Protocol for the evaluation of the short- and medium-term effects of an early literacy tutoring intervention

Paul Bingley, Jens Dietrichson, Mette Friis-Hansen, Vibeke Myrup Jensen, & Julie Kaas Seerup

This version: 2023-01-25

1. Introduction

This protocol describes the evaluation of the short- and medium-term effects of a literacy program for kindergarten-age children in Denmark, called *Læseklar* (meaning reading-ready in Danish), and the evaluation of the implementation of the project. The project is funded by TrygFonden (grant nr 128499) and based at VIVE – The Danish Center for Social Science Research.

The evaluation will use a research design consisting of a school-level cluster-randomized trial and a randomized encouragement design. The design aims to answer the following four research questions (RQs):

1. What are the short- and medium-term effects of Læseklar on literacy skills and well-being?
2. Are there peer effects of Læseklar on literacy skills and well-being?
3. Are there effects of Læseklar on parental investments and attitudes toward school?
4. How cost-effective is Læseklar?

We plan to measure outcomes from the end of kindergarten up to 2 years after within the project period. Furthermore, the research design allows for longer-run follow-ups, which we describe in this protocol and will conduct, should funding become available.

Below we present the background for the evaluation and the program theory, including the target group, the program content and how we believe the intervention may affect students. Section 2 describes the primary research design, analyses statistical power, describes how we will estimate cost-effectiveness, as well as describes an alternative research design, which will be used if the primary design fails.

1.1. Background

Being able to read well is a fundamental skill. Yet, on average around 20 percent of 15-year-olds are not proficient readers in the OECD countries (OECD, 2016, 2019). Denmark does better than the average, but the proportion of not proficient readers is still a very substantial 15 percent (OECD, 2016; Christensen et al., 2019). At the same time, Denmark has one of the world's highest expenditures per child on pre- and primary school education (e.g., OECD 2017a, b). Finding cost-effective methods to help students with, or at risk of, reading difficulties should therefore be a high priority for researchers and educational policy makers.

This project evaluates an early literacy intervention in the form of a multi-sensory tutoring program, Læseklar, which targets the students most at risk of reading difficulties in the first year of primary school in Denmark (kindergarten or grade zero). The program combines components that have shown promising results in the international literature: one-to-one and small group tutoring, a focus on phonemic awareness, phonics, decoding, and multi-sensory learning methods.

1.2. Program theory

Special educator Anna Aldenius Isaksson, who is an advisor on the project, developed the original Swedish version of Læseklar ("Läsklar"; see lasklar.se and Bøg et al., 2021). Together with Isaksson, teachers and reading specialists in Hvidovre municipality developed the Danish version in a separate VIVE project (Seerup et al., 2022). Below, we first describe the target group and the characteristics of and likely causes of reading difficulties. We then describe how the intervention might work: first the content and the methods used in the program, and then how and why the content and methods may improve the skills of our target group. Lastly, we discuss short- and longer-run effects on the target group, and how peers and parents may be affected.

1.2.1. The target group and reasons for reading difficulties

Læseklar targets students with or at-risk of reading difficulties in the early years of school. Virtually all children need help to learn the alphabetic principle, i.e., the insight that the visual symbols of the writing system (graphemes) represent the sounds of the language (phonemes; Castles et al., 2018). However, for many reasons, some children may be at risk of reading difficulties and therefore need extra help. Examples include dyslexia, language and cognitive deficits, language minority and recent immigrant status (Bøg et al., 2021), and having attention (e.g., Dion, Roux, Landry, Fuchs, Wehby, & Dupéré, 2011), and working memory deficits (e.g., Follmer, 2018h; Jacob & Parkinson, 2015, Castles et al., 2018).

Regardless of the ultimate causes of reading difficulties, phonological awareness and letter-sound knowledge are the areas where children with reading difficulties are most often impaired (Share, 1995). Phonological awareness allows children to distinguish phonemes, the distinct units of sound that are the smallest building blocks of language, and combine them into larger units such as syllables. Phonics training gives children their knowledge of letters and the ability to combine letters or letter patterns with sounds. Explicit instruction in these areas may be especially important for children at risk of reading difficulties (e.g., Lundberg & Hoien, 1996). The hypothesis that interventions including instruction in phonological awareness, phonics, and decoding can improve the literacy skills of at-risk children has extensive support in the earlier literature (e.g., Dietrichson et al., 2021; Ehri et al., 2001; Snow et al., 1998; Snowling & Hulme, 2011).

Although Læseklar can be used in later grades, the Danish version of Læseklar has primarily been developed with kindergarten students in mind. Our intervention will be implemented in the second, i.e., spring semester of kindergarten. To get as close as possible to how the program would be implemented outside a research project, schools will play a major role in determining which students are at risk and will receive Læseklar. We describe the screening procedure we will use and the schools we will target for recruitment in section 2.

1.2.2. Program description and how the intervention might work

This section describes how the program will be implemented in participating schools randomized to the treatment group and how the intervention might work. We describe the randomization process and timeline in section 2.

Implementation. We will ask schools to start training with the program at the start of the spring semester in kindergarten. Results from an evaluation of the program in Sweden showed short-term improvements of a measure related to self-efficacy and that the intervention was implemented before motivational problems had become prevalent (Bøg et al., 2021). Implementing interventions before students start seeing themselves as “bad readers” may be an advantage, as the intervention then does not have to address demotivation. As children normally start primary school in the fall of the year they turn six both in Denmark and Sweden (Pontoppidan et al., 2018), we therefore believe that kindergarten is an opportune time to intervene.

Results from our Danish efficacy trial, in which the training started in the fall semester of kindergarten for the tested students, indicated that schools included some students, who became stable decoders without training with Læseklar (Seerup et al., 2022). Two related reasons for this inclusion may be that the test results at the start of kindergarten are somewhat noisy, and that students are not used to being in school and therefore seem more at risk of reading difficulties in the first months of the school year than they actually turn out to be. We will therefore ask schools to implement the program at the start of the spring semester. Most schools in the Swedish trial implemented the program in the spring semester of kindergarten.

The intended frequency and duration of the program is 3-4 times a week for 8-10 weeks, over a total of 30-35 sessions where each session lasts 10-15 minutes. However, this dosage can and should be adapted to the individual child (i.e., students with greater difficulties may need more sessions). Instruction is given one-to-one (for the students with the greatest difficulties) or in small-groups of 2-3 students by a tutor. In previous evaluations of the program, tutors have been regular teachers or special educators (Bøg et al., 2021; Seerup et al., 2022). In this project, teachers, special educators, pedagogues, and teacher’s aides may serve as tutors. The participating schools may choose the type of tutors that is best suited given their organizational context.

The tutors receive a manual and can see three short films illustrating the method. The films show a tutor with long experience of using the program with children with very few letter-knowledge skills. The tutors can contact the research-team for support at any time, and we will contact each school after the start of the training to check that the implementation is progressing according to plan.

Content and methods. The aim of Læseklar is threefold: First, to make the students realize that words denote both content (e.g., “lion” represents the animal) and has a form made up of a number of phonemes. Second, to help students grasp the connection between letters and sounds. Third, and most important, to show students how to decode simple words (i.e., to translate written letter combinations into sounds).

The program includes three types of sessions. In the first type, the tutor uses clay figurines representing 24 Danish language sounds and letters, a box with a compartment for each letter, and a laminated sheet with the fingerspelling alphabet. A session starts with the tutor selecting three figurines. The student’s (or students’, if the session is conducted in a small group) task is to help the figurine to move into its designated compartment, or “house”. To do so, the student needs to figure out a code. The first part of the code is the sound of the first letter in the figurine’s name, for example, /l/ in “lion”, and the second part consists of fingerspelling the letter. The student should make the /l/ sound, while at exactly the same time fingerspelling the letter. Then the student moves the figurine into its house, which is marked with the corresponding letter. After the three figurines have moved into their houses, the procedure is repeated once and then the session ends. During the next session, the student repeats the previously completed figurines, and then continues with a couple of new ones. For students experiencing difficulties remembering letter-sound combinations, tutors can use a mirror so that students see how their mouths look and feel when saying letter names and sounds. This multi-sensory method can also be used in the second and third type of session.

The second type of session starts after the student has learnt about 12-15 of the letter sounds, which typically takes less than a month. The tutor uses images representing short words, which are glued onto the back of the pages in a small booklet. The tutor acts as a secretary for the student, who sounds out the phonemes in the word represented by the image. As the student sounds out the phonemes, the tutor writes the word on the front of the page. If needed, the student uses fingerspelling to aid the memory of the letter-sound combinations. Using the booklet, the student can practice by themselves or with their parents, as it is possible to verify whether words are read correctly by simply turning the page and check the image.

The third type of session aims to build decoding speed and fluency by practicing the decoding of short syllables with the help of flashcards. The student's task is to decode as many syllables as possible during one minute. The tutor registers the result on a chart and then the same syllables are repeated in the same order two more times in the same session. This procedure usually yields a clear improvement. Making improvements visible for the students with the help of the chart usually motivates them to practice more. The second and third type of session are often used interchangeably, which decreases the risk that the student gets bored. It is also possible to go back to the first type of session, when words contain letter-sounds that were not among the first 12-15 that the students learned.

