Early Childhood Research Quarterly 57 (2021) 12–26

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

journal homepage: www.elsevier.com/locate/ecresq

Implementing toddler interventions at scale: The case of "We learn together"

Dorthe Bleses[a,b,c,d,*], Peter Jensen[b,c,d], Anders Højen[a,b,c,d], Pauline Slot[e], Laura Justice[f]

[a] School of Communication and Culture, Aarhus University, 8000, Aarhus C, Denmark,
[b] TrygFonden's Centre for Child Research, Aarhus University, 8000, Aarhus C, Denmark,
[c] Department of Economics and Business Economics, Aarhus University, 8000, Aarhus C, Denmark,
[d] Centre for Integrated Register-based Research, CIRRAU, Aarhus University, Aarhus, Denmark
[e] Department of Development and Education of Youth in Diverse Societies, Utrecht University, Netherlands
[f] Department of Educational Studies, The Ohio State University, OH, USA

ARTICLE INFO

Article history:
Received 8 July 2020
Revised 20 April 2021
Accepted 24 April 2021

Keywords:
School readiness intervention
Toddler classrooms
Replication of first stage trial
At-scale effectiveness study
Language
Math
Social-emotional outcomes

ABSTRACT

The first years of life are characterized by rapid learning in several school readiness domains, including language, math, and social-emotional skills, all of which are important for later childhood outcomes and academic achievement in school. In this research, we investigate the effects of an early-stage evidence-based school readiness intervention and add-on elements to support teachers' implementation under real-life circumstances to explore the readiness of this intervention for scaling up. We replicated the findings of a prior trial of this intervention, and demonstrated that a 20-week curriculum with instructional content and supportive tools for teachers to be more explicit and intentional in their interactions with children can be implemented successfully under real-life circumstances and result in positive effects on targeted language and math skills.

Introduction

The first years of a child's life are characterized by rapid learning of a variety of skills that are important for early childhood outcomes and later achievement in school. Children's early experiences at home serve to shape these early skills (Phillips et al., 2017). Consequently, the significant variability among children in their early home experiences, partially driven by family socioeconomic status (SES) and sociocultural background, contributes to substantial disparities among children in early skill development during infancy and toddlerhood (e.g., Fernald, Marchman, & Weisleder, 2013; Hoff, 2013). For instance, 1 recent study showed that significant disparities in language skill distinguish 2-year-old children from less- and more-advantaged homes within a low-SES sample, with a medium-sized effect size ($d = 0.44$; Justice, Jiang, Bates, & Koury, 2020). To be clear, we view these early gaps in skill development to be a product of society's systematic failures to provide equitable opportunities for learning to all children. Consequently, early childhood researchers can develop, test, and scale practices that can create more equitable early learning opportunities for all children.

To this end, this early gap in skills for advantaged vs less advantaged very young children has led to extensive research activity focused on improving learning opportunities for young children in early childhood education (ECE) settings. Provision of quality preschool represents 1 avenue for providing enriching experiences to all young children, and studies find that it has overall net-positive impacts on children's language, literacy, math, and social-emotional skills in the early years (see Duncan & Magnuson, 2013). Similarly, evaluations of more tailored interventions have also shown positive outcomes (e.g., Chambers, Cheung, & Slavin, 2016).

A key limitation of much of the work focused on identifying the benefits of ECE programming is that little attention has been directed towards *scaling* findings from efficacy studies into larger-scale effectiveness studies, or exploring factors that support implementation under real-life conditions to assess the readiness of interventions for scaling up (Greenwood, Schnitz, Carta, Wallisch, & Irvin, 2020; Schindler, McCoy, Fisher, & Shonkoff, 2019; Walker et al., 2020). In the present paper, we address this limitation by investigating the effects of an early-stage evidence-based

* Corresponding author.
E-mail address: bleses@au.dk (D. Bleses).

https://doi.org/10.1016/j.ecresq.2021.04.008

"school readiness intervention" that targets early skill development in toddlers (Bleses, Jensen, Slot, & Justice, 2020), an age range seldom studied in large-scale intervention studies (Burchinal, Magnuson, Powell, & Hong, 2015; Greenwood et al., 2020; Larson et al., 2020; Walker et al., 2020). Our particular interest is assessing impacts of this intervention when used at scale under real-life circumstances in the context of add-on elements to support implementation.

Early school readiness is important for later outcomes

Schoolreadiness skills are a set of competencies including language, math, and social-emotional skills that start to emerge in toddlerhood. Vocabulary and complex language skills develop rapidly during the toddler years (Fenson et al., 2007), as do early numeracy skills (i.e., knowledge about number and quantity; Clements & Sarama, 2011; Duncan et al., 2007) and content-specific language relevant for math, including words referring to quantities (e.g., "more" or "less,"), spatial language (e.g., "above" or "beneath") or names of shapes (e.g., "triangle" or "square," see Purpura & Reid, 2016). The various schoolreadiness domains show significant inter-correlations during the toddler and preschool years (Duncan et al., 2007; McClelland et al., 2007; Slot & von Suchodoletz, 2018; Slot, Bleses, & Jensen, 2020) and show consistent relations with children's future academic achievement (Cameron, Kim, Duncan, Becker, & McClelland, 2019; Nix, Bierman, Domitrovich, & Gill, 2013; Son, Choi, & Kwon, 2019).

Skill gaps in such school readiness domains are apparent in the second year of life and appear to be quite stable over time (Bornstein, Hahn, & Putnick, 2016; Hammer et al., 2017; Justice et al., 2020). These gaps are linked to social-demographic factors including SES and immigrant background, pointing to the importance of providing high-quality experiences in the early years of life to children in these subgroups of families (Fernald, Marchman, & Weisleder, 2013; Gilkerson et al., 2018; Hammer et al., 2017; Hoff, 2013). To promote early skill development, many broad ECE programs and more specific curricula have been developed and evaluated, as described in several recent meta-analyses; generally, these produce positive short-term effects, and in some cases positive long-term effects, but with large variation among offerings (e.g., Chambers, Cheung, & Slavin, 2016; Duncan & Magnuson, 2013; Greenwood, Schnitz, Carta, Wallisch, & Irvin, 2020; Walker et al., 2020). These recent reviews highlight several current limitations to this body of work.

Interventions targeting infants and toddlers

The first significant gap is that most of the evidence of efficacy and effectiveness of ECE programming is focused on preschool-aged children. Much less is known about the impacts of interventions targeting infant and toddler childcares, (e.g., Greenwood, Schnitz, Carta, Wallisch, & Irvin, 2020; Walker et al., 2020). Moreover, existing research on the youngest children has mainly tested the effectiveness of various professional development interventions (e.g., teacher coursework, onsite coaching, video-based feedback) aimed at improving the quality of interactions in center- and family-based childcare without a specific skill-based curriculum that focuses on developing specific school readiness domains (e.g., Helmerhorst, Riksen-Walraven, Fukkink, Tavecchio, & Deynoot-Schaub, 2017; Werner, Vermeer, Linting, & Van IJzendoorn, 2018). For instance, Moreno and colleagues (Moreno, Green, & Koehn, 2015) showed that professional development consisting of a combination of coursework and coaching demonstrated patterns of improvements in teacher practices related to the quality of interactions, but children's skill development was not directly targeted in the intervention.

In recent years, however, 2 randomized controlled trial (RCT) studies have evaluated more content-specific interventions across different school readiness domains, but with different foci. The first study evaluated the effect of a curriculum ("Responsive Early Childhood Curriculum,"or RECC), that supports learning in both ECE teachers and children (Landry et al., 2014). The RECC curriculum supports quality (e.g., by encouraging teachers to respond contingently to children's signals) and provides age-appropriate, stimulating activities to promote language, early literacy, and math development in toddlers as well as social-emotional curriculum. The study showed increased process quality in intervention classrooms but significant effects only on children's social and emotional development. This finding highlights that both specific instructional content and professional development focusing on providing a rich learning environment may be needed to promote language and math development for this age range (Landry et al., 2014). The second study tested the effect of a toddler school readiness intervention "Play and learn" (now named "We learn together") with a strong content-specific focus based on sequence and scope, similar to those tested in preschools (Bleses, Jensen, Slot, & Justice, 2020). "We learn together" is based on a scope of instruction targeting language and math sequenced over a 20-week curriculum period. Supportive tools for teachers to help them be more explicit and intentional in their interactions with children were provided, but they were asked to develop activities of their own choice within a theme-based instructional framework to sustain some teacher discretion in implementation, following the positive results with a previous intervention that used a similar approach (Bleses et al., 2018a). The intervention study demonstrated positive, mainly medium- to large-sized effects on targeted language and math skills (d = 0.19–0.80) among toddlers.

A second significant gap concerns real-world implementation of interventions by non-researchers (cf. Durlak & DuPre, 2008) and scaling up of interventions in early-learning settings. A recent systematic review of early language interventions found that there is an urgent need to increase knowledge about the effectiveness of interventions taken to scale. The term "at scale" refers to population-level implementation of (effective) interventions (Fagan et al., 2019). In fact, the authors of the review concluded that most studies are efficacy studies with small samples including mainly at-risk children and very few studies had more than 200 participants (N = 8) (Walker et al., 2020). Effectiveness studies of interventions implemented at scale based on representative samples are critical to evaluating for whom an intervention works the best (e.g., Schindler, McCoy, Fisher, & Shonkoff, 2019). Such research is particularly important as recent research has indicated that intervention effects are reduced when implemented at scale. For instance, 2 RCT studies in Denmark investigated the effectiveness of a systematic, explicit language and literacy intervention that was implemented across Denmark under real-world conditions involving large, heterogeneous samples of teachers and children (Bleses et al., 2018a, Bleses et al.,2018b). Although positive effects were reported, effects seem to be affected by degree of fidelity of the intervention (Bleses et al., 2018a, Bleses et al., 2018b).

