies: winter 2015.
- Screening the results from the literature search: winter 2015-2016.
- Extraction of data from eligible research reports: winter and spring 2016.
- Statistical analysis: summer and fall 2016.
- Preparation of the final review report: fall 2016.

PLANS FOR UPDATING THE REVIEW

The lead reviewer will be responsible for updating the review approximately every 3-5 years.

AUTHOR DECLARATION

Authors' responsibilities

By completing this form, you accept responsibility for preparing, maintaining and updating the review in accordance with Campbell Collaboration policy. The Campbell Collaboration will provide as much support as possible to assist with the preparation of the review.

A draft review must be submitted to the relevant Coordinating Group within two years of protocol publication. If drafts are not submitted before the agreed deadlines, or if we are unable to contact you for an extended period, the relevant Coordinating Group has the right to de-register the title or transfer the title to alternative authors. The Coordinating Group also has the right to de-register or transfer the title if it does not meet the standards of the Coordinating Group and/or the Campbell Collaboration.

You accept responsibility for maintaining the review in light of new evidence, comments and criticisms, and other developments, and updating the review at least once every five years, or, if requested, transferring responsibility for maintaining the review to others as agreed with the Coordinating Group.

Publication in the Campbell Library and in Campbell Systematic Reviews

The support of the Coordinating Group in preparing your review is conditional upon your agreement to publish the protocol, finished review, and subsequent updates in the Campbell Library. The Campbell Collaboration places no restrictions on publication of the findings of a Campbell systematic review in a more abbreviated form as a journal article either before or after the publication of the monograph version in *Campbell Systematic Reviews*. Some journals, however, have restrictions that preclude publication of findings that have been, or will be, reported elsewhere and authors considering publication in such a journal should be aware of possible conflict with publication of the monograph version in *Campbell Systematic Reviews*. Publication in a journal after publication or in press status in *Campbell Systematic Reviews* should acknowledge the Campbell version and include a citation to it. Note that systematic reviews published in *Campbell Systematic Reviews* and co-registered with the Cochrane Collaboration may have additional requirements or restrictions for co-publication. Review authors accept responsibility for meeting any co-publication requirements.

I understand the commitment required to undertake a Campbell review, and agree to publish in the Campbell Library. Signed on behalf of the authors:

Form completed by: **Date:**

APPENDIX A – CRITERIA FOR SCREENING

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 1-4 are 'No'. If the answers to question 1-4 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 purpose to improve academic achievement and where academic goals are the primary focus of the intervention?*
   Yes – include
   Uncertain – include
   No – stop here and exclude

   Question guidance: Interventions should explicitly aim to improve academic achievement 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 result in improved academic performance or a higher skill level in a specific academic task.

2. *Are the participants in the intervention students in a regular primary or secondary school (grades K-12)?*[6]
   Yes – include
   Uncertain – include
   No – stop here and exclude

   Question guidance: A regular primary and secondary school 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. Furthermore, studies of preschool or other early childhood interventions should be excluded. Studies of interventions in tertiary education, such as universities, colleges, technical training institutes, community colleges, nursing schools, research laboratories, centres of excellence, and distance learning centres should also be excluded.

[6] We will screen for this review simultaneous with the screening for the parallel review regarding students in grades K-6 (for title registration see Eiberg, Due Knudsen, Sonne-Schmidt & Klint Jørgensen, 2014). In this simultaneous screening on title and abstract we will not separate studies with respect to focus on primary or secondary school. This separation will be done during the full text screening.

3. *Did the intervention take place in school during the regular school year in an OECD 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 intervention should be performed during the regular school year and in school(s), with schools being a stakeholder in the intervention. Interventions performed during e.g. summer or winter breaks should be excluded. If one part of the intervention is performed in school, and another outside of school, the intervention should be included.

4. *Is the study a primary impact study reporting quantitative outcomes published in or after 1980?*
   Yes – include
   Uncertain – include
   No – stop here and exclude

   Question guidance: The study should be primary research, reviews should be excluded. Purely qualitative research should also be excluded. The study should be published in or after the year 1980 to be included.

Second level screening based on full text:
Repeat, if necessary, questions 1 – 4 based on full text. Exclude the study if the answer is 'No' to one or more of these questions; otherwise continue with questions 5-7 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.

5. *Is the intervention aimed at raising academic achievement for individual students or groups of students that are categorized as having academic difficulties or being at-risk of such difficulties?*
   Yes – include
   Uncertain – include
   No – stop here and exclude

Question guidance: To be included, interventions should be targeting certain students and/or student groups identified in the study under consideration by their observed academic achievement (e.g., low academic test results, low grade point average or students with specific academic difficulties such as learning disabilities), or because they are deemed at-risk of academic difficulties on the basis of their educational, or social background (e.g., children from families with low socioeconomic status, children placed in care, students from diverse ethnic/cultural backgrounds, second language learners). Interventions targeting students with physical learning disabilities (e.g. blind students), students with dyslexia/dyscalculia, and interventions that are specifically directed towards students with a certain neuropsychiatric disorder (e.g. autism, ADHD) should be excluded. Interventions applied to improve the common learning environment at school level in order to raise academic performance of all students should be excluded, regardless of the characteristics of the student population.