Conceptual framework: How the intervention might work. The first type of session aims to help students understand that the audio stream from the spoken language can be divided into a number of letter-sounds and thereby learn how to differentiate them from one another. Phonological awareness and letter-sound knowledge are important building blocks of children's literacy development (e.g., Bus, van IJzendoorn, & Pellegrini, 1995; Ehri et al., 2001; Snow et al., 1998; Snowling & Hulme, 2011), and are often seen as co-requisites for successful reading acquisition (Share, 1995). Moreover, Ehri et al., (2001) suggest that phonemic awareness training is more effective when letters are used in the training.

The second and third type of session aim to help students put the letter-sounds together to form words and to do so fluently, i.e., to build decoding skills. Decoding skills are often thought to be a necessary step towards reading comprehension (e.g., Hoover & Tunmer, 2018) and are difficult for children to learn on their own (Castles et al., 2018). Share (1995) argues that proficient decoding may act as a self-teaching mechanism. Learning to decode early may facilitate memory representation of words and increase exposure to print, and therefore aide the acquisition of vocabulary, which may help with comprehension (Blachman et al., 2014). Thus, by helping students decipher the alphabetic code early on, the program aims to provide an important stepping-stone to reading comprehension.

The small-group instruction allows for an individualized pace and support, frequent feedback from the tutor, and social interaction with a supportive "role model". These are features emphasized by pedagogical and psychological theories of learning (e.g., Bandura, 1986; Vygotsky, 1978; Hattie & Timperley, 2007). Furthermore, the instructional mode provides ample opportunities to practice the regulation of behavior, as well as concentration, attention, and working and short-term memory

(Seerup et al., 2022). These features are shared by most tutoring programs and may, at least partly, explain why tutoring is in general quite effective (e.g., Juel, 1996; Neitzel et al., 2020). Reviews comparing the effects of targeted (selected or indicated) interventions for at-risk students often find the largest effect sizes for intensive, small-group instruction by adults (e.g., Dietrichson et al., 2017, 2020, 2021; Fryer, 2017; Inns, Lake, Pellegrini, & Slavin, 2019; Slavin, Lake, Davis, & Madden, 2011).

The multi-sensory methods are more specific to Læseklar. Placing the figurines into their houses and using fingerspelling may activate the kinaesthetic, visual, tactile, and auditory senses. There are several possible reasons why such multi-sensory stimulus may strengthen memory capacity and retrieval. Working with the figurines creates a storyline about the figurines moving into their houses, which may activate the episodic memory (the memory of auto-biographical events; e.g., Tulving, 2002). Through the lens of dual-coding theory (Paivio, 1969; Clark & Paivio, 1991), multi-sensory stimulus allows information to enter the memory system through multiple processing channels. As the information can be more easily “chunked” into short-term memory, processing information through multiple channels may mitigate limitations on the capabilities of each individual channel (Shams & Seitz, 2008; Shams et al., 2011).

A related hypothesis is that multisensory stimulus may help students overcome limitations on their working memory (WM) capacity. In WM tests that ask subjects to hold a set of items (e.g., letters) available for some mental operation (e.g., reporting the letters after a delay), performance typically declines with increasing set size and when the subject concurrently engages in an unrelated distractor task. The decline is usually larger when the sets increase with, or the distractors consist of, similar items (e.g., Cocchini, Logie, Della Sala, MacPherson, & Baddeley, 2002; Conlin & Gathercole, 2006; Conlin, Gathercole, & Adams, 2005; Cowan, Saults, & Blume, 2014). Leading theories of why WM capacity is limited explain these phenomena by reduced interference between the memory representations of the items, or that dissimilar items draw resources from more than one WM domain and are therefore less likely to lead to the depletion of the resources (Oberauer et al., 2016).

In Læseklar, the process of moving figures into their houses may create, e.g., visual and tactile representations, i.e., *dissimilar* representations, that help students keep the two items of a letter and a sound in their WM long enough to connect the two (Bøg et al., 2021). Using fingerspelling may increase WM capacity so that students can simultaneously keep two or three letter-sound combinations in memory long enough to form words. Keeping two or three elements in memory may not sound like much but at the start of primary school students often have limited WM capacity (Demetriou, Spanoudis, & Shayer, 2014), and students at risk of reading difficulties tend to have lower WM capacity (e.g., Castles et al., 2018).

The program may also strengthen student motivation and self-efficacy. Helping the figurines find their houses may be more enjoyable and motivating than just thinking about how the beginning of a word sounds. Social interaction with a supportive role model may also influence motivation and self-efficacy (Juel, 1996). Lastly, by intervening already in kindergarten and first grade, that is, hopefully before reading difficulties become an entrenched part of the students’ view of their own ability (e.g., Butler, 1999; Muenks & Miele, 2017; Nicholls, 1978, Walgermo, Frijters, & Solheim, 2018), the development of motivational-emotional vulnerability may be avoided.

1.2.3. Potential effects

Short-term effects. Læseklar aims to improve phonological awareness, phonics, and, in turn, decoding skills. The overarching aim of the program is to help children decipher the alphabetic code and become able to decode simple words. A measure of decoding skills is therefore our most important outcome, but we also expect to find effects on related literacy outcomes and, later on, more general measures of reading skills (e.g., reading comprehension). However, it is important to note that while Læseklar may improve decoding skills and thereby mitigate the risk of reading difficulties, decoding should not be equated with the ability to read fluently or with reading comprehension. For students who have grave difficulties, it seems unlikely that Læseklar will eradicate these difficulties and such students may need more than just one early intervention. As we discuss below, successful early interventions may however make later interventions more successful (e.g, Stanovich, 1986; Cunha & Heckman, 2007).

There may also be effects on other types of skills and traits. As interventions may help at-risk students avoid long periods of failure, improving at-risk students' literacy skills may be a boost for their confidence and self-esteem (Toste et al., 2020). Reading difficulties have been linked to the development of motivational-emotional vulnerability (Stanovich, 1986; Poskiparta et al., 2003) and reduced self-regulation (Connor et al., 2016), and there is a strong association between the development of academic achievement and school motivation (Archambault et al., 2010; Taylor et al., 2014; Toste et al., 2020). The meta-analysis of Toste et al. (2020) furthermore indicates that reading achievement is a stronger predictor of motivation than vice versa.

It has been suggested that it is not until children are six or seven years old that they develop a normative concept of ability and start seeing outcomes as diagnostic of their current ability (e.g., Nicholls, 1978; Butler, 1999; Muenks & Miele, 2017). Furthermore, prior to formal schooling children are seldom confronted by situations where their achievements are systematically compared to achievements of their peers (Poskiparta et al., 2003), but already in the first year of school the poorest readers may develop a concept of themselves as weaker readers (Walgermo et al, 2018). To protect their self-image, students may reduce their reading efforts as reading is associated with poor performance on their part (ibid.). That is, to avoid thinking about themselves as weak readers, children avoid reading. Such a vicious circle may be difficult for children to break on their own (Poskiparta et al, 2003). Morgan et al. (2008) and Chapman et al. (2000) suggest that early experiences of failure in learning to read have a lasting impact on a child's self-beliefs and result in persistent weak reader self-concepts.

At-risk students in the first year of school are on the other hand highly motivated to learn how to read (Bøg et al., 2021; Walgermo et al., 2018). Thus, as motivation and self-concept are both crucial to student learning (Walgermo et al., 2018), academic interventions implemented before social comparisons have entrenched children's view of their own abilities may improve both academic outcomes, motivation, and self-efficacy.

Training and learning literacy skills may also affect other cognitive skills (e.g., Goldstein, 1976; Ritchie & Tucker-Drob, 2018; Seerup et al., 2022; Share, 1995), such as executive functions (EF). Core EF generally include inhibitory control, WM, and cognitive flexibility (e.g., Diamond, 2013). Inhibitory control is the control of one's attention, behavior, thoughts, and emotions. WM is the processes involved in the control, regulation, and active maintenance of task-relevant information in the service of complex cognition (Castles et al., 2018). Cognitive flexibility is the ability to change perspective and adjust in the face of changed circumstances (Diamond, 2013). Closely related concepts like self-regulation involves the application of all core executive function skills (McClelland & Cameron, 2012; Hernandez et al., 2018).

Recent reviews have found a substantial association between children's EFs and their academic achievements (e.g., Blair & Razza, 2007; Follmer, 2018; Jacob & Parkinson, 2015; Peng & Kievit, 2020). Theoretically, such effects may be expected for at least three reasons (Dietrichson et al., 2022): First, neurocognitive skills like EF develop when they are used (Zelazo, 2015), and learning to read involves the use of EF skills (Castles et al., 2018). Second, training academic skills may create new cognitive routines that are useful also for solving EF tasks (Gathercole et al., 2019). Third, learning academic skills increases domain-specific knowledge, which in turn may improve performance on EF tasks (Oberauer, 2018).