Implementation fidelity as a barrier for intervention fidelity

To understand intervention effectiveness at scale, it is crucial to examine more precisely how fidelity of implementation may relate to child outcomes and how fidelity can be improved when applied at scale (Biel et al., 2020; Schindler, McCoy, Fisher, & Shonkoff, 2019). In a recent review of implementation research, Hsueh, Halle, and Maier (2020) distinguished between "implementation fidelity" and "intervention fidelity." Implementation fidelity refers to the degree to which the implementation infrastructure (e.g., professional development and different types of support) are provided

as intended and represent elements of an *outward focus* on implementation. The term intervention fidelity, on the other hand, is used to describe the degree to which the intervener delivers the intervention as intended. Intervention fidelity provides a framework for *inward examination* of a program's theory of change or implementation processes, including the level of fidelity needed to achieve intervention effects. Intervention fidelity is regarded as a multidimensional construct and involves assessment of implementation dosage, adherence, as well as quality (see also Durlak & DuPre, 2008). Biel et al. (2020) explored implementation and intervention fidelity as reported in studies of language interventions in ECE settings. Even though such intervention studies increasingly reported aspects of fidelity of programs, around 42% did not report fidelity or investigate associations of fidelity with child outcomes (Biel et al., 2020). The majority of studies assessed implementation fidelity using reporting (by checklists (56%) or frequency logs (14%)), whereas approximately 30% used observation measures.

In sum, even though early childhood is a period of major growth in language, math, and social-emotional skills, intervention studies targeting this age range are scarce, and there is a need for more high-quality intervention research targeting toddlers. Furthermore, current intervention research is mainly based on efficacy studies but more knowledge on *if* and *how* intervention effects can be reached at scale is needed.

Supporting implementation of interventions at scale

Approaches to promoting the implementation of at-scale interventions are still only emerging. Here, we focus on promising approaches. The first approach is to use theory of change models as active tools for examining not only if a program is effective but also under which conditions a program works or not and for whom. One example of such a model is the IDEAS (Innovate, Develop, Evaluate, Adapt and Scale) Impact Framework (Schindler, McCoy, Fisher, & Shonkoff, 2019), which includes 4 components. First, *program strategies* are the actions the program takes to achieve change, which includes program dosage (quantity of sessions) and program fidelity (extent to which the delivered program matches the intended program). Moreover, program strategies include actions that promote the implementation quality of program strategies, such as standardization of the intervention and standardization of staff training. Second, *program targets* are the skills and behaviors that the program strategies are attempting to change. Third, *expected outcomes* are the real-world changes in skills and behaviors, and fourth, *moderators* are the factors that affect intervention effects (Schindler, McCoy, Fisher, & Shonkoff, 2019).

The second approach is to use implementation frameworks to support implementation at scale, which is a rather new effort in ECE settings. One framework specifically developed for ECE settings is the Integrated Stage-Based Framework for Implementation of Early Childhood Programs and Systems (Metz, Naoom, Halle, & Bartley, 2015), which is based on previous syntheses in implementation science literature but adapted to ECE settings. The framework is based on 4 stages (*exploration, installation, initial implementation and full implementation*) with 3 core elements across each of the stages. These include (1) building and using implementation teams to actively lead implementation efforts, (2) using data and feedback loops to drive decision-making and promote continuous improvement and (3) developing a sustainable implementation infrastructure that supports capacity for individuals, organizations and communities (Metz et al., 2015). There is evidence from other scientific fields, in particular the health sciences, that show that such processes will improve implementation fidelity and thereby increase effectiveness of interventions, when implemented at scale (Metz et al., 2015).

Interpreting effects of large-scale intervention studies in education

Typically, when evaluating the magnitude of effect sizes, Cohen's suggested labels are used (0.2 = small; 0.5 = medium; and 0.8 = large). However, these standards are based on a few small-scale tightly controlled laboratory experiments in the 1960s and may have limited applicability to present education sciences. Recent meta-analyses of well-designed field experiments of education interventions indicate much lower effects are to be expected. Cohen's *d* is therefore considered an outdated and outsized standard for what constitutes meaningful effect sizes in education (Bailey, Duncan, Cunha, Foorman, & Yeager, 2020; Kraft, 2020).

Recently, Kraft (2020) developed benchmarks for interpreting effect sizes in education based on analyses of 747 RCTs of preschool to grade 12 education interventions and proposed the following benchmarks: Less than 0.05 = small; 0.05 to less than 0.20 = medium; and 0.20 or greater = large effect size. Moreover, when interpreting the potential impact and policy relevance of effect sizes, parameters like research design, cost and scalability are relevant. A low-cost large-scale intervention with large effects has, for instance, potentials for huge impact compared to a small-scale invention targeting specific groups of children.

Taking a toddler intervention to scale

In the present study, we tested the toddler school readiness classroom-based intervention "We learn together" (Bleses, Jensen, Slot, & Justice, 2020) at scale by using theory of change as a tool and by adapting components from intervention frameworks to a Danish context. The effectiveness of the intervention was tested under routine circumstances in an earlier early-stage RCT involving 87 childcare centers and 1116 toddlers (Bleses, Jensen, Slot, & Justice, 2020).

Here, we highlight several key findings from that trial. First, the intervention study demonstrated effect sizes in the magnitude of 0.19–0.80 standard deviation (SD) units on targeted language and math skills. Second, there was variation in teacher's implementation fidelity, even though the intervention was very well received by the involved teachers and municipalities. Third, differential effects were found for children. For children of less-educated mothers, the treatment had no significant effect for vocabulary (in contradiction to children of high and mid-high educated mothers), and for children with a non-Western background the treatment no significant effects were found for any outcomes. Given the latter finding, it is important to assess the extent to which these limitations are related to fidelity of the intervention and if they can be reduced by supportive implementation elements. Moreover, not all elements of the intervention support at-scale use. For instance, the professional development provided to implementing teachers was delivered by the research team, which is generally not feasible for at-scale use.

The theory of change for "We learn together" using the IDEAS framework can be seen in Table 1 (the intervention is described in detail in the method section), and in the following, we identify the measures taken to support scale up use of "We learn together." These includes (1) community engagement strategies; (2) standardization of staff training by the development of a train-the-trainer model of the professional development course for local implementation; and (3) introduction of local implementation teams that make use of implementation data.

Community engagement strategies to support local organization and implementation of the intervention in each municipality included an initial organization readiness workshop in the municipalities. Organizational readiness concerns the prerequisites of the participating municipalities to participate in the intervention and to implement sustainable practice changes after the interven-

Table 1
Components of the theory of change of "We learn together."

	Elements
Program strategies Program dosage	• 2 large-group, 2 small-group sessions and 1 individual conversation per week.
Standardization of program strategies	• Weekly planner with theme based instructional framework with aligned program targets. • Posters supporting rich conversations and use of responsive strategies. • Standardized staff training delivered by consultants. • Use of real-life implementation data. • Implementation teams (monthly meeting in each classroom or family-based care unit).
Service offered to municipality consultant	• 2-days-train-the-trainer course in the teacher professional development course, developed and offered by the research team.
Service offered to the staff and leaders	• Standardized teacher professional development course, developed by the research team, and offered by the municipality consultant. • One teacher per childcare unit was offered a course to support implementation during the course of the intervention.
Level of implementation fidelity	• Dosage is measured via weekly implementation notes at the level of the individual child.
Program targets Improved classroom instruction	• Use of targeted instructional content (language, math), measured via weekly classroom implementation notes and child progress checklists. • Use of instructional activities measured via standardized questionnaire. • Use of rich language and responsive strategies measured via two video-based observations of classroom instruction (around week 10–12).
Expected outcomes	• Improved instructional content, richer language and more frequent use of instructional strategies • Improved outcomes in language (vocabulary and language use), math language and numeracy.
Moderators	• Child and family characteristics (gender, age, ethnicity, parent's education, family type, parents' employment status and housing type). • Program dosage. • Number of implementation team meetings.

tion. These workshops, attended by representatives of municipality management, day-to-day management and selected staff, aimed to identify potential implementation drivers and barriers in the local organization, which could be used to support implementation. All the leaders in the childcare centers also participated in the staff' course and in an additional 1-day course focusing on their role in supporting the intervention implementation in their childcare centers. Finally, the research team and the municipalities were collaborating on the development of the add-ons to the original intervention to support engagement (Olswang & Goldstein, 2017).

A train-the-trainer course targeting educational consultants in the municipalities was developed. The educational consultants were introduced to and instructed in delivering the 2-day professional development course, which trained the staff for the implementation of "We learn together" (see the method section for a description of the elements in the course). The local trainers were able to embed the new practice in other municipality programs, which reinforced the relevance of the intervention to the teachers. The "train-the-trainer" model was also chosen to support subsequent sustainability of the intervention.

Implementation teams were used during the course of the intervention to support implementation fidelity and the inclusion of all children in the intervention activities. As teachers in Danish toddler childcare have very little experience with implementing evidence-based practices and using implementation data, we developed a "light" version of implementation teams. Supported by a lead colleague who was responsible for planning the classroom meetings, each classroom was expected to meet 5 times (once in each 4-week theme) to discuss the implementation of the intervention. The implementation meetings were based on data from (1) weekly implementation notes completed by the staff that included information about content, child exposure and child progress on the targeted domains (once in each 4-week theme); and (2) brief videos of implemented activities. The lead colleague received a 2-day course held by the research team to facilitate the implementation meetings: one was on the content of the intervention and the other one focused on their role in facilitating dialogues about the implementation.