6. *Does the study report outcomes of standardised tests in reading or mathematics?*
   Yes – include
   Uncertain – include
   No – stop here and exclude

   Question guidance: The primary outcome variables should be either standardised reading (e.g. vocabulary, comprehension) tests or standardised mathematics tests (e.g. mathematical problem-solving, arithmetic and numerical reasoning, grade level math), or both.

7. *Is the study a RCT, QRCT or QES with a control or a comparison group?*
   Yes – include
   Uncertain – include
   No – stop here and exclude

   Question guidance: Eligible types of control groups include waitlist controls and no-treatment controls. Eligible types of comparison groups include alternative treatments. Studies using single group pre-post comparison should be excluded. RCT: randomised controlled trials, including cluster randomised trials. QRCT: quasi-randomised controlled trials (i.e., participants are allocated by means such as alternate allocation, person's birth date, the date of the week or month, case number or alphabetical order). QES: quasi-experimental studies, such as e.g. matching designs, statistical controls, difference-in-differences, and regression discontinuity designs.

APPENDIX B – CODE BOOK

1. Report characteristics

1.1. Authors
1.2. Publishing status
1.3. Publication year
1.4. Outlet/Type of publication

2. Study characteristics

2.1. Study location (country)
2.2. Study design (RCT, QRCT, QES)
  2.2.1. Describe treatment assignment
2.3. Number of separate sites included in the study (classrooms, schools, districts)
  2.3.1. If multiple sites, describe if there were differences in assignment between sites?

3. Participant characteristics

3.1. Specify the target group of the intervention, e.g. students with specific learning difficulties, low achievement, low SES etc.
3.2. Gender (share of girls)
3.3. Age distribution (min, max, mean)
3.4. Grade distribution (min, max, mean)
3.5. Ethnicity/Cultural/Language background
3.6. Socioeconomic status (share low income, share low parental education, share low status parental occupation)

4. Intervention characteristics

4.1. Name of intervention
4.2. Instructional methods
  4.2.1. Describe the instruction methods used in the intervention (e.g. tutoring, cooperative learning etc), and any differences between treatment and control groups regarding these methods. State explicitly if there are no differences.
4.3. Content domain
  4.3.1. Describe the content domain targeted by the intervention (vocabulary, mathematical problem-solving), and any differences between treatment and control groups regarding the content they are instructed in. State explicitly if there are no differences.
4.4. Intervention site
  4.4.1. If not only in school, where?
4.5. How is the intervention delivered?
  4.5.1. Group size (e.g. 1:1, 1:2,...)
  4.5.2. Intervention implementer
  4.5.3. Is the implementer trained?
4.6. Duration of intervention in weeks (intended, received)
4.7. Frequency of intervention in sessions (intended, received)
4.8. Intended intensity of intervention in hours per week (intended, received)
4.9. Implementation quality (questions from Wilson, Lipsey, Tanner-Smith, Huang, & Steinka-Fry, 2010)
  4.9.1. Was the implementation of the program monitored by the author/researcher or program personnel to assess whether it was delivered as intended? (Yes/No/Cannot tell)
  4.9.2. Based on evidence or author acknowledgment, was there any uncontrolled variation or degradation in implementation or delivery of treatment, e.g., high dropouts, erratic attendance, treatment not delivered as intended, wide differences between settings or individual providers, etc.? Assume that there is no problem if one is not specified (yes (describe below)/ possible (describe below)/ no, apparently implemented as intended)
  4.9.3. Describe implementation problems, if any.

5. Control/comparison characteristics

5.1. What is the nature of the control/comparison condition?

- Controls do not receive any intervention/treatment/ service (*if yes, continue to section 6)*

- Wait-list controls (*if yes, continue to section 6)*
- Comparison intervention (*if yes, questions regarding participant characteristics and intervention characteristics should be answered for all treatments*)

6. Outcome measurement

6.1. Measurement timing
6.2. Name of standardised test (*repeat for all outcomes*)
6.3. Subject of standardised test (mathematics, reading, *repeat for all outcomes*)
6.4. Content domain(s) of test (e.g. vocabulary, algebra etc, *repeat for all outcomes*)
6.5. Number of outcome assessment periods (*repeat for all outcomes*)
6.6. Who performs the tests?

7. Sample size

7.1. Sample size used in analysis for outcome measurement (*repeat for all outcomes and groups*)

8. Outcomes

8.1. Outcome (*repeat for all outcomes and measurements*)
- Dichotomous outcome
- Continuous outcome
- High score / 1 is desirable
- High score / 1 is not desirable
- Numeric outcome (e.g. mean, beta-coefficient, F-test, t-test)
- Standard deviation (incl which groups the standard deviation is sourced from)
- Estimation method (e.g. raw means, adjusted means, regression adjusted etc)

APPENDIX C – RISK OF BIAS TOOL

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]		

[a] 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.[7] 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, 2008). 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).

[7] This risk of bias model was introduced by Prof. Reeves at a workshop on risk of bias in non-randomised studies at SFI Campbell, February 2011. The model is a further development of work carried out in the Cochrane Non-Randomised Studies Method Group (NRSMG).

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.

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] judgement) 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[8]	☐	Irrelevant	☐	☐
Other:	☐	☐	☐	☐

[8] See User guide for unobservables.

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-randomised 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 an intervention.

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.

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.