Longer-run effects. At some point, nearly all students learn the decoding skills targeted by Læseklar. If Læseklar simply shifts the learning of decoding to an earlier time-point without affecting the further acquisition of literacy skills, then medium- and long-term effects on the treated students seem less likely. However, there are reasons to believe that the further acquisition of literacy and other skills will be affected.

First, literacy skills may exhibit a "Matthew effect"; i.e., that skills beget skills (Stanovich, 1986). For example, being able to decode allows students to read on their own. As decoding may improve word recognition skills, such practice may improve word reading efficiency (Share, 1995). More efficient word reading may imply that students read more, which ought to improve vocabulary and language comprehension. In turn, reading comprehension may improve (Hoover & Tunmer, 2018). Better decoding skills may also make it easier for students to follow and benefit from the regular reading instruction in class, which could trigger a similar type of chain reaction.

Second, as suggested by the reciprocal association between academic achievement and school motivation (Archambault et al., 2010; Taylor et al., 2014; Toste et al., 2020), improving motivation, self-efficacy, and self-esteem may create longer-run effects. As students who cannot read after first grade are highly over-represented among poor readers throughout compulsory school (e.g., Denton, et al., 2004; Francis et al., 1996), early interventions may prevent students from experiencing long periods of school failure. Without earlier interventions, later interventions may therefore have to be more extensive to succeed (Cook et al., 2014; Vaughn et al., 2015). Both these lines of argument suggest that there may be dynamic complementarities, that is, later interventions may have larger effects, if they can build on successful earlier interventions (e.g., Cunha & Heckman, 2007; Knudsen et al., 2006).

Third, if the training and learning of literacy skills have lasting transfer effects on more general cognitive skills, such effects may help students' further learning. Gathercole et al. (2019) argue that training one set of cognitive skills (in their case WM) provides benefits to other skills when the training involves learning new cognitive routines that can be applied to novel or not-yet-learned tasks involving the other skill. Melby-Lervåg and Hulme (2010) provide an example of transfer of literacy training to memory skills. They find that second graders improve serial and free recall after training phoneme awareness and vocabulary, and argue that improved phonological and semantic memory representations of the words may explain the pattern of results. Literacy interventions furthermore increase domain-specific knowledge. Oberauer et al. (2018) cite evidence that knowledge from experience has substantial effects on the performance of WM tests. Manifestations of this effect include prior learning improving "chunking" (i.e., combining items into larger sets), that known words are easier to remember than unknown words, and that repetition improves performance.

Improved cognitive skills would be useful in all academic subjects, including the further learning of literacy skills.

Peer effects. Targeted interventions may not only affect the students receiving the interventions, peers in the same classrooms and schools may also be affected. Plausible mechanisms include teachers being able to keep instruction at a higher level and that improving the achievement of at-risk students may make them less disruptive, thus creating a better classroom climate and better performance overall (e.g., Lavy et al., 2011; Carrell et al., 2018; Berlinski et al., 2022). Furthermore, if the program prevents some targeted at-risk students from developing reading difficulties, special education resources that otherwise would be used to help targeted students could be used to help non-targeted students. As schools may allocate additional special education resources across grades, such effects may extend to peers in other grades than the targeted students.

The only study we are aware of that has estimated the peer effects of a targeted intervention similar to ours is a very recent study by Berlinski et al. (2022). They estimate the peer effects of a third grade tutoring intervention targeting the literacy skills of low-achieving students in Colombia. The intervention improved the literacy skills of the targeted students by 0.36 standard deviations, and with 0.11 standard deviations for their peers. Estimating a linear-in-means model of peer effects, they report that a one standard deviation increase of low-achieving students' literacy scores increases the scores of higher-achieving peers by 0.7 standard deviations. Peer-effects may also be long lasting (Carrell et al., 2018), motivating us to examine peer effects also at follow-up.

Parental time investments and attitudes. Parental behavior may also be affected by the intervention. Studies using natural or randomized experiments have found that parents offset school grants by lowering their own spending (Das et al., 2013), reduce their efforts when their children attend better schools (Pop-Eleches & Urquiola, 2013), decrease their time investments upon learning from school inspections that their school is of high quality (Greaves et al., 2021), and that parents choose lower quality schools after successful preschool interventions (Currie & Thomas, 2000; Gensowski et al., 2020). We therefore risk observing fade-out of intervention effects. It is also possible that Læseklar will lift students to a level where it is easier for parents help. In this case, the parents may increase their time investments and strengthen the response of the program. Either way, for a comprehensive examination of the program effects and cost-effectiveness, it will be important to capture the parental response to the program.

School responses. Schools may also change their behavior after the intervention, which may in turn influence the longer-run effects. Such side effects are expected because schools often promote the lowest achieving students more than higher achieving students (Schroeders et al., 2016; Bai et al., 2020). For example, schools assign students to special education services based on measures of academic progress (Raudenbush, 2001). Blachman et al. (2014) found that 43 percent in the treatment group had received special education services compared to 60 percent in the control group 10 years after a tutoring intervention that had large post-test effect sizes. If schools withdraw support after a successful intervention, fade-out should be expected. Thus, for the estimation and interpretation of longer-run effects it will be important to track how schools react to the intervention.

1.3. Evidence of short-term effects of Læseklar

Two RCTs of Læseklar have been conducted, one in Sweden and one in Denmark. Both involved members of the research team and both used a waiting list design in which individual students were randomized to receive the program in a first period or in a second period. By design, these two trials estimated short-term effects.

Bøg et al. (2021) included 161 students deemed to be most at risk of reading difficulties in kindergarten and first grade in 12 Swedish schools. The treatment group showed large and statistically significant improvements compared to the control group on the two pre-registered primary outcome measures: a standardized test of decoding (Hedges' g = 1.07) and a standardized test of letter knowledge (g = 1.03). The improvements were also statistically significant on measures of phonological awareness (g = 0.56) and self-efficacy (g = 0.57), but not on measures of enjoyment and motivation. The program appeared cost-effective relative to other tutoring programs in the short-term.

In the first phase of this project, we conducted an efficacy trial of the Danish version of Læseklar in 2019/2020 (Seerup et al., 2022). The trial originally included 13 schools and we randomized 117 students to a treatment and a waiting list control group. However, our post-testing was interrupted by the COVID-19 induced lockdown of schools in Denmark. The interruption meant that we could only obtain post-tests for 55-60% of the students (depending on the test).

The smaller than planned sample size made the results uncertain and no individual effect estimate was statistically significant. Our quantitative analysis indicates beneficial but smaller than expected effects on our two primary outcomes, decoding and letter knowledge. We found an effect size of 0.15 on the decoding test and an effect size of 0.35 on the letter knowledge test. The effect sizes were relatively large on some secondary outcomes: tests of impulse control (0.52), phonological awareness (0.32), and consistency of attention (0.32). The effect sizes on tests of WM, verbal memory, and inattention were close to zero. Tests conducted by the schools themselves at the end of kindergarten, after the lockdown ended and after some schools had begun to give the program to waiting list students, indicated relatively high average decoding scores among the participating students. We also found suggestive evidence that students who trained more with the program were less likely to be classified as at risk of reading difficulties afterwards, and that schools that implemented the program well showed better results.

We followed the implementation of the Danish efficacy trial closely. Interviews indicate that teachers were unanimous in their appreciation of the logic and pedagogy of Læseklar. However, observations and "diaries" filled out by the teachers implementing the program revealed some implementation problems. Based on both the quantitative and qualitative analyses, our best, not mutually exclusive, explanations for the smaller effects in Denmark compared to Sweden are:

- While our full sample is balanced on all pre-tests, there are imbalances in favor of the control group in the reduced sample tested before the lockdown. These imbalances may have contributed to the smaller effects.
- Implementation problems in some schools implied that some treatment group students did not receive the full program. In particular, a relatively large proportion did not receive sufficiently many decoding training sessions (type 2 and 3 sessions), which may explain the smaller than expected effects on this test.
- Although the effect on the letter knowledge test is not small, it is smaller than in the Swedish trial. One explanation seem to be that the Danish schools focused more on letter knowledge than the Swedish schools and that the Danish children knew more letters before the

intervention. Many children know nearly all letters at the post-test. The kind of drastic improvements seen in the Swedish trial were therefore harder to achieve.

We discuss how we will mitigate risks of the first two problems in this trial in section 2.1.5. The last explanation however implies that the post-test effects may be smaller in Denmark, even if we solve the first two problems. We discuss the statistical power of the trial in section 2.4.

2. Research design

This section describes the research design we will employ to answer RQ1-4, and the implementation analysis. We first present an overview of the design and then discuss ethical considerations, the steps in the research design and program implementation. Thereafter, we describe the outcome measures, the analysis plan, statistical power calculations, cost-effectiveness calculations, and an alternative design, which we will use if the primary design fails.

2.1. The research design

Our primary objective within the project period is to estimate the short- and medium-term effects of Læseklar. Our design will also provide the set-up for the evaluation of long-term effects, beyond the project period.