Present study

The present study describes evaluation of the at-scale version of "We learn together" in a large-scale effectiveness study based on a heterogeneous sample of children. We first examined the extent to which it was possible to *replicate* the findings of the early-stage intervention study under real-world conditions. Following Schindler, McCoy, Fisher, & Shonkoff, 2019, we investigated factors related to the readiness of the intervention for scaling up described above, specifically *potential moderators of intervention effects* including for whom does the intervention work and *program strategies* (the fidelity of the intervention). The "We learn together" intervention is a classroom-based intervention, where teachers work together with the overall goal of including individual children in at least 2 large group activities, 2 small group-activities, and 1 individual conversations per child each week. Number of activities provided by the team of teachers is only registered at the child level and not at the teacher level. We therefore operationalize intervention fidelity (dosage of implementation) as the total number of activities *individual* children participated in (see more elaborated description below). We use the same measure (number of activities an individual child is included in) to explore how intervention exposure is associated with child outcomes following earlier Danish RCT studies (Bleses et al., 2018a, 2018b). We also describe adherence to the intervention (a different aspect of intervention fidelity) using observation data. As the train-the-trainer course may have an impact on how well the intervention was introduced to the teachers with potential impact on intervention fidelity and adherence, we also evaluated aspects of implementation fidelity, that is, the quality of the local professional development course.

Five research questions were addressed: (1) What are the main effects of the intervention "We learn together" on children's language, math, and social skills? Based on the early-stage study (Bleses, Jensen, Slot, & Justice, 2020), we hypothesized that the implementation of "We learn together" would result in positive main effects on language and math outcomes and potentially also social outcomes. (2) To what extent are the impacts of "We learn together" conditional on child and parent characteristics? As in the early-stage study study, we expected that child and parent characteristics would moderate intervention effects. (3) What is the intervention fidelity of "We learn together" when carried out in a real-world setting, and how are child, teacher and center level characteristics related to fidelity? Based on results from other large-scale RCTs in a Danish context (Bleses et al., 2018a , 2018b), we hypothesized that fidelity would be variable across teachers, although we had no prior expectations of how child, teacher and center characteristics would predict fidelity, as this was not investigated in

the early-stage study. (4) To what extent is intervention exposure associated with child outcomes? We hypothesized, based on earlier Danish RCT studies, that higher exposure would be associated with higher child outcomes. (5) Does the use of implementation teams improve fidelity? Based on the implementation literature (e.g., Metz et al., 2015), where the positive role of implementation teams has been documented, we hypothesized that more use of implementation teams would promote higher intervention fidelity.

Methods

Design

The intervention was evaluated using a cluster-randomized controlled trial running for 2 periods (i.e., 2 periods of 20-weeks over years including baseline- and pretest measures, implementation and posttest measures). Randomization occurred at the level of center-based units and family-based childcare units, respectively, and occurred prior to the start of the first period. In the first period, the design was equivalent to a standard cluster-randomized controlled trial with all childcare centers, allocated to either a treatment group or a "business as usual" control group. All results reported below are based on data solely from the first period.

Participants

The number of participants was determined using *a priori* power calculations based on the goal of detecting an effect size of $d = 0.2$ with a power of 80% and a standard significance level of 5%. This resulted in the enrollment of 255 childcare centers of which 111 were center-based care units and 144 were family-based care units (we use the term childcare center to refer to both types of childcare settings). The childcare centers were recruited from 13 different municipalities in Denmark.

All children in the included childcares participated in the interventions. For the main analysis of the present study, we only included children aged 18 months or older at the time of the pretest. The enrolled childcare centers did serve children aged below 18 months; these children also participated in the intervention, but no reliable outcome measures were obtained for these children due to their young age and the lack of appropriate outcome measures. For the toddlers, 2170 children were enrolled in the study and completed both pretest and posttest assessments. Data for these children were collected between August 2016 and March 2017. Baseline characteristics of participating childcare centers, children and teachers (we use the term teacher to refer to any adult participating in the intervention) are shown in Table 2.

By using the personal identification number of the Danish civil registration system, we were able to obtain information from administrative registers on child and family background through Statistics Denmark. The register data provided detailed information on parents' education, immigrant status, country of origin, income, employment status, and other characteristics. Balancing tests of baseline characteristics of the child and parents showed that these were distributed evenly across the control and treatment groups.

In the participating childcare centers, we also collected data on teachers through a questionnaire (N = 696). The baseline characteristics of the teachers are shown in Table 2. Balancing tests of baseline characteristics of the teachers showed that these were distributed evenly across the control and treatment groups.

During the intervention period, there was a constant inflow and outflow of children in the childcare centers, typically because children often transferred to preschool shortly before the age of 3 years. In addition to the main analytic sample, we therefore also have children who either left or entered the childcare centers during the intervention period. Since we evaluated the intervention in a large-scale trial under real-world conditions and therefore had to rely on the cooperation of the childcare centers (and their teachers), we did not impose data collection requirements on incoming children and did not collect data for children who left before the intervention finished. Thus, these groups of children are characterized by missing either pretest or posttest measures and for those in the intervention group by not having been exposed to the full intervention. Hence, we did not include leavers and new arrivers in analyses.

The final analytic sample for the investigation of the children's outcomes consists of all children aged 18 months or over who had both pretest and posttest scores for a given outcome measure. Summary statistics of these outcome measures are shown in Table 3, separately for the control and the treatment group. In general, the means show a clear increase in the score from the pretest to the posttest, implying improved outcomes for the children over time. Balancing tests showed that the pretest outcome measures are balanced across control and treatment groups.

Attrition

Since this study is based on a large-scale trial under real-world conditions, we experienced attrition among the participants, both at the cluster level and at the individual level. First, there was cluster-level attrition with 6 childcare centers not providing any posttest data (3 in the control group and 3 in the treatment group; 4% and 2% of the recruited centers, respectively). One childcare center closed during the intervention period, while the other 5 did not collect posttest data for unknown reasons. Second, in the remaining centers, there was individual-level non-response, either at pretest or at posttest (or both). Overall, pretest data was available for 3588 children (94% of the recruited sample), while both pretest and post-test data was available for 2170 children (57% of the recruited sample and 60% of the pretest sample; see Table S1 for further information).

Pretest and/or posttest data may be missing for children due to illness or other absences at the time of testing. However, the main reason for the individual-level attrition between the pretest and the post-test was that children left the childcare center to transition to preschool and in fewer cases because of residential mobility, as families with young children are often in need of different (and larger) housing arrangements. When the families change their residential location, they typically move their children to a childcare center near their new home.

We assessed the attrition rates of the study by using the standard of the US Department of Education, What Works Clearinghouse (WWC, Clearinghouse, 2017). The WWC's attrition standard applies to the combination of overall and differential attrition, where the latter is the difference in rates of attrition for the control and treatment groups. The combinations are classified as resulting in tolerable, potentially tolerable, or unacceptable levels of potential bias. For our study, the combinations of overall and differential attrition from randomization to posttest fell in the region where the threat of bias is tolerable (both at cluster level and at individual level), even under cautious assumptions (even though the description above indicates that more optimistic assumptions may be appropriate). From pretest to posttest, the combinations were also in the tolerable region.

Furthermore, WWC also points to the risk of bias due to children entering childcare centers after the time of random assignment (joiners). However, as described previously, we did not include any joiners in the analytic sample and therefore this did not pose a risk of bias in our study.

Table 2
Baseline characteristics of childcares, children, and teachers (by group) and balancing tests.

	Control	Treatment	Balancing tests
Childcares			
No. of centers/groups	66	183	
No. of classrooms/family-based units	180	499	
No. of children	594	1576	
No. of teachers	167	529	
	Mean	Mean	*P* value
Child characteristics			
% girls	47.6	51.5	0.24
Age (mean in months)	23.1	23.4	0.13
% Danish origin	89.9	89.0	0.75
% Western origin	2.7	3.1	0.73
% non-Western origin	7.4	7.9	0.83
% Maternal education: low	15.5	16.4	0.67
% Maternal education: low-mid	32.3	36.0	0.11
% Maternal education: high-mid	32.2	29.8	0.36
% Maternal education: high	16.0	14.0	0.48
% Maternal education: missing	4.0	3.9	0.89
% Paternal education: low	16.3	19.1	0.16
% Paternal education: low-mid	48.0	44.9	0.29
% Paternal education: high-mid	18.4	18.8	0.83
% Paternal education: high	12.0	11.7	0.91
% Paternal education: missing	5.4	5.6	0.88
% Parents married or cohabiting	85.6	85.9	0.89
% Parents without employment	12.5	11.2	0.60
% Owner-occupied housing	63.2	62.6	0.88
Teacher characteristics			
% female	97.6	98.5	0.44
Age (mean in years)	47.9	47.2	0.41
% >10 years of experience	65.3	66.2	0.83
% BA degree	37.7	34.6	0.46

Notes: Data for children are only based on children aged 18 months or over at the time of the pre-test and with both pre-test and post-test assessments. Data for teachers are only based on teachers who answered both the pre-test and the post-test questionnaire. Balancing tests are for equality of means in the control and treatment group. For child characteristics, the p-values are from regressions of row variables on a treatment indicator variable with standard errors adjusted for clustering at the childcare center/group level. For teacher characteristics, the balancing tests are two-sample test of proportions (*z*-test), except for age, which is a two-sample *t*-test with unequal variances.

Table 3
Pre-test and post-test scores for children's outcomes.

	N	Control Pre-test Mean(SD)	Post-test Mean(SD)	Treatment Pre-test Mean(SD)	Post-test Mean(SD)
Productive vocabulary (0–70)	2170	26.7(19.3)	47.2(17.7)	27.2(18.5)	49.3(18.1)
Receptive vocabulary (1–2) (0–39)	881	18.1(9.8)	24.4(6.7)	19.0(9.4)	25.6(6.5)
Receptive vocabulary (2–3) (0–39)	937	23.3(6.9)	25.0(5.2)	23.5(6.6)	25.9(5.2)
Language use (0–10)	2170	3.6(3.0)	6.4(2.9)	3.6(2.8)	6.4(2.9)
Math language (0–72)	2159	10.5(10.7)	24.7(15.3)	10.2(9.9)	30.8(15.7)
Numeracy (0–30)	2159	3.6(5.0)	10.3(6.9)	3.6(4.8)	11.9(7.1)
Empathy (0–18)	2039	12.2(4.5)	14.9(3.3)	12.0(4.2)	15.2(3.3)
Self-regulation & cooperation (0–12)	2039	7.9(2.7)	9.2(2.5)	7.8(2.6)	9.3(2.4)

Notes: For receptive vocabulary, (1–2) indicates that the pre-test is from age-dependent test 1 and the post-test is from age-dependent test 2. Similarly, (2–3) indicates that the pre-test is from age-dependent test 2 and the post-test is from age-dependent test 3. The other numbers in parentheses are the range of the measure.

Procedure

Intervention

The "We learn together" intervention is framed to build caregiver and child capacities. The intervention was designed taken later scalability into account, both in terms of key elements of the intervention and the cost-effectiveness of the intervention. The intervention is brief (20 weeks), includes few critical elements with teacher discretion in implementation (see below) and can be implemented within the existing ECE structure without additional financial or human resources. The intervention is classroom-based and teachers are trained to collaborate towards including individual children in a specific number of activities each week (2 large group activities, 2 small group-activities, and 1 individual conversations per child per week). Explicit costs include paying for substitutes while the staff participates in the 2-days professional development course (14 hours) and expenses to intervention materials. The total cost for implementing the intervention has been calculated to 120 US dollars per individual (Rosholm et al., 2021). The intervention was originally developed in close collaboration with

teachers and educational consultants in the involved municipalities.

As is commonly found in interventions aimed at improving child language learning outcomes (Greenwood, Schnitz, Carta, Wallisch, & Irvin, 2020), the "We learn together" intervention is based on naturalistic conversation models that emphasize the child's interest and initiations as opportunities to model and prompt language use during and across daily contexts. Similarly, the intervention captures several of the different elements that Weiland et al.'s exploratory review (Weiland, McCormick, Mattera, Maier, & Morris, 2018) identified as particularly promising. The intervention material (the curriculum and materials) was standardized at a level that allows others to implement the intervention as intended in the efficacy trial intervention (Bleses, Jensen, Slot, & Justice, 2020) and the main elements are:

The intervention is based on naturalistic conversation models and incorporates teacher choice. To promote fidelity to intervention components, effective curricula are often manualized via (semi)scripted lessons that teachers implement with the children (Weiland et al., 2018). However, recent research finds that teacher-implemented curricula effects are heightened if teachers are provided with some discretion in implementation (Bleses et al., 2018a, Bleses, Jensen, Slot, & Justice, 2020). Therefore, instead of receiving scripted lessons, teachers in each classroom were asked to develop activities of their own choice to address targeted skills and to use a set of strategies to enrich and differentiate their interactions with individual children. These interactive strategies included contingently responding to children's comments and engaging children in extended conversations (Phillips et al., 2017), which are important predictors of child development.

Posters and videos that support the use of specific practices promoting enriched conversations were provided to the teachers. The strategies were visualized on posters to be put on the wall: (1) "The high-quality conversation" poster included strategies like open-ended questions, repetition and expansion; (2) "The Learning new words" poster included word learning strategies like relating new words to known words, and providing examples from the same or related semantic categories inspired by Beck and McKeown (2007); and (3) "The learner's ladder" poster included examples of targeted low and high support strategies to scaffold children's learning, based on Justice et al. (2010). Moreover, optional intervention materials were developed and made available for the teachers. The materials included pictures of target words, both as small and larger pictures (so they could be used in, e.g., memory games) and in the form of posters which included pictures of all target words for language and math, respectively, one for each theme. The intervention material also included 5 books, which matched each of the 5 themes (see below).

The intervention has a significant focus on teaching specific instructional content. The intervention includes a twofold scope of instruction targeting language (general vocabulary and language use) and math (math vocabulary and numeracy skills) that was sequenced over the 20-week curriculum period. The sequence and scope was developed based on empirical databases (e.g., Jørgensen, Dale, Bleses, & Fenson, 2009). Examples of the sequence of scope includes (1) *general vocabulary*: thematic words (typical words for objects, events and actions, e.g., "beach," "wet"), words for feelings (e.g., "happy," "angry"), time (e.g., "before," "now") and space ("down," "under"); and (2) *math vocabulary*: words for numbers, shapes (e.g., "round," "square"), sizes (e.g., "big," "short"), and patterns (e.g., "dots," "stripes"). To support teachers in their planning of activities, pictures and posters of target words were provided but it was voluntary to use them.

Social-emotional skills are generally well-supported in a Danish context (Slot, Bleses, Justice, Markussen-Brown, & Højen, 2018), and were therefore not directly targeted in the intervention, but we measured effects on social-emotional development to capture potential spill-over effects (see Table 5).

The intervention was organized in 5 content themes, which were developed in collaboration with teachers during the development of the intervention. The themes were chosen to be engaging and relevant for toddlers and included: "My daily life"; "The weather"; "Animals and nature"; "My family"; and "Out in the world." Each theme lasted 4 weeks to give teachers and children time enough to engage deeply in theme-relevant activities. To guide intervention implementation, teachers in each classroom received a weekly planner offering a theme-based instructional framework by which to plan these activities in alignment with the week objectives; the planner provided a suggested dosage of 2 large-group and 2 small-group activities per week and 1 individual conversation per child each week (among others to minimize effects of implicit biases in engaging children in activities). Exposure at the individual child level in these 3 types of activities were registered on the weekly planner during the week to support the completion of implementation notes (see below). Another different type of activity was called "exploration zones", where teachers arranged interesting areas indoor or outdoor (e.g., with pictures, toys, outdoor materials etc.) to promote interest and curiosity in children and when children initiated playful activities, adults were encouraged to join in and interact with the children.

The intervention uses technology and real-time data to support implementation. Based on research indicating the importance of self-evaluation and reflection when implementing instructional practices (Crawford, Zucker, Van Horne, & Landry, 2017), each classroom completed implementation notes on a technology-based logging tool and monitored children's gains from the intervention to support reflection on implementation. The implementation logs represented the experiences of individual children in each classroom. The self-reported implementation notes also tracked elements of adherence to the intervention (e.g., which educational activities teachers used, the extent to which the learning objectives were targeted).

Teacher professional development. All staff (staff with or without the 3.5-year pedagogy bachelor degree, which is obligatory to become a certified ECE teacher in Denmark) in the childcare centers were introduced to the intervention in a 2-day professional development course by the local educational consultant. As is the case in many intervention studies, the main elements in the introduction course were sharing information about the intervention and incorporating modeling of central intervention strategies (Biel et al., 2020). The course included background knowledge and explicitly described and illustrated the targeted domains, how to reference these explicitly during activities using specific practices to enrich and differentially support children. The course, moreover, provided opportunities to actively apply and generalize the learned content using video-recordings; and supporting metacognition (e.g., reflection and self-monitoring) throughout the course. These strategies have been shown, along with the adult's active involvement in the learning process, to promote learning in teachers (Bier et al., 2019). The training also provided teachers with all materials necessary to fully implement the intervention.

Description of business-as-usual classrooms

As we use business-as-usual classrooms as a control group, as mentioned above, we provide a brief introduction to the Danish ECE context. Denmark has a universal daycare system which is heavily subsidized and on average 90% of 1- and 2-year-old children are enrolled in childcare (Ministry for Children & Social affairs, 2018a). With regard to structural characteristics, the teacher-child ratio is relatively low (1 teacher to about 3.5 children in center-based care and 1 teacher to about 4–5 children in family-based care). In center-based care, approximately 60% of teachers

have a 3.5-year pedagogical bachelor degree, whereas in family-based care only a minority (8%) has a formal educational background (Ministry of Children & Social Affairs, 2018b). In both cases, some local variation can be found. The ECE area is regulated by the "childcare legislation" [Dagtilbudsloven], which seeks to promote coherence and continuity between the childcare centers resulting in rather uniform centers in terms of content. As part of the childcare legislation, a broad learning curriculum was implemented in 2004, which has an explicit focus on strengthening the educational learning environment throughout the day focusing on topics and content cover the 6 curriculum themes (e.g., personal and social-emotional development, language, math and science). The legislation is aimed at a broad concept of learning through free play, creativity, and outdoor activities within a social and inclusive context (Bauchmüller, Gørtz, & Rasmussen, 2014), which is in accordance with the Danish ECE tradition. Per this tradition, there is a strong and common belief among teachers that learning occurs in social interactions and in play situations rather than in structured instructional situations, such as circle time or academic activities (Broström, Johansson, Sandberg, & Frøkjær, 2012).

Measures

Child outcomes

To measure child outcomes, we used 1 standardized test and 3 teacher-report instruments that were completed at pretest (before the intervention) and posttest, approximately 7 months after the intervention was initiated. Each of the teacher-reported instruments took on average about 8–10 minutes to complete, thus teachers spent about 30 minutes per child completing the instruments. The standardized test was administered by staff, who received extensive training in reliable administration.

For *language,* we used 2 different measures. To test receptive vocabulary, a test of Danish receptive vocabulary was adapted from the English Computerized Comprehension Task (CCT; Friend & Keplinger, 2008; Friend, Schmitt, & Simpson, 2012). The original CCT consists of 41 items and was developed for 16–30-month-olds. On each trial, the child sees 2 pictures denoting nouns, verbs, or adjectives on a touch screen. One picture denotes the target word (e.g., dog), and 1 picture is a foil. The test administrator asks the child to "Touch the dog!" The result is scored as correct, incorrect, or no response. Correct responses were reinforced by a brief target-related sound, for example, Bow-wow! for "dog." The original CCT contains relatively easy, medium, and hard items. For the purpose of the present study, we needed a test with harder items to be able to test up to the age of 35 months without ceiling effects. In order not to increase the length of the test beyond 2-year-olds' typical window of attention, a decision was made to develop 3 age-related difficulty levels for the ages 18–23, 24–29, and 30–35 months. Each of the 3 age-dependent tests consisted of items ranging from easy to hard, the assumed item difficulty being based partly on items in the original English CCT, partly on results from the Danish CDI-study of productive vocabulary (Bleses et al., 2008), and partly on pilot tests. The interface for the Danish adaptation were iPads rather than touch screens. For the final analyses, we used the sum of correct answers.