Estimating the effects of school interventions is subject to several threats to the internal validity of the estimates, which are specific to or worsened when the aim is to estimate the medium- and long-term effects. Our research design randomizes schools to treatments, which will mitigate selection bias, as the treatment assignment is independent of school characteristics. As treatment and control groups are assigned simultaneously and are affected by the same temporal effects, this design also mitigates temporal and regression to the mean effects (Torgerson et al., 2011). Other threats to the internal validity of the effect estimates are not necessarily mitigated by randomization alone. These threats include within-school selection of students, cross-overs and spill-overs between the treatment and control groups, blinding of participants and testers, students and schools leaving the sample (i.e., attrition), Hawthorne and John Henry-effects, and, related to all of these problems, obtaining adequate statistical power. We explain how our design mitigates these threats when we discuss the mitigating features below.

We will combine a cluster-randomized encouragement design with a cluster-randomized controlled trial (cluster-RCT) in which schools are randomly assigned to either the treatment or the control group. The steps of the design are described below and further elaborated in the coming sections:

1. Ethical considerations
2. Determine the set of eligible schools.
3. Random invitation procedure
4. Assignment of schools to the treatment and control group
5. Program implementation

2.1.1. Ethical considerations

The design implies that some schools will not receive the program from the project. This may be an ethical problem. However, note that we will not deny any student help. Schools that do not receive the program can and should go on helping their at-risk students as they usually do. We will just not subsidize Læseklar in all schools. Furthermore, the project does not have resources to provide all schools in Denmark with the program. Therefore, we believe that first prioritizing a set of schools that probably have more students that need extra help (the eligible schools, see next section) and then randomly determining which among the eligible schools that we will invite to participate is a sound ethical procedure.

Among the schools that accept the invitation to participate (participating schools), we will randomly assign one group to receive the program from the project (treatment schools) and one group that will not (control schools). We will compensate control schools for participating by a payment of 1000 DKK per class and by giving the control schools an offer to buy the program at the discounted price the project pays, once the post-tests have been conducted. Hence, all participating schools will get something extra, which we believe will be beneficial to them. Therefore, as we expect all participating schools and students to be better off because they are participating in the project, and no school or student to be worse off than they would have been, had there not been a project, we believe that the suggested research design is ethically sound.

We will inform all parents in participating schools about the intervention beforehand and all participation in the project will be voluntary. That is, parents will have the opportunity to decline participation for their children (including being tested), and no child will be tested against their will. We have received an approval from VIVEs internal review board (nr 2022/4 dated June 22, 2022).

2.1.2. Eligible schools

The target group for Læseklar is the students most at risk of reading difficulties in kindergarten. As participating schools will have to invest personnel time in the project, we prioritized inviting schools that were likely to have many at-risk students. However, the exact number of at-risk students in each school could not be known until schools were recruited and the pre-tests were conducted. To determine the eligible set of schools, we used data on educational characteristics of parents that are highly correlated with the national reading test in 2nd grade. Specifically, for the first round of invitations, we selected the schools in Denmark in which no more than half of the parents have a tertiary education (using data from the kindergarten classes in the school year 2019/2020). This cutoff yielded an eligible sample of 394 public schools (out of 1,135 for which we could obtain sufficient information). Our targeted number of participating schools is 80 (see section 2.4. for a statistical power analysis and motivation of this number).

2.1.3. Randomization of invitations and invitation procedure

Among eligible schools, we used the random number generator in Stata to assign each school a ranking number. The randomization was conducted by members of the research team who were blind to school identity. We randomly ranked the 394 schools and this rank determined the order we invited the schools. Number 1-300 ended on the invitation list while the final 94 schools were listed as 100% untreated. Thus, we did not invite them.

For administrative reasons, we will recruit over at least two school years. In spring 2022, we started inviting schools with ranking numbers 1-80 (April 2022), and added ranking number 81-120 by June 2022 as we had a low response rate to our invitation. In total, 16 schools agreed to participate and were randomized to the treatment and control groups.

In the second year, we will invite schools with rank 121-300 and schools that declined the invitation in the first year. Depending on the response rate, we may also have to increase the number of eligible schools. It is also possible that the response rate is higher than expected. As our budget determines the number of schools we can give the program to, should we end up with more schools than the targeted 80 we may choose to randomly assign less than 50% of participating schools to the treatment group.

We invite schools by e-mailing the school manager responsible for kindergarten (typically, the school principal) and, if needed, conduct follow-up phone contacts, stating clearly that participation is conditional on schools accepting to end up in the control group. As mentioned in section 2.1.2., the control schools will receive a compensation of 1,000 DKK per class for being in the control group, and may buy the program at a discounted price after the post-test. These features ought to decrease the risk that control schools dropping out upon learning that they did not become treatment schools.

We conducted the randomization for the first round and notified the schools of their treatment status mid-November. We will use a similar procedure in 2023. The treatment schools are expected to implement the program in the spring semester of 2023 or of 2024, which will give schools at least two months to prepare the implementation (and usually more, as they do not have to start early January). That is, regardless of when their school was recruited, students in treatment schools will train with the program in the second semester of kindergarten.

2.1.4. Assignment of schools to the treatment and control group

Among the schools that accepted the invitation to participate in 2022, we randomized the treatment (receipt of Læseklar) at the school level. We conducted a stratified randomization to increase the chances of obtaining a balanced sample and increase statistical power (Athey & Imbens, 2017).[1] As the recruitment procedure stretches over at least two school years and schools recruited the first year will receive the intervention before the recruitment in the second year begins, schools were stratified by recruitment year. Within each recruitment-year strata, we used a pair-matching procedure similar to the one suggested by Bai (2022) for non-clustered experiments. We ranked schools by the proportion of students at risk of reading difficulties according to test scores of literacy and language skills, and then matched schools with adjacent ranks into pairs. The pre-test scores comes from language and literacy tests conducted by all Danish schools at the beginning of the first semester of kindergarten (typically in September or October). Many Danish schools use the same test (Socialstyrelsen, 2017). However, two recruited schools used another test and we therefore used scores standardized by each tests' normed sample to make test scores comparable across schools. We then used the random number generator in Stata to determine the assignment to the treatment and control group in each pair. The assignment probability was 50% for each school. We will use the same procedure for schools recruited in 2023.

[1] In terms of reducing the risk of imbalances and increasing power, a minimization procedure (e.g., Torgerson & Torgerson, 2007; Kasy, 2016), or a combined minimization-randomization procedure (Scott, McPherson, Ramsay, & Campbell, 2002; Xu, Proschan, Lee, 2016), may be preferable to stratified randomizations. However, our number of randomization units is larger than the 50 units for which Torgerson & Torgerson (2007) suggest minimization would be most useful, and the properties of frequentist inference procedures seem less well developed for minimization than for pair-matching/stratification designs (Xu, Proschan, & Lee, 2016).

The main advantage of a school-level randomization compared to both class- or student-evel randomizations is the lower risk of treatment cross-overs and spill-overs to the control group. Schools may of course want to adopt a program they hear about and perceive is successful but the risk should be greatly reduced compared to a situation where schools would get direct access to such a program through the project, which they would if randomization was conducted within schools. In our case, the between-school spillover risk is very low. The program we are evaluating is not available on the open market in Denmark and schools that received the program material through our efficacy trial are not allowed to spread the program to other schools.

Control schools may on the other hand decide to use other programs instead of Læseklar. We will examine whether this is the case. Whether adoption of other programs is a major problem for our evaluation or not, depends on the reasons why schools adopt other programs. Our objective is to evaluate the effect of Læseklar in comparison to treatment as usual in the eligible Danish schools. Therefore, if schools adopt the program they would have adopted, if there had not been a study, such adoption is part of treatment as usual and not a problem for our evaluation. On the other hand, if schools adopt another program *because* they became part of our study, then we are not comparing Læseklar to treatment as usual. This discussion points to a more general problem of treatment and control groups being affected not only by the intervention but by the act of participating in a research study itself (e.g., Hawthorne and John Henry-effects; Glennerster & Takavasha, 2013). The solution of blinding treatment and control schools to treatment status is not feasible in our case as we need to inform schools, parents, and children about the study. Instead, we will examine this issue by interviewing/surveying control schools about whether they adopted other programs because they are a part of the study. We may also use the randomized invitation to schools to examine this issue.

The main disadvantage of school-level randomization compared to a lower level randomization is the reduced statistical power. We describe below how our research design seeks to recruit and retain enough schools to ensure adequate power. A further complication in our case is that we, as mentioned, will make the program available to control schools at a reduced price after the first year. However, as we promote the program to target kindergarten, it is less likely that the program will be used in first grade. Nonetheless, as first grade is after the post-test but before the first follow-up, we will follow up on the control-schools that choose to buy the program and ask them about which grades have used the program. That control students may receive the program after the post-test also complicates the analysis of the effects at follow-up. Despite this complication, we believe the advantages in terms of recruitment and retention of offering control schools the subsidized program after the post-test clearly outweigh the disadvantages. We describe our analysis plan for the follow-up outcomes in section 2.3.