To assess the productive vocabulary and language use, we administered a brief teacher-based standardized checklist, the CDI-Educator (Bleses, Jensen, Højen & Dale, 2018c). The CDI-Educator is based on well-developed and validated parent report measures (MacArthur-Bates Communicative Development Inventories; Fenson et al., 2007) and a Danish adaptation of a short version of the instrument (Vach, Bleses, & Jørgensen, 2010) was adapted for the early childhood education setting. CDI-Educator has a 70-item vocabulary checklist with 9 categories of content (sound effects and animal sounds, animals and things, food and drink, body parts, small household items, furniture and rooms and places (to go), people and routines, action words, descriptive words, and particles). Additionally, the checklist includes 5 questions concerning the child's use of decontextualized language with respect to objects and actions distant from the here and now (e.g., whether the child at any time speaks about earlier episodes and persons, who are not present or about something that will happen in the future). For our final analyses, the vocabulary summary score was calculated adding up the number of words the child could produce. For the language use summary score, the response categories were converted to points and summed up across the 5 questions ("not yet" [0], "sometimes" [1] or "often" [2]). The CDI-Educator has been standardized on a total of 5097 children aged 18–34 months (cf. Bleses et al., 2018c). Test-retest correlations (0.68 correlation for vocabulary and 0.54 for language use) and internal consistency measures (Cronbach's alpha: 0.98 for vocabulary and 0.88 for language use) demonstrate reliability. External validation was established through the Danish Receptive and Productive One-Word Picture Vocabulary Test (r = between *0.*43 and 0.65 for vocabulary and language use). Measurement properties of the instrument is reported in a validation study (Bleses et al., 2018c).

For *math language* and *numeracy*, we used a researcher-developed teacher-administered checklist as no standardized measure for toddlers was available at the time. The checklist evaluates 2 dimensions of early math development, that is, children's comprehension and use of math language (numeracy, 10 items). The response categories were converted to points and summed up across the relevant questions ("not yet" 0], "sometimes" [1], "often" [2] or "always" [3]). This instrument included items that were more proximal to the intervention targets. Internal consistency was assessed for math language and numeracy; all items had an item-total correlation exceeding 0.50. Cronbach's alpha was high for both math language and numeracy, around 0.95, demonstrating high internal consistency. Correlations with the 2v scores of the CDI-Educator (locabulary, anguage use) were also substantial (r = 0.60 and 0.73), demonstrating that the math checklist also exhibits good external concurrent validity (see Bleses, Jensen, Slot, & Justice, 2020).

For*social-emotional skills*, the Danish adaptation (Sjoe et al., 2020; Sjö 2019) of the standardized questionnaire, *Social-Emotional Assessment/Evaluation Measure (SEAM)–Research Edition* (Squires, 2014) was applied. SEAM has 10 domains critical to social-emotional skills: empathy, healthy interactions, expression of emotions, regulation of social-emotional responses, cooperation, sharing and engaging, regulation of attention and activity level, independence, self-image, and adaptive skills. Each of the benchmarks has 4 response categories: 0 = *not true*, 1 = *rarely true*, 2 = *somewhat true*, and 3 = *very true* (a high score indicates a positive aspect of social-emotional development). Based on Rasch analyses, the 10 benchmarks were converted to 2 overall indices: (1) The empathy index, assessing the child's ability to communicate own feelings and to read and understand others' feelings; and (2) The self-regulation & cooperation index, assessing the child's ability to regulate and cooperate, and the child's adaptability (the reliability coefficients range from 0.82 to 0.91, see Sjö et al., 2019).

Intervention fidelity

Two types of information were used to track intervention fidelity: (1) Ongoing completion of weekly implementation logs on a web-based platform completed by the teachers in each classroom (per classroom) or family-based care and; (2) Observational ratings. The implementation notes were completed by each classroom on a weekly basis. The teachers in each classroom were advised to reflect together on each of the questions in the implementation notes. The teachers had to log on themselves to the IT platform

to complete the notes. The implementation notes also documented child exposure, defined as the number of times each child participated in large and small group activities per week and individual conversations. We operationalize intervention fidelity (dosage of implementation) as the total number of activities (large group, small group, individual conversations) across the intervention period which teachers engage each individual child in. We use this measure to test how intervention exposure is associated with child outcomes.

Teachers also reported the adherence to the intervention, defined as the extent to which the intervention components were implemented: What type of activities in large and small group settings was carried out in the classroom that week (it was possible to mark activities already listed or add other activities), the extent to which they used the provided intervention material (there was a list of all the intervention materials and the teachers marked which they have used), and the extent to which the teachers in the classroom judged that they were addressing the learning objectives in interactions with the children ("not at all" [1], "sometimes" [2], "often" [3]). These self-reported ratings of the adherence to the intervention are only used descriptively.

Observational ratings were applied to a subset of childcare centers to assess adherence to the critical elements of the intervention and participants' responsiveness based on video recordings in the middle of the 20-week program. Classrooms were selected so that municipalities, type of childcare (center-based vs family-based childcare, and condition (intervention vs control classroom) were equally represented.

In each classroom, a video of a large-group and a small group activity was collected by research staff. In terms of focus, 58% of activities had a focus on language, with the remaining focused on math language and numeracy. Research staff was blinded to the condition and coded the videos using a checklist developed to observe teacher's adherence to 5 different intervention elements. The first was the use of sequence and scope. Videos were coded for: (1) how often the teachers mentioned the learning target of the particular week in which the video recording took place (1–5+ times); (2) how many times the teachers used each of the 7 responsive strategies from "The high-quality conversation" (1–5+ times); (3) how many times the teachers used the 5 word learning strategies from "Learning new words" (1–5+ times); and (4) how many times the teachers used the 6 low and high support strategies from "The Learners Ladder" (1–5+ times); and (5) how many times the teachers used strategies specific to language modeling ("low" [1], "mid" [2], "high" [3]). In addition, the level of child engagement was coded. ("low" [1], "mid" [2], "high" [3]). These observational assessments of the adherence to the intervention are only used descriptively. Inter-rater reliability was not assessed.

Moreover, the implementation fidelity of the local professional development course in each municipality was evaluated using self-reporting and observation. The quality of the local course was addressed in a survey completed by the municipality consultants and included questions about the extent to which the consultants completed all elements of the course as intended, the extent to which they were able to relate theory and practice, the degree to which the teachers understood the content and whether they were motivated and engaged in the intervention. Direct observations of a subset of the local courses were made using a protocol that was developed to assess fidelity of the course and the engagement of the educational consultants as well as teachers.

Child progress checklists

Teachers tracked children's individual progress toward the learning objective within each of the learning domains in language (general vocabulary and language use) and math (math vocabulary and numeracy skills) using an informal assessment indication of whether the child "never," "sometimes," or "often" demonstrated the skill. The progress checklist, which was available at the IT-platform, was completed once for each theme for all individual children (that is 5 times) . The teachers were introduced to the progress checklists during the professional development course. Moreover, the completion of implementation notesand progress of children was discussed during meetings in the implementation teams.

Analytic strategy

To estimate the effects of the intervention compared with the business-as-usual control group, we use a value-added specification of a regression model specified as:

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \delta I_i + \gamma y_i^{pre} + u_i,$$

with y_i as the post-test outcome measure for child i and y_i^{pre} as the pretest outcome measure. Including this measure allows us to control for unobserved characteristics that might be correlated with prior achievement. $\boldsymbol{X}$ is a vector of explanatory variables that include child and family characteristics (gender, age and ethnicity of the child, parental education level, parents' marital status and employment status). These variables are included to control for individual differences that might influence the outcome measures and to increase the statistical precision of the estimates. We also included municipality fixed effects. I_i is an indicator variable for whether the child is enrolled in an intervention center, which makes δ our main parameter of interest. Outcomes are normalized with the standard deviations of the relevant pretest for the control group, which means that estimates of δ can be interpreted as effect sizes (Cohen's d).

Since the randomization occurs at the childcare level, we adjust the standard errors of the estimates to take account of this clustering and the hierarchical structure of the dataset. The cluster-adjusted standard errors also account for any other non-independence of children within childcare centers. At the childcare center level, intraclass correlations (ICCs) ranged from 0.03 for language use to 0.08 for receptive vocabulary and at the classroom/family-based unit level, ranged from 0.16 for receptive vocabulary to 0.25 for math language. These ICC values indicate that the majority of variance was between children, with fractions ranging from 0.75 on math language to 0.84 on receptive vocabulary. In general, we analyzed the data at the child level with childcare centers as the clustering unit, since the intervention was implemented at the childcare level (including possible spillover effects between classrooms and family-based caregivers, respectively). However, we did run a number of sensitivity checks to investigate if it made any difference for the results if we also took account of clustering at a lower level (i.e., classrooms or family-based caregivers), either by adjusting standard errors or by using multilevel models (HLM). This did not result in changes to the results reported below (at most, the changes in the magnitude of estimated effects were marginal) for any of the analyses. We estimated separate models for each child outcome. Missing data was handled by the maximum likelihood method (FIML).

Results

Main effects of "We learn together" on children's developmental outcomes

The first research question examined the main effects of the intervention "We learn together" on children's language, math, and social skill in center-based and family-based childcare centers after 1 period of exposure to the intervention. We estimated separate models for each of the 7 different child outcomes for the

Table 4
Estimated treatment effects of the intervention (effect sizes).