What stratification procedure to use is not obvious and depends on considerations about for example attrition and statistical power (see e.g., McKenzie, 2022). In particular, attrition at the randomization unit level (schools in our case) may reduce the power advantages of pair-matching designs (but not necessarily, see Chaisemartin & Ramirez-Cuellar, 2022). We expect little to no school-level attrition. For the primary outcomes at follow-up, we will obtain data from administrative registers to which schools have to report. Regarding the post-test, all control schools have incentives to stay in the experiment until the post-test (see section 2.2) and the only thing they have to do obtain the participation payment is to allow their students to be tested, which is handled by research staff and is a pre-requisite for participating. Therefore, we believe the potentially greater statistical power of pair-matching is more important in our setting (e.g., Bruhn & McKenzie, 2009, Imai, King, & Nall, 2009, and Bai, 2022, present theoretical and empirical evidence in favor of pair-matching).

2.1.5. Program implementation

Implementation starts the fall of 2022. Schools randomized to the treatment group received the program material after randomization. We recommend that schools start training with the program at the beginning of the spring semester, and no later than early March so that all targeted at-risk students have time to receive the full program and can be post-tested before the end of kindergarten.

In the efficacy trial, the students we could post-test started training already in the first semester of kindergarten. As the teachers had relatively short time to observe the children when making the selection, the early start may be one explanation of why some children in the control group learned to decode quite a few words before receiving the program. As teachers have had more time to observe the students in school and know more about their literacy skills, a later start date should make the selection of the at-risk students less noisy. In addition, starting later will give schools more time to prepare the organization of the training and tutors will have more time to learn how to deliver the program.

Selection of students. In early January, we ask treatment schools to select the students they plan to train with the program. For control schools we use the results from the pre-test to define at risk students and by the end of the school year, we will ask the teachers if some students had any form of special support. This post-randomization selection of students would risk introducing bias, if the effects were estimated on the sample of the selected students in treatment and control schools. However, our primary analytical strategy will not use this sample (we discuss the consequences for our analytical strategy of this selection further below). Asking schools to select which students that will/would have received the program has the major advantage that when schools will implement the program outside of a research study, this is how the selection will be conducted. We believe mimicking this implementation process as closely as possible is a key factor for external validity.

Selected students should be the ones considered most at-risk of reading difficulties in their respective schools. We will not restrict the group of eligible students further than that the reason for their difficulty cannot be something that invalidates the methods used in the program. Certain parts of the program would for example be difficult to use with blind or deaf students. Furthermore, students should not be able to decode when the training starts, and should have difficulties with phonological awareness and learning letter-sound combinations. If schools are uncertain of which students who should receive the program, we will provide support and recommendations of tests they can use to determine which students that are most in need of help. Such support is also available for schools from the program developer when they buy the program outside of a research study.

We will recommend that schools train with 2-3 students per class. The motivation for this number is the experiences from the efficacy trial: working with many students increased the risk of implementation problems, made it more difficult to replace tutors when teachers were on sick leave, and increased the risk of including students, who learned to read without extra tutoring.

Tutors. Personnel in each school will act as tutors. Schools can choose which personnel group and persons that conducts the tutoring, as long as the persons have experience of working with students. We expect that teachers, pedagogues, and special educators will be the most common tutors, but also teacher's aides may be tutors.

A problem we encountered in the efficacy trial was that schools did not provide back-up for absent teachers. We will therefore recommend that schools just extend the training period in shorter periods of teacher absence and recommend that the schools train extra tutors. For example, reading specialists ("læsevejledere"), who typically work across classrooms could either back up teachers, or if the reading specialists are implementing the program to start with, teachers or teacher's aides can back up the reading specialists. This set-up will help to ensure that students most at risk of reading difficulties receive the program and that schools have the resources needed to implement the program well.

Sessions. Schools may organize the sessions in a way that is best suited to their local conditions. A common choice in the efficacy trial was to conduct Læseklar-sessions during lessons where there are two adults present in the classroom (e.g., a teacher and a pedagogue) or when all classes have joint lessons and there are several teachers present. Then a teacher or pedagogue conducts the Læseklar-session, while the others conduct the regular lessons.

Program material. We described the program material and conceptual framework more in detail in section 1.2.2. In short, tutors receive a manual and three illustrative videos. The tutors are also invited to a start-up webinar and they can contact VIVE personnel for support. Our experience from the Swedish and Danish efficacy trial is that requests for more extensive method support during implementation are rare and can, in case they arise be handled over the phone.

We will offer schools the program material, the tutoring training course, support for tutors, and the testing without cost. As we believe that it is important for the longevity of the program that schools handle their internal implementation and organization processes themselves, all other expenditures, including the implementation time, will be borne by the schools themselves. As mentioned, we will offer control schools the program for purchase at a subsidized price. We believe this is an important recruitment advantage as well as a way of avoiding resentful demoralization, which occurs when the control group is dissatisfied because they did not receive an intervention (Torgerson et al., 2011).

Support. Our design is an effectiveness trial and we therefore want to get as close as possible to how schools would implement the program outside of a research study. Monitoring and support will therefore be kept to a minimum. However, based on the observations and interviews conducted with the tutors during and after the efficacy trial (Seerup, 2022), we will provide the opportunity to attend a start-up webinar and administer a follow-up contact by e-mail/telephone about 3-4 weeks after schools have started the training. In particular, this follow-up will emphasize that the transition from training letter-sound correspondences (session type 1) to training decoding (session type 2) should be made early, which not all schools did in the efficacy trial. This type of support could, according to the program developer, be a standard procedure of program implementation, should the program be effective.

Implementation fidelity. Examining implementation fidelity will be important in the short- and medium-term. One motivation is that fidelity data will help us interpret the effects from the RCT, including how control schools react to being in the control group. Lack of effects may be a sign of lacking fidelity, not a sign of an ineffective program.

Regarding the implementation fidelity, we will ask tutors to keep a "diary" of the sessions for each student. We developed and used such diaries in the efficacy trial, which provided useful information about both the number and type of sessions for each student.

We will inquire if control group students were given other interventions or extra support services, as this information is crucial for the interpretation of the effects. If control students also receive an intervention, we expect smaller effects than if they do not.

2.2. Pretest, posttest and follow-up tests

We will test students using pre-tests conducted before randomization, post-tests conducted at the end of kindergarten, and follow-up tests at different intervals after the end of the intervention. We describe these in turn below.

2.2.1. Pre-tests

Since 2009, it has been compulsory for Danish schools to screen their students at the start of kindergarten. As pre-test we primarily use the national pre-school reading test for kindergarten. The test is developed by Danish and international researchers (Bleses, Jensen, Makransky, Dale, Højen and Vach) and inspired by several existing language proficiency tests such as the MacArthur-Bates Communicative Development Inventory test (Bleses et al., 2017). The Danish proficiency test consists of three to seven different sub-tests such as word knowledge, language comprehension, rhymes, and letter knowledge. These sub-tests all reflect important communicative development skills related to reading later in life (Ministry of Education, 2019). For each child and sub-test, a total score is calculated on the basis of these questions and this total score is age and gender adjusted, i.e., compared to a large random sample of children of same gender and same age (in months).

We expect that most schools will use this test, but we will accept schools as participants as long as they currently use or are willing to use a norm-referenced test. In that way, we can standardize the pre-test score against the norm-referenced sample and use this standardized score in the analysis. We will use the pre-literacy and language subscales from the national pre-school reading test for kindergarten, and similar scales from other tests. As mentioned, we will collect the results of pre-tests from all students in the participating schools before randomization.

In the first recruitment and randomization round (fall 2022), 14 schools used the national pre-school reading test for kindergarten and two schools used three norm-referenced subscales from Læsevaluering på begyndertrinnet (Borstrøm & Klint Petersen, n.d.): letter knowledge, phonological awareness, and vocabulary. The first two are measures of pre-literacy, and the last includes partly similar content to the word knowledge sub-test in the national pre-school reading test for kindergarten.

2.2.2. Post-tests

Our primary post-intervention outcome at the end of kindergarten will be a measure of decoding skills from a subtest of a standardized dyslexia test widely used in Denmark (the 30 real words in “Elbros ordlister” from “Ordblinderisikotesten”; Gellert & Elbro, 2016). The test score is the number of correctly read words (i.e., a number between 0 and 30). All kindergarten students in treatment and control schools will be tested on this test, which takes about five to eight minutes to conduct per student. The post-tests are conducted by testers hired by the project. We will strive to keep testers blind to the treatment status of the children. This procedure mitigates ascertainment bias, which could occur when those testing and marking are aware of the group allocation of the students from which the tests originate. However, we know from experiences in other projects that it may be difficult to keep schools from revealing their treatment status.

We will furthermore use subscales from the national survey of student well-being as secondary outcome measures. All students in public schools take this survey annually from Kindergarten to 3rd grade (Ministry of Education, 2018b).

2.2.3. Follow-up tests

We will use the compulsory Danish national tests in reading conducted in the first semester of second grade as our primary outcome at the first follow-up test. This test is currently mandatory for all public schools. Further compulsory national tests in reading are conducted in 3rd, 4th, 6th, and 8th grade. We will not be able to analyze the results from these tests within the project period. However, the tests are constructed in the same way as the one in the second grade, and they can in principle be analysed in a similar fashion as the test in second grade. That is, the description and the analytical strategy presented below is valid also for them. Should we have the resources, we will proceed to analyze the later national tests, as well as other longer-run outcomes available in the Danish administrative registers when this information becomes available. We do not describe these register variables more in this protocol, but they include for example measures of educational attainment, earnings, and health care consumption.

The national tests consist of three subtests: decoding, text comprehension, and language comprehension but only the final score is reported. This score is our primary follow-up outcome measure. The tests are taken on and corrected by a computer, thus making both testing and marking blind to treatment status. The test takes approximately 45 minutes. The first five minutes is a hyphenation test and the last 40 minutes are 60 questions related to language proficiency.

The compulsory nature of the national test increases the possibilities to track students who move from participating to non-participating schools and therefore may decrease attrition bias, which is a great advantage for long-run studies. However, there are valid reasons for students not taking the tests, and in practice, not all students do (on average 91% did in 2019/2020; Ministry of Education, 2020). We will use other register data to examine if students that do not take the tests differ from those that do and whether there is differential attrition between the treatment and control group, and between invited and non-invited schools.

We will continue to use subscales from the national survey of student well-being as secondary outcome measures at follow-up.

Parental compensational effects are common, and a potential threat to finding long-run effects of Læseklar. At follow-up, we plan to examine parental investments in children's reading-skills through two channels. First, we will use the national register for library use. In this register, all borrowed books from public libraries are registered and can be linked to the individual parent including other information such as theme (science) and age-categories (children books versus adult books). Second, we will, if possible, survey parents in the treatment and control schools. We will use questions about how much time parents (or older siblings) spend in a regular week helping children with homework and reading to or with their child. We will ask separate questions about reading homework and other subjects to capture substitution of parental time investments between subjects. We will also ask questions about parental attitudes to school. As much as possible, we will re-use questions used in other research projects about primary school. For example, we may use surveys of the consequences of the 2014 national reform (Arendt et al. 2017) to allow for a comparison with a representative sample of schools.

2.3. Analysis plan

This section presents the strategy we will use to estimate the effects at the post-test, conducted at the end of kindergarten, and the follow-up tests. We discuss issues with multiple hypothesis testing, which affect both the post and follow-up estimation, in section 2.3.3.

2.3.1. Estimation of post-test effects

We will use the stratified randomization of participating schools to the treatment and control groups to estimate the effects of the program. On the school level, the randomization will, in expectation, make the assignment of treatment unrelated to school and student characteristics. A comparison of school means would thus be unbiased. However, to increase statistical power and because we are interested in estimating the effects for different groups of students, we will use student level data in the estimations. As discussed in section 2.1.5., the selection of students that receive treatment within treatment schools is not random, and the students the control schools would have selected, had they received treatment, cannot be directly observed. We will therefore rank students by their pre-test scores within each school, and then run regressions based on equation (1) for different quantiles. For our primary analysis of the effects on low-achieving students, we will run the following regression for students with the 20% lowest pre-test scores:

$$y_{i,post} = \alpha + \beta D_i + X_i\gamma + \varepsilon_i \qquad (1)$$

where $y_{i,post}$ is the outcome variable for student i, D_i is an indicator equal to 1 for all students in treatment schools and 0 for students in control schools, X_i is a vector of the pre-test scores, α is an intercept, β is the effect estimate, and ε is an error term. As D_i is based on the assignment to treatment, not on actually implementing the program, this strategy estimates an intention-to-treat (ITT) estimate of the effect. We will include all students with observations on $y_{i,post}$, and impute missing pre-test scores. We will also show results from regressions without covariates.[2]

We will follow the recommendations in Chaisemartin & Ramirez-Cuellar (2022) for pair-matched and clustered randomizations and cluster the standard errors on the pair-level. Regular asymptotic procedures may be sensitive if the number of clusters is small or if clusters have different sizes (e.g., Cameron & Miller, 2015; Mackinnon & Webb, 2017), and may either overestimate or underestimate the standard errors in pair-matched designs (Bai, Romano, & Shaikh, 2019; de Chaisemartin & Ramirez-Cuellar, 2022). de Chaisemartin & Ramirez-Cuella (2022) found that their procedure produced less reliable results with fewer than 20 pairs. Although our number of pairs will be higher, we will, as recommended by e.g., Athey & Imbens (2017), Bind and Rubin (2020), and Young (2018), also report p-values from randomization tests.

Our primary interest at post-test is the students with the 20% lowest pre-test scores in each school, of which all or a large proportion of students will receive Læseklar. Similar regressions can be run also for higher-achieving students, and any effect on higher-achieving students will primarily be interpreted as peer effects (peer effects may however take longer to fully develop). As we will use samples fully determined by the pre-test, this strategy avoids biasing the results from using the

[2] As described in section 2.1.4., we intend to conduct a balanced randomization with equally many treatment and control schools. As including covariates when analyzing imbalanced randomizations may create bias (Freedman, 2008), should there be imbalances we will include interactions between the treatment indicator and covariates as suggested by Lin (2013) to test if our results are sensitive to this problem.

potentially endogenous selection of students to treatment within treatment schools. The downside is that we do not get an estimate for the students actually receiving the program.

We will use instrumental variables (IV) strategies to obtain estimates of the effects on students receiving Læseklar and the peer effects. We will use the random assignment to treatment and control schools and the pre-test scores to create instruments for actually receiving the program and having a peer who received the program. The parameter of interest represents the local average treatment effect (LATE) of receiving Læseklar in the subsample of compliers (Imbens & Angrist, 1994). That is, the students that would have received Læseklar, had their school been randomized to treatment. This subsample is not necessarily directly observable but it is, in our case, a highly relevant subsample (see e.g., Marbach & Hangartner, 2020, for methods of describing compliers and noncompliers). If no one in the control group receives the program, then the LATE is equal to the average treatment effect on the treated (ATET; e.g., Abadie & Cattaneo, 2018). To evaluate more directly if there are significant peer effects, we will use the randomization and two rank variables to develop two instruments for the two potentially endogenous variables that indicate students in treatment schools who received and did not receive the program, respectively. For example, by interacting the treatment indicator with an indicator for being in the lowest 20% of the achievement distribution and an indicator for being in the upper 80% we would get two instruments for the two endogenous variables. As the pre-tests and indicators are non-random, we will use the methods described by Borusyak and Hull (2021) to account for the non-random exposure to the random treatment assignment.

Regular two-stage least squares (2SLS) estimates are biased in finite samples. However, Andrews and Armstrong (2017) showed that unbiased estimators can be constructed given that the sign of the coefficient on the instrument in the first stage regression is known. As the relation to uptake of the program is evident for both the treatment indicator and the rank in our case, we will use their method to obtain an unbiased estimate.

Additional secondary analyses may include examining heterogeneity by distribution regressions (e.g., Chernozhukov et al., 2013), and using similar methods as Berlinski et al. (2022) to estimate the linear-in-means peer effects.

2.3.2. Estimation of follow-up effects

At the follow-up in second grade (and at potential later follow-ups outside of the project period), we will conduct a similar analysis as the one described in section 2.3.1. The primary outcome is the test score from the national test in reading conducted in the first semester of second grade. We may also leverage the full sample of eligible schools (i.e., treatment schools, control schools, and non-invited schools) and may in some analyses use historical data from the national tests. That is, the national tests were first conducted in 2010 and historical results for earlier cohorts could be used to increase statistical power (conditional on data availability for schools). Furthermore, as both treatment schools and control schools will be more likely to use Læseklar for other cohorts than the one we study in kindergarten, we may also use tests conducted by these cohorts to improve statistical power. However, as data availability is unclear at the time of writing, we cannot describe these analyses in detail.

2.3.3. Multiple hypothesis testing

The post and follow-up tests include testing multiple hypotheses. Testing multiple hypotheses risks increasing the familywise error rate (FWER), i.e., the risk of rejecting at least one true null hypothesis above the desired level of statistical significance (Romano & Wolf, 2005), and risks increasing the false discovery rate (FDR), i.e., the expected proportion of falsely rejected null hypotheses (e.g., Benjamini & Hochberg, 1995). To mitigate these problems, we have clearly designated one primary outcome at both post and follow-up test. If we find statistically significant effects on secondary outcomes, we may examine the FDR.

2.4. Statistical power calculations

We used Stata 16 (StataCorp, 2019) and the command *clustersampsi* (Hemming & Marsh, 2013) to calculate the minimum detectable effect (MDE) for our primary outcome (i.e., not adjusted for multiple hypothesis testing). We focus on the post-test and the primary analysis, and provide some tentative analysis for the peer effects. A power analysis for the follow-up outcomes require assumptions about quite a few more unknown parameters, which makes it difficult to assess power at the time of writing.