Outcome	Receptive vocabulary	Productive vocabulary	Language use	Math language	Numeracy	Empathy	Self-regulation
	0.13**	0.10***	0.03	0.59***	0.34***	0.10***	0.06
	(0.05)	(0.03)	(0.04)	(0.07)	(0.07)	(0.04)	(0.05)
N	1818	2170	2170	2159	2159	2039	2039

Note: Standard errors in parentheses. Standard errors are adjusted for clustering at center/group level.
Missing values are handled by using the ML method (FIML). All estimates are from separate value-added models with covariates and municipality fixed effects included. The effect sizes are in terms of standard deviations of the pre-test for the control group.
** $P < 0.05$.
*** $P < 0.01$.

overall sample (recall that receptive vocabulary is measured with the CCT test, productive vocabulary and language use are measured with CDI-Educator, social-emotional skills are measured with SEAM and math language and numeracy skills are measured with the researcher-developed math checklist). These results are displayed in Table 4. As can be seen from the table, the overall effects were significantly positive for the targeted skills, such as receptive (d = 0.13) and productive vocabulary (d = 0.10), math language (d = 0.59) and numeracy (d = 0.34), while no significant effects were found for language use. The strongest effects were found for the 2 math outcomes. Interestingly, we found significant effects on empathy (d = 0.10) but not on self-regulation and cooperation. To test whether the effects were different for the 2 types of care, we included interaction terms between type of care and the intervention in the models (results not shown). There were no significant differences in effects between the 2 types of care (with P values ranging from 0.34 to 0.97) for the 7 outcome measures.

Moderation effects of child and parent characteristics

For the second research question, which concerned whether the treatment effects were moderated by child and parent characteristics, we estimated models that included interaction terms between various child and parent characteristics and the intervention in the models; each interaction term was added in a separate model. The child characteristics examined included gender and age and the parent characteristics included ethnic background, parents' education, parental cohabitation status, parents' employment status, and parental housing status. Neither pretest scores nor child or parent characteristics moderated the treatment effects, that is, no differential effects could be established. The 1 exception occurred when comparing children with employed parents with children with parents on public transfer income or with no income (the latter group consisting of 251 children), for which there was 1 significant interaction for math language ($b = -0.34$, $SE = 0.17$, $P < 0.05$).

Intervention fidelity of the intervention

The third research question addresses the fidelity of "We learn together" when implemented in a real-world setting. The recommended level of activities for each child each week was participation in 2 large group activities, 2 small group activities and 1 individual conversation with a teacher. As can be seen in Table 5, our results show that on average teachers met the expected level of engaging children in activities but with large variation across children (the analysis is based on completed implementation notes, which reported individual child exposure on a weekly basis (N = 11,505 of 13,960, corresponding to 82%). For the weekly recommendation of 2 large-group activities, each child participated in an average of 2.7 activities (SD = 1.5; interdecile range 1.1–4.4). For the weekly recommendation of 2 small-group activities, each child participated in an average of 2.0 activities (SD = 1.3; interdecile range 0.8 to 3.5). Finally, for the weekly recommendation of 1 individual conversation, individual children experienced an average of 6 (SD = 4.5; interdecile range 1.4–12) individual conversations with their teachers each week. Based on implementation notes, 70% of teachers noted that they established exploration zones very frequently and that the children used these to initiate playful activities.

Table 5
Exposure of children to activities, mean per week (N = 1574 children).

	Mean	SD	10th percentile	90th percentile
Large group activities	2.7	1.5	1.1	4.4
Small group activities	2.0	1.3	0.8	3.5
Individual conversations	6.0	4.5	1.4	12.0

Notes: The table only includes children in the treatment group, as the exposure for children in the control group per definition is equal to 0. The mean per week is calculated for each child as the total exposure divided by the duration of the intervention period (i.e., 20 weeks). The total exposure is the total number of activities in each category over the full intervention period (for the first period). Since not all children have data on all 20 weeks in the intervention period, the calculated mean per week may be downward biased. If the missing data reflect that the child did not participate or that no activities took place in those weeks, then the calculated mean per week is not downward biased.

We examined whether the large variation in exposure to "We learn together" was related to child and parent characteristics (the child's gender, age, ethnic background, parental education level, parental cohabitation status, parents' employment status and parental housing status) but did not find any significant differences at the child level. We also investigated the partial correlations (controlling for childcare type) between childcare characteristics (aggregated measures of teacher age, teacher educational level and experience) and an aggregated measure of exposure across children by regressing child exposure on the childcare characteristics (N = 156 childcare centers of 183, corresponding to 85%). We found no significant correlations between child exposure and percentage of teachers with a BA education ($t(153) = -0.12$, P = 0.91) or teachers experience ($t(153) = 1.26$, $P = 0.21$). In contrast, we found a significantly positive correlation between child exposure and an aggregated measure of teacher age in the childcare centers ($t(153) = 2.14$, $P = 0.03$), indicating that older teachers implemented more activities than younger teachers.

Based on the self-reported implementation notes, it appears that in large and small group activities (which was chosen by the teachers themselves), the most frequent types of activities that the teachers engage children in were shared book reading, singing activities and nursery rhymes, games with the picture cards, and creative activities, that is, very typical activities for infant-toddlers. The intervention materials were, on average, used in between 60% and 75% of weekly activities, even though it was completely voluntary to use the materials. In particular, the teachers used the provided picture cards of target words. Moreover, 41% of the teachers indicated that they addressed the specific learning targets of the intervention to a very high extent, approximately half of the teachers (49%) indicated that they to some extent addressed the specific targets where less than 10% indicate that they either not or to a

Table 6
Estimated treatment effects of the intervention (effect sizes), by treatment exposure.

Outcome	Receptive vocabulary	Productive vocabulary	Language use	Math language	Numeracy	Empathy	Self-regulation
High exposure	0.18***	0.16***	0.07	0.75***	0.48***	0.12***	0.11*
	(0.06)	(0.04)	(0.05)	(0.10)	(0.07)	(0.04)	(0.06)
Low exposure	0.03	−0.04	−0.07	0.33***	0.01	0.04	−0.05
	(0.07)	(0.05)	(0.05)	(0.10)	(0.10)	(0.05)	(0.06)
Difference	0.15***	0.20***	0.14***	0.42***	0.47***	0.08*	0.16**
	(0.06)	(0.05)	(0.05)	(0.07)	(0.08)	(0.04)	(0.05)
N	1818	2170	2170	2159	2146	2039	2039

Notes: Standard errors in parentheses. Standard errors are adjusted for clustering at center/group level.
Missing values are handled by using the ML method (FIML). Each column contains estimates from a separate value-added model with covariates and municipality fixed effects included. The effect sizes are in terms of standard deviations of the pre-test for the control group.
* $P < 0.10$.
** $P < 0.05$.
*** $P < 0.01$.

limited extent addressed the learning targets. Measures of adherence to the content of the intervention were coded from videos collected in each classroom mid-way through the implementation period of a smaller part of the sample (N = 106 video recordings). Analyses showed that treatment teachers used the targeted learning objectives and responsive strategies (on average 4.5 vs 3.0 times per activity) and word learning strategies (on average 4.4 vs 0.2 times per activity) to a higher extent than teachers in the control condition did (on average 4.5 vs 3.0 times). Neither teachers in the treatment or in the control groups used the Learner's Ladder often (on average 0.8 vs 0.2 times).

As the implementation of the local professional development course is a significant element of taking interventions to-scale, we also examined the extent to which this course was implemented as intended. The local professional development course was provided by 19 local educational consultants who all participated in the train-the-trainer course. All educational consultants reported high compliance with the structure and content of the local professional development course; the consultants reported that they to a very high (42%) or a high degree (58%) carried out the course as intended. Additionally, they reported high level of motivation and engagement of the teachers (73%). Between 60% and 75% reported that the structure and content of the local professional development course was clear and understandable for the teachers and supported learning. The direct observation of a limited number of local professional development courses (N = 9) supported the trainers' self-evaluations. One exception was related to informing participants about the study design and the use of real-life data where the compliance was somewhat lower in half of the courses. The overall conclusion is, therefore, that the educational consultants were able to conduct the professional development course in a way that was close to the intention and seemingly prepared the teachers satisfactory for the "We learn together" intervention.

Moderation effects of intervention exposure

In the fourth research question, we investigated whether individual variation in intervention exposure was associated with child outcomes. We did this by estimating the effects of the intervention in models where we had added measures of exposure. The analysis is based on completed implementation notes that reported individual child exposure on a weekly basis (N = 11,505 of 13,960, corresponding to 82%). To obtain robust estimations, we summed all activities regardless of type and split the sample by the median number of activities. Hence, our measure of exposure only distinguished between low and high exposure. The results are shown in Table 6, which contains the estimated treatment effects for each of the 2 levels of exposure and the difference between them.

As can be seen from the differences in the table, high exposure was associated with significantly higher outcomes than low exposure, except for empathy (the differences range from $d = 0.08$ to $d = 0.48$, depending on outcome). Even for those outcomes where there were no overall intervention effects (language use and self-regulation & cooperation), the differences between high and low exposure showed a significant advantage for children with high exposure ($d = 0.14$ and $d = 0.16$, respectively). In sum, this reflected the same pattern for the group of children with high exposure as the overall effects showed, whereas for children with low exposure, there were no significant effects except for math language. For this outcome, there was still a significantly positive effect with low exposure, but the magnitude was only half the effect size of high exposure. It should be acknowledged that all results for exposure are only correlational, as differences in exposure were not part of the experimental design.

Effects of implementation teams on fidelity

Finally, in accordance with the fifth research question, we investigated the association between implementation teams and intervention fidelity. Based on the completed implementation notes (N = 993 of 1445, corresponding to 67%), we found that on average 2.6 ($SD = 1.5$) of the 5 planned implementation team meetings were carried out but with substantial variation between childcare centers (range 0–5). Correlating the number of implementation team meetings and intervention fidelity (aggregated child exposure within childcares), there was generally no associations between intervention fidelity as measured by child exposure (we found a marginally significant correlation with more than 1 team meeting (vs 1 or none) and child exposure ($t(260)=1.84$, $P = 0.07$)).

Discussion

The present study adds to the emerging literature concerning the scaling up of early interventions in childcare and is a unique contribution to the literature given the focus on the effectiveness of a schoolreadiness intervention for toddlers implemented at scale and tested on a heterogeneous and representative sample. Given the importance of early intervention to mitigate early skills gaps, this study contributes highly valuable knowledge of effective intervention components and what works for whom as well as knowledge about how interventions are implemented at scale, how intervention dosage affects child outcomes and how measures to improve intervention fidelity under real-world circumstances work. We highlight several major contributions of the study.

The study has meaningful and policy relevant main effects when implemented at-scale

Based on the early-stage study of the intervention evaluated in this paper (Bleses, Jensen, Slot, & Justice, 2020), we hypothesized that the implementation of "We learn together" would result in positive main effects on language and math outcomes and potentially also social outcomes. The most important contribution of this study is that we demonstrated that it was indeed possible to replicate the findings from the first stage efficacy trial of "We learn together" and achieve population-level short-term effects at scale for 5 out of 7 outcomes. With the exception of language use, all targeted language and math outcomes increased significantly more in the intervention group than the control group. Moreover, a significant spill-over effect on empathy was found, conceivably because the children, as a function of the intervention, participated in, on average, 2 weekly small-group activities, which can support peer interactions to a higher extent than large-group activities. Morever, the focus on words for feelings may also have supported children's ability to communicate their own feelings and to read and understand others' feelings. It is important to note that in contradiction to the early-stage trial, where the assessment of child outcomes relied entirely on teacher report instruments, we included a standardized test of receptive vocabulary, which indicated slightly higher effects than for the teacher reported domains. Compared to the early-stage trial, the intervention effects of the current at-scale study were somewhat reduced (between 0.1 and 0.5 of a standard deviation). This may be surprising as we found that the implementation exposure was higher in the current study compared to the first stage trial (see below).

We speculate that there may be several different sources of this somewhat surprising result. Even though the early-stage trial already was an effectiveness study, we did introduce changes when taking the intervention to scale, which might have affected child outcomes. First, there was less self-selection of childcare centers, as in 6 municipalities all childcare centers participated. Second, and perhaps because of less self-selection, children in the current samples had lower pretest scores compared to children in the early-stage study. Third, the train-the-trainer model, entailing that the local professional development course was provided by educational consultants rather than scientific staff, may have caused lower quality in terms of teaching the content of the intervention.

Nevertheless, the effect sizes still ranged from 0.10 to 0.59 which, according to a newer benchmarking model developed by Kraft (2020), can be characterized as medium and large-sized effect sizes. Following Kraft's additional criteria for evaluating impact, it is of note that this intervention additionally (1) was evaluated in a large-scale sample; (2) is a low-cost intervention, which was implemented without additional resources from the municipalities and (3) already is designed for scaling to nationwide use. Consequently, the present findings suggest that the intervention has potentials for large impact at a population level. The results of the intervention are, based on Kraft's benchmarks, highly meaningful and relevant for policy makers.

Child and parent characteristics do not moderate effects

We were able to identify for whom the intervention worked by comparing intervention effects for different groups of children. As in the early-stage project, we expected limited effects of child and parent characteristics on intervention effects and this hypothesis was confirmed. Using the rich available register data, we found that child characteristics, such as gender and age, and parent characteristics, such as ethnicity, education, and employment status did not significantly affect treatment effects, that is, no differential effects could be established. Consequently, compared to the early-stage trial, we no longer find significantly lower treatment effects for children of mothers with no or low education or for immigrant children. This finding is similar to the effectiveness studies of language interventions in Danish preschools, which were also tested in heterogenous samples (Bleses et al., 2018a, 2018b).

The intervention fidelity was acceptable even at at-scale implementation

Based on results from other large-scale RCTs in a Danish context (Bleses et al., 2018a; 2018b), we hypothesized, that not all childcares would meet the recommended level of intervention fidelity, that is, not all children were exposed to the recommended number of activities. However, the results of the study revealed that teachers met, on average, the recommended level for child exposure in the intervention, though substantial variation was found across children. Furthermore, the variation in exposure across children was unrelated to child or parent characteristics indicating that the teachers were not implicitly biased in terms of including children in the intervention activities, possibly as a result of a more systematic focus on an inclusive practice in the professional development course. In fact, the intervention fidelity was higher compared to the first stage trial: almost 1 additional group activity (on average 2.7 large group activities vs 1.8 in the first trial and on average 2.0 small group activity vs 1.1 in the first trial), but in particular the number of individual conversations increased substantially (on average 6.0 vs 1 in the first trial). The only childcare level characteristic that predicted child exposure was teacher age, such that older teachers resulted in higher exposure. This suggest that an important factor underlying the effectiveness of the "We learn together" intervention, is that it was feasible to design an intervention for toddler childcares that could be implemented at scale at a satisfactory level.

The satisfactory level of intervention fidelity may be related to several factors. One of the design characteristics of the intervention, that is, providing teachers with discretion and flexibility in the implementation of the intervention, may have contributed to the satisfactory intervention fidelity. Even though the intervention has some clearly identified core components that are empirically validated (e.g., sequence and scope, responsive strategies), the implementation is designed to flexibly match various local contexts and child populations as teachers themselves were asked to develop activities of their own choice, which may have supported the implementation flexibility. The lack of prepared lessons and scripting seemingly did not prevent teachers to pick up the curriculum as hypothesized by Weiland et al. Weiland, McCormick, Mattera, Maier, & Morris, 2018. On the other hand, the lack of scripting might have caused more implementation variability across different activities, but we have no data to investigate whether this was the case. Furthermore, the measures taken when preparing the "We learn together" intervention presumable enhanced the implementation of the intervention.

Suggestively, the on average high level of intervention fidelity indicate that the program strategies which were more organizational in nature were successful in making the intervention ready for scaling up. The train-the-trainer model was evaluated very positively by the local municipality consultants who believed it was possible for them to integrate the intervention efficiently in the local educational context to a higher extent compared to researcher-held courses and this local embedment may have encouraged the teachers' implementation of the intervention. A qualitative evaluation among the educational consultants suggests that the initial readiness workshop focusing on implementation readiness in each municipality provided them with important knowledge about strengths and difficulties of implementing this specific intervention in their local childcare centers, which they could target in the lo-

cal professional development course. Along the same line, the organizational support that was provided as part of the intervention study, including the inclusion of principals in the professional development course, may have supported the day-to-day implementation. Principals completed a survey post the intervention, which suggested that many participated in classroom planning meetings and gave informal coaching, both at meetings and when teachers implemented the intervention focusing on the teachers' learning and the use of rich language and responsive strategies.

Positive child outcomes are associated with intervention exposure

An additional contribution of this study is, that we examined how intervention fidelity is related to child outcomes. We hypothesized based on earlier Danish RCT studies that higher exposure would be associated with higher outcomes and this hypothesis was confirmed. Even though we found that teachers, on average, engaged children in activities at the recommended level there was substantial differences in exposure at the level of the individual child and this difference in dosage was associated with child outcomes. In fact, we found no associations with child outcomes unless the recommended dosage was met, indicating that this dosage is necessary to be able to expect effects. Furthermore, for children who were exposed to the recommended level of exposure, we found significant associations not only for all targeted skills except for language use but also for both the untargeted domains, further supporting the effectiveness of this toddler intervention even under at scale implementation. Results of the present work suggest that variation in implementation dosage is highly associated with individual children's benefits of the intervention. It is, therefore, essential that we learn more about individual teacher and system level factors that affect intervention fidelity in order to improve fidelity under real-life circumstances.

Implementation teams did not improve intervention fidelity

One initiative identified as being important for implementation fidelity that was implemented in the current study was the so-called implementation teams, which were supposed to have 1 meeting per each theme (i.e., every fourth week). We predicted that the use of implementation teams would improve fidelity. However, the findings did not clearly support this prediction; in that we only found a marginally significant association between use of implementation teams and child exposure. We speculate that there are several reasons for the lack of associations with use of implementation teams. First of all, it was difficult for the classrooms to complete the implementation team meetings as, on average, only half of the meetings were carried out. A subsequent evaluation indicated that structural conditions, in particular lack of preparation time and sick listing have had an impact on the frequency of meetings. Second, many teachers were not familiar with using on-line data and the head teachers noted in an evaluation that the training they were provided with did not prepare them sufficiently to use these data to support implementation. Nevertheless, the main conclusion from a subsequent evaluation was that teachers experienced that the implementation team meetings had increased knowledge sharing and exchange of experience, which resulted in a deeper understanding of the "We learn together" intervention and a greater awareness of how to use the language strategies with individual children. The implementation teams created a space for reflecting on and evaluating the instructional practice, which were conceived as very constructive and inspiring. Looking ahead, the use of implementation teams in relation to interventions are promising but more attention to structural conditions and training of lead teachers is necessary.

Limitations

The findings of this study importantly advance our understanding of scaling up interventions. However, several limitations should be noted. First, for efficiency purposes – and the lack of available appropriate standardized tests for this age group and general problems testing these domains in toddlers - we mainly used published teacher reported standardized measures, which have demonstrated high internal reliability and good or acceptable external validity concurrent with standardized test. Although, there are concerns about the validity of providing estimates of children's development using teacher ratings, it is a common methodology in large-scale research involving the assessment of thousands of children (e.g., National Center for Early Development and Learning's Multi-State Study of Pre-Kindergarten, National Center for Early Development and Learning, National Institute for Early Education Research State-Wide Early Education Programs Study, and the Twins Early Development Study). A number of studies have shown that teacher reports are valid and reliable measures of children's behavioral (Bishop et al., 2003) and academic skills (National Center for Education Statistics [NCES], 2002; Justice et al., 2009). Similarly, again for reasons of efficiency, we use a self-reported measure of intervention fidelity to capture dosage and adherence, which may have limited our possibility to measure full associations with child outcomes. However, the differential patters of associations indicate that this measure overall worked well as a rough indicator of intervention fidelity. Note though that use of a standardized assessment resulted in slightly higher effects suggesting that using teacher-ratings might have underestimated the effects. Second, long-term follow-up data are not available, and thus it is unknown if the intervention results will contribute to longer-term skill improvement. Finally, even though the results of the early-stage study were replicated, it remains unclear whether the intervention effects would successfully generalize to other contexts like the US, as the high level of teacher education and higher teacher-child ratios in Denmark may have affected the implementation positively.

Conclusion

In conclusion, the present study is one of the first to evaluate the effects of a toddler schoolreadiness intervention when implemented at scale, and as a whole the study confirms that the low-cost "We learn together" intervention is ready to be scaled up. An inexpensive program that produces medium-to-large effect sizes and therefore economically valuable outcomes, has the potentials for making good policy. The study has provided some knowledge about the hypothesized causal mechanisms underlying the interventions effects via analyses of moderators but future research is needed to fully unpack the black box of the "We learn together" intervention to get even closer to the effective components. The study also demonstrated that taking effective interventions to scale involves challenges. Based on an analysis of implementation of interventions in 5 public systems (behavioral health, child welfare, education, juvenile justice and public health), Fagan and colleagues (Fagan, 2019) find several factors that affects scale-up across systems from higher system level factors as public awareness and support of effective interventions like "We learn together" to lower-level factors like the skills of the workforce. Such barriers of implementation also need to be addressed in future studies to improve fidelity and child outcomes.

Authors' contributions

Dorthe Bleses: Conceptualization, methodology, writing – original draft preparation, project administration, funding acquisition; Peter Jensen: Methodology, data curation, formal analysis, writing

– original draft preparation. Anders Højen: Conceptualization, writing – reviewing and editing. Pauline Slot: Conceptualization, writing – reviewing and editing. Laura Justice: Conceptualization, writing – reviewing and editing.

Acknowledgments

This research was supported from a grant by from the Ministry of Children and Education and TrygFonden. We wish to thanks the participating municipalities, teachers, and children.

References

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., & Yeager, D. S. (2020). Persistence and fade-out of educational-intervention effects: mechanisms and potential solutions. *Psychological Science in the Public Interest, 21*, 55–97. https://doi.org/10.1016/170/.1157279/1502096120096210598145848.

Bauchmüller, R., Gørtz, M., & Rasmussen, A. W. (2014). Long-run benefits from universal high-quality preschooling. *Early Childhood Research Quarterly, 29*, 457–470. https://doi.org/10.1016/j.ecresq.2014.05.009.

Beck, I. L., & McKeown, M. G. (2007). Increasing young low-income children's oral vocabulary repertoires through rich and focused instruction. *The Elementary School Journal, 107*, 251–271. https://doi.org/10.1016/j.ecresq.2017.07.005.

Biel, C. H., Buzhardt, J., Brown, J. A., Romano, M. K., Lorio, C. M., Windsor, K. S., & Haghish, E. (2020). Language interventions taught to caregivers in homes and classrooms: a review of intervention and implementation fidelity. *Early Childhood Research Quarterly, 50*, 140–156. https://doi.org/10.1016/j.ecresq.2018.12.002.

Bleses, D., Højen, A., Dale, P. S., Justice, L. M., Dybdal, L., Piasta, S., & Haghish, E. (2018a). Effective language and literacy instruction: evaluating the importance of scripting and group size components. *Early Childhood Research Quarterly, 42*, 256–269.

Bleses, D., Højen, A., Justice, L. M., Dale, P. S., Dybdal, L., Piasta, S. B., & Haghish, E. (2018b). The effectiveness of a large-scale language and preliteracy intervention: The SPELL randomized controlled trial in Denmark. *Child Development, 89*, e342–e363.

Bleses, D., Jensen, P., Højen, A., & Dale, P. S. (2018c). An educator-administered measure of language development in young children. *Infant Behavior and Development, 52*, 104–113.

Bleses, D., Jensen, P., Slot, P., & Justice, L. (2020). Low-cost teacher-implemented intervention improves toddlers' language and math skills. *Early Childhood Research Quarterly, 53*, 64–76. https://doi.org/10.1016/j.ecresq.2020.03.001.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T., & Basbøll, H. (2008). The Danish communicative development inventories: validity and main developmental trends. *Journal of Child Language, 35*, 651–669. https://doi.org/10.1017/S0305000907008574.

Bornstein, M. H., Hahn, C. S., & Putnick, D. L. (2016). Stability of core language skill across the first decade of life in children at biological and social risk. *Journal of Child Psychology and Psychiatry, 57*, 1434–1443. https://doi.org/10.1111/jcpp.12632.

Broström, S., Johansson, I., Sandberg, A., & Frøkjær, T. (2012). Preschool teachers' view on learning in preschool in Sweden and Denmark. *European Early Childhood Education Research Journal, 22*, 590–603. https://doi.org/10.1080/1350293X.2012.746199.

Burchinal, M., Magnuson, K., Powell, D., & Hong, S. S. (2015). Early childcare and education. In *Handbook of child psychology and developmental science* (pp. 1–45).

Cameron, C. E., Kim, H., Duncan, R. J., Becker, D. R., & McClelland, M. M. (2019). Bidirectional and co-developing associations of cognitive, mathematics, and literacy skills during kindergarten. *Journal of Applied Developmental Psychology, 62*, 135–144. https://doi.org/10.1016/j.appdev.2019.02.004.

Chambers, B., Cheung, A. C., & Slavin, R. E. (2016). Literacy and language outcomes of comprehensive and developmental-constructivist approaches to early childhood education: a systematic review. *Educational Research Review, 18*, 88–111 July 2nd 2020. https://doi.org/10.1016/j.edurev.2016.03.003.

Clearinghouse, W. W. (2017). *Standards handbook (version 4.0)*. Washington, DC: Institute of Education Sciences Retrieved from https://ies.ed.gov/ncee/wwc/handbooks .

Clements, D. H., & Sarama, J. (2011). Early childhood mathematics intervention. *Science (New York, N.Y.), 333*, 968–970. https://doi.org/10.1126/science.1204537.

Crawford, A., Zucker, T., Van Horne, B., & Landry, S. (2017). Integrating professional development content and formative assessment with the coaching process: the Texas school ready model. *Theory Into Practice, 56*, 56–65. https://doi.org/10.1080/00405841.2016.1241945.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., & Brooks-Gunn, J. (2007). School readiness and later achievement. *Developmental psychology, 43*, 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428.

Duncan, G. J., & Magnuson, K. (2013). Investing in preschool programs. *The Journal of Economic Perspectives, 27*, 109–132. https://doi.org/10.1257/jep.27.2.109.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American journal of community psychology, 41*, 327. https://doi.org/10.1007/s10464-008-9165-0.

Fagan, A. A., Bumbarger, B. K., Barth, R. P., Bradshaw, C. P., Cooper, B. R., Supplee, L. H., & Walker, D. K. (2019). Scaling up evidence-based interventions in US public systems to prevent behavioral health problems: challenges and opportunities. *Prevention Science, 20*, 1147–1168. https://doi.org/10.1007/s11121-019-01048-8.

Fenson, L., Marchman, V., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates communicative development inventories. users guide and technical manual* (2nd ed.). Baltimore: Paul H. Brookes Publishing Co.

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental science, 16*, 234–248. https://doi.org/10.1111/desc.12019.

Friend, M., & Keplinger, M. (2008). Reliability and validity of the computerized comprehension task (CCT): Data from American English and Mexican Spanish infants. *Journal of child language, 35*, 77–98. https://doi.org/10.1017/s0305000907008264.

Friend, M., Schmitt, S. A., & Simpson, A. M. (2012). Evaluating the predictive validity of the computerized comprehension task: comprehension predicts production. *Developmental psychology, 48*, 136. https://doi.org/10.1037/a0025511.

Gilkerson, J., Richards, J. A., Warren, S. F., Oller, D. K., Russo, R., & Vohr, B. (2018). Language experience in the second year of life and language outcomes in late childhood. *Pediatrics, 142*, Article e20174276.

Greenwood, C. R., Schnitz, A. G., Carta, J. J., Wallisch, A., & Irvin, D. W. (2020). A systematic review of language intervention research with low-income families: a word gap prevention perspective. *Early Childhood Research Quarterly, 50*, 230–245. https://doi.org/10.1016/j.ecresq.2019.04.001.

Hammer, C. S., Morgan, P., Farkas, G., Hillemeier, M., Bitetti, D., & Maczuga, S. (2017). Late talkers: a population-based study of risk factors and school readiness consequences. *Journal of Speech, Language, and Hearing Research, 60*, 607–626. https://doi.org/10.1044/2016_JSLHR-L-15-0417.

Helmerhorst, K. O., Riksen-Walraven, J. M. A., Fukkink, R. G., Tavecchio, L. W., & Deynoot-Schaub, M. J. G. (2017). Effects of the caregiver interaction profile training on caregiver–child interactions in Dutch childcare centers: a randomized controlled trial. *Child & youth care forum, 46*, 413–436. https://doi.org/10.1007/s10566-016-9383-9.

Hoff, E. (2013). Interpreting the early language trajectories of children from low-SES and language minority homes: implications for closing achievement gaps. *Developmental psychology, 49*, 4–14. https://doi.org/10.1037/a0027238.

Hsueh, J., Halle, M. T. G., & Maier, M. (2020). *An overview of implementation research and frameworks in early care and education research* (pp. 177–194). New York, NY: Foundation for Child Development. In. *Getting it right: Using implementation research to improve outcomes in early care and education*.

Jørgensen, R., Dale, P. S., Bleses, D., & Fenson, L. (2009). CLEX: a cross-linguistic lexical noms database. *Journal of child language, 37*, 419–428.

Justice, L. M., Turnbull, K. L., Bowles, R. P., & Skibbe, L. E. (2009). School readiness among children with varying histories of language difficulties. *Dev. Psychol., 45*, 460–475. https://doi.org/10.1037/a0014324.

Justice, L. M., Jiang, H., Bates, R., & Koury, R. (2020). *Language disparities related to maternal education emerge by two years in a low-income sample* Manuscript revised and resubmitted.

Justice, L. M., McGinty, A. S., Cabell