All power calculations use the following assumptions: we use standard levels of significance (0.05) and power (0.8). In our main scenarios, we use a within-school intra-cluster correlation (ICC) of 0.10 (based on Hedges & Hedberg, 2007). The average cluster-size (school) is set to 50, based on about 20 students per class and 2-3 classes per school (see e.g., Beuchert, Dietrichson, & Jensen, 2021). The 20% lowest scoring students at pre-test is therefore 10. The number of schools at post-test is assumed to be 80, split evenly between treatment and control schools. Lastly, in analyses with pre-tests, we assume a baseline correlation between post-tests and covariates of 0.4, corresponding to an $R^2 = 0.16$. This level of explained variation is much lower than in the two earlier RCTs of Læseklar, Bøg et al. (2021) and Seerup et al. (2022). As we do not stratify within schools, which the two earlier studies did, and the pre-test we have access to was conducted longer before the start of the intervention compared with the two earlier studies of Læseklar, we expect less explained variation. However, the assumption may be conservative.

In the primary analysis, the MDE is 0.25 assuming that all of the 20% lowest scoring students receive Læseklar. As we will recommend schools to work with 2-3 students per class, the assumption that all of 20% lowest scoring students receive Læseklar may not to hold. Assuming instead that 70% of the 20% lowest scoring students actually received Læseklar and zero peer effects, we need the effect on the students that actually received Læseklar to be 0.25/0.7 = 0.36 for the weighted average to be 0.25. This effect on actually receiving students would be slightly below the average effect of tutoring studies in kindergarten (unpublished analysis using data from Dietrichson et al., 2021) and much lower than the effect of Læseklar on decoding estimated in Bøg et al. (2021; ES = 1.07). We therefore believe we will have adequate power in the primary analysis.

The MDE on the 80% highest scoring students is 0.20. The same ratio of direct effects to peer effects as in Berlinski et al. (2022) would imply that we need a direct effect of 0.65. Although this would be a large effect size, it is clearly below the short-term effects found in the Swedish trial (Bøg et al., 2021).

2.5. Cost-effectiveness

Should there be statistically significant effects of the treatment, we will measure cost-effectiveness as the costs per unit increase in effect size per student. We will calculate the effect size (Hedges' g) as (Lipsey & Wilson, 2001):

$$(2) \qquad g = \left(1 - \frac{3}{4N - 9}\right) \times \frac{\hat{\beta}}{SD_{post}}.$$

The term in the parenthesis is a small sample size correction, $N = n_T + n_C$ is the total sample size, n_T is the number of treated students and n_C is the number of control students, $\hat{\beta}$ is a treatment effect estimate and SD_{post} is the unadjusted post-test standard deviation in the treatment and control group (estimated, e.g., by a multilevel model to account for the clustering of students in schools).[3]

We will calculate the costs of implementing the program for schools; that is, not including the costs of developing the intervention. We will follow the framework for cost-effectiveness analysis of educational interventions developed in e.g., Levin & Belfield (2015) and Hollands et al. (2013, 2016), and calculate the costs for the following "ingredients": personnel costs, school facilities, and materials and equipment. To calculate personnel costs, implementing personnel will be asked to keep track of their time spent learning the method, preparing the sessions, and conducting the sessions using a "diary" developed in the efficacy trial.

If there are peer effects on students in the treated cohorts in treated schools, positive or negative, these effects will be included in the analysis. Similarly, if the program affects parental time investments, these effects will be accounted for.

Effects on non-treated cohorts are also conceivable, especially in the longer run, and we need to check whether the achievement of non-treated cohorts are affected. We will use the same estimation strategies as described earlier to estimate the effects on cohorts in which no student received the program but who are in school at the same time as the treated cohorts. For example, when we estimate the program effects using the second grade national test in Danish for the treated cohorts we can also examine effects on the compulsory third grade test in Danish for earlier cohorts who were too old to receive the program in kindergarten.

Finding effects also on non-treated cohorts may be a sign that schools have reallocated resources. For example, if Læseklar prevents some students from developing reading difficulties later on, this would free up resources that can be used for students with reading difficulties in other cohorts. Such effects should then be included in the cost-effectiveness analysis and in this case, the school is the relevant level for the cost-effectiveness analysis. As schools are normally separate units of account in the municipal budgets, we will then try to obtain total school costs from participating municipalities. To the school costs, we add the costs of all resources provided by the project to each school (program material, support), which we keep track of ourselves. We are interested in the costs of implementing Læseklar per treated student. The extra cost incurred by the intervention in treatment schools would then be equal to the estimate of the treatment effects on total school costs plus resources provided by the project (which are not in the school accounts), divided by the number of students in the treatment schools.

[3] It is possible to calculate the small sample correction exactly (see Hedges, 1981). However, we want to compare our effect sizes to those reported in related reviews and meta-analyses, which often use this formula and the post-test standard deviation (e.g., Dietrichson et al., 2017, 2021). We therefore think that the effect sizes are more comparable when calculated by this formula.

2.6. Alternative research design: Matching and difference-in-differences

If we cannot recruit enough schools, we will expand the sample of control schools and students for the follow-up tests using a matching technique. We will match treatment schools and students to schools and students from the sample of schools that did not receive an offer to participate on the basis of register data (e.g., previous school-level test results, demographics, socioeconomic status). We will use the entropy balancing algorithm developed by Hainmueller (2012), which minimizes covariate mean imbalances as well as more moments of the distribution. We will use the weights from the algorithm in a weighted, cross-sectional difference-in-differences analysis that includes all available national test cohorts to date.

References

Abadie, A., & Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics*, *10*, 465-503.

Andrews, I., & Armstrong, T. B. (2017). Unbiased instrumental variables estimation under known first-stage sign. *Quantitative Economics*, *8*(2), 479-503.

Arendt, K.S.; K. Baunkjær; & B. S. Rangvid (2017): *Forældres oplevelser af skolen i folkeskolereformens tredje år*. Kommenteret talrapport. København: VIVE - Det Nationale Forsknings- og Analysecenter for Velfærd.

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

Bai, Y. (2022). Optimality of matched-pair designs in randomized controlled trials. *American Economic Review*, *112*(12), 3911-3940.

Beuchert, L., Dietrichson, J., & Myrup Jensen, V. (2020). *Forsøg med Alternativ Ressourceanvendelse - Baggrundsnotat med forslag til modeller*. Notat, København: VIVE - Det nationale forsknings- og analysecenter for velfærd.

Bind, M. A., & Rubin, D. B. (2020). When possible, report a Fisher-exact P value and display its underlying null randomization distribution. *Proceedings of the National Academy of Sciences*, *117*(32), 19151-19158.

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, *78*(2), 647-63.

Bleses, D., Jensen, P., and Højen, A. (2017). *CDI-Sprogvurdering. Valideret redskab til vurdering af sprogudvikling hos børn på 18-34 måneder*. Report.

Borusyak, K., & Hull, P. (2021). *Non-random exposure to exogenous shocks*. Working Paper.

Bus, A. G., Van Ijzendoorn, M. H., & Pellegrini, A. D. (1995). Joint book reading makes for success in learning to read: A meta-analysis on intergenerational transmission of literacy. *Review of Educational Research*, *65*(1), 1-21.

Bøg, M., Dietrichson, J., & Isaksson, A. A. (2021). A multi-sensory tutoring program for students at-risk of reading difficulties: Evidence from a field experiment. *Journal of Educational Research*, *114*(3), 233-251.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317-372.

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5-51.

Chapman, J. W., Tunmer, W. E., & Prochnow, J. E. (2000). Early reading-related skills and performance, reading self-concept, and the development of academic self-concept: A longitudinal study. *Journal of Educational Psychology, 92*(4), 703-708.

Chernozhukov, V., Fernández-Val, I., & Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, *81*(6), 2205-2268.

Christensen, V. T. (ed.) (2019). *PISA 2018 – Danske unge i en international sammenligning*. Copenhagen: VIVE – The Danish Center for Social Science Research.

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, *3*(3), 149-210.

Connor, C. M., Day, S. L., Phillips, B., Sparapani, N., Ingebrand, S. W., McLean, L., ... & Kaschak, M. P. (2016). Reciprocal effects of self-regulation, semantic knowledge, and reading comprehension in early elementary school. *Child Development*, *87*(6), 1813-1824.

Currie, J., & Thomas, D. (2000). School quality and the longer-term effects of Head Start. *Journal of Human Resources*, *35*(4), 755–774.

Das, J., Dercon, S., Habyarimana, J., Krishnan, P., Muralidharan, K., & Sundararaman, V. (2013). School inputs, household substitution, and test scores. *American Economic Journal: Applied Economics*, *5*(2), 29–57.

de Chaisemartin, C., & Ramirez-Cuellar, J. (2022). *At what level should one cluster standard errors in paired experiments, and in stratified experiments with small strata?* (No. w27609). National Bureau of Economic Research.

Demetriou, A., Spanoudis, G., & Shayer, M. (2014). Inference, reconceptualization, insight, and efficiency along intellectual growth: A general theory. *Enfance*, *3*, 365-396.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168.

Dietrichson J., Filges T., Seerup J. K., Klokker R. H., Viinholt B. A., Bøg, M. & Eiberg, M. (2021). Targeted school-based interventions for improving reading and mathematics for students with, or at-risk of, academic difficulties in grade K-6: A systematic review. *Campbell Systematic Reviews*, *17*(2), e1152.

Dion, E., Roux, C., Landry, D., Fuchs, D., Wehby, J., & Dupéré, V. (2011). Improving attention and preventing reading difficulties among low-income first-graders: A randomized study. *Prevention Science*, *12*(1), 70-79.

Ehri, L., Nunes, S. R., Willows, D. M, Valeska Schuster, B., Yaghoub-Zadeh, Z, & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly*, *36*(3), 250–287.

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology*, *88*(1), 3–17.

Freedman, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, *40*(2), 180-193.

Follmer, D. J. (2018). Executive function and reading comprehension: A meta-analytic review. *Educational Psychologist*, *53*(1), 42-60.

Gathercole, S. E., Dunning, D. L., Holmes, J., Norris, D. (2019). Working memory training involves learning new skills. *Journal of Memory and Language*, *105*, 19-42.

Gellert, A. S. & Elbro, C. (2016). *Rapport om projekt vedrørende tidlig identifikation af elever i risiko for udvikling af alvorlige afkodningsvanskeligheder (herunder ordblindhed)*. Københavns Universitet: Center for Læseforskning.

Gensowski, M., Landersø, R., Bleses, D., Dale, P., Højen, A., & Justice, L. (2020). *Public and parental investments and children's skill formation* (Study Paper No. 155). Copenhagen: The Rockwool Foundation.

Glennerster R., & Takavarasha K. (2013). *Running randomized evaluations: A practical guide*. Princeton, NJ: Princeton University Press.

Goldstein, D. M. (1976). Cognitive-linguistic functioning and learning to read in preschoolers. *Journal of Educational Psychology*, *68*(6), 680–688.

Greaves, E., Hussain, I., Rabe, B., & Rasul, I. (2021). Parental responses to information about school quality: Evidence from linked survey and administrative data. *Economic Journal*, forthcoming.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate re-weighting method to produce balanced samples in observational studies. *Political Analysis*, *20*(1), 25-46.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107-128.

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

Hollands, F. M., Pan, Y., Shand, R., Cheng, H., Levin, H. M., Belfield, C. R., ... & Hanisch-Cerda, B. (2013). *Improving early literacy: Cost-effectiveness analysis of effective reading programs*. New York: Center for Benefit-Cost Studies of Education, Teachers College, Columbia University.

Hollands, F. M., Kieffer, M. J., Shand, R., Pan, Y., Cheng, H., Levin, H. M. (2016). Cost-effectiveness analysis of early reading programs: A demonstration with recommendations for future research. *Journal of Research on Educational Effectiveness*, *9*(1), 30-53.

Hoover, W. A., & Tunmer, W. E. (2018). The simple view of reading: Three assessments of its adequacy. *Remedial and Special Education*, *39*(5), 304–312.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467-475.

Jacob, R., & Parkinson, J. (2015). The potential for school-based interventions that target executive function to improve academic achievement: A review. *Review of Educational Research*, *85*(4), 512-552.

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis*, *24*, 324–338.

Lavy, V., Paserman, M. D., & Schlosser, A. (2011). Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom. *Economic Journal*, *122*(559), 208-237.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Annals of Applied Statistics*, *7*(1), 295-318.

Marbach, M., & Hangartner, D. (2020). Profiling compliers and noncompliers for instrumental-variable analysis. *Political Analysis*, *28*(3), 435-444.

McClelland, M. M., & Cameron, C. E. (2012). Self-regulation in early childhood: Improving conceptual clarity and developing ecologically valid measures. *Child Development Perspectives, 6*(2), 136– 142.

McKenzie, D. (2022). https://blogs.worldbank.org/impactevaluations/why-i-am-now-more-cautious-about-using-or-recommending-matched-pair-randomization.

Melby-Lervåg, M., Hulme, C. (2010). Serial and free recall in children can be improved by training: Evidence for the importance of phonological and semantic representations in immediate memory tasks. *Psychological Science*, *21*(11), 1694-1700.

Morgan, P. L., Fuchs, D., Compton, D. L., Cordray, D. S., & Fuchs, L. S. (2008). Does early reading failure decrease children's reading motivation?. *Journal of Learning Disabilities*, *41*(5), 387-404.

Ministry of Education (2020). Resultaterne fra de obligatoriske nationale test 2020. Retrieved from: file:///C:/Users/B058956/Downloads/201104-Notat-om-resultaterne-fra-de-obligatoriske-nationale-test-2020.pdf.

Ministry of Education, M. (2019). *Vejledning til sprogvurdering 3-6*. Copenhagen: Technical report, Ministry of Education.

Ministry of Education (2021). Ministeriets sprogvurderingsværktøjer. https://www.uvm.dk/dagtilbud/viden-og-udvikling/materialer-maalrettet-it-leverandoerer/ministeriets-sprogvurderingsvaerktoejer

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity?. *Psychological Bulletin*, *142*(7), 758.

Oberauer, K., Lewandowsky S , Awh E , Brown G D A , Conway A , Cowan N , et al. (2018). Benchmarks for models of short-term and working memory. Psychological Bulletin,*144*(9), 885.

OECD (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris: PISA, OECD Publishing. Retrieved from http://www.oecd.org/publications/pisa-2015-results-volume-i-9789264266490-en.htm.

OECD (2017a). Public spending on childcare and early education. OECD family database. http://www.oecd.org/els/family/database.htm.

OECD (2017b). *Education at a Glance 2017: OECD Indicators*. OECD Publishing, Paris. http://dx.doi.org/10.1787/eag-2017-en

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review*, *76*(3), 241-263.

Peng, P., & Kievit, R. A. (2020). The development of academic achievement and cognitive abilities: A bidirectional perspective. *Child Development Perspectives*, *14*(1), 15–20.

Pop-Eleches, C., & Urquiola, M. (2013). Going to a better school: Effects and behavioral responses. *American Economic Review*, *103*(4), 1289–1324.

Ritchie, S., & Tucker-Drob, E. M. (2018). How much does education improve intelligence? A meta-analysis. *Psychological Science*, *29*(8), 1358–1369.
Romano, J. P., & Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, *100*(469), 94-108.
Scott, N. W., McPherson, G. C., Ramsay, C. R., & Campbell, M. K. (2002). The method of minimization for allocation to clinical trials: a review. *Controlled Clinical Trials*, *23*(6), 662-674.
Seerup, J. K., Dietrichson, J., Jensen, V. M., & Bingley, P. (2022). *Tutoring young children at risk of developing reading difficulties: Effects of an early tutoring intervention on literacy skills and executive functions*. Unpublished manuscript.
Shams, L., & Seitz, A. (2008). Benefits of multisensory learning. *Trends in Cognitive Science*, *12*(11), 411–417.
Shams, L., Wozny, D. R., Kim, R., & Seitz, A. (2011). Influences of multisensory experience on subsequent unisensory processing. *Frontiers in Psychology*, *2*(264), 1–9.
Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, *55*(2), 151-218.
Snow, C., Burns, S., Griffin, P., & the Committee on the Prevention of Reading Difficulties in Young Children, National Research Council (1998). *Preventing reading difficulties in young children*. Washington DC: The National Academies Press.
Snowling, M., & Hulme, C. (2011). Evidence-based interventions for reading and language difficulties: Creating a virtuous circle. *British Journal of Educational Psychology*, *81*(1), 1–23.
Socialstyrelsen (2017). *Vejledning til Sprogvurdering 3-6*. Retrieved from: https://bedstsammen.kk.dk/sites/bedstsammen.kk.dk/files/vejledning_-_sprogvurdering_3-6_-_06122017.pdf.
Stanovich, K. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360–407.
Torgerson, C. J. & D. J. Torgerson (2007). The use of minimization to form comparison groups in educational research. *Educational Studies*, *33*(3), 333-337.
Toste, J. R., Didion, L., Peng, P., Filderman, M. J., & McClelland, A. M. (2020). A meta-analytic review of the relations between motivation and reading achievement for K–12 students. *Review of Educational Research*, *90*(3), 420-456.
Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology*, *53*, 1-25.
Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard university press.
Walgermo, B. R., Frijters, J. C., & Solheim, O. J. (2018). Literacy interest and reader self-concept when formal reading instruction begins. *Early Childhood Research Quarterly*, 44, 90-100.
Xu, Z., Proschan, M., & Lee, S. (2016). Validity and power considerations on hypothesis testing under minimization. *Statistics in Medicine*, *35*(14), 2315-2327.
Young, A. (2018). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *Quarterly Journal of Economics*, *134*(2), 557–598.
Zelazo, P. D. (2015). Executive function: Reflection, iterative reprocessing, complexity, and the developing brain. *Developmental Review*, *38*, 55-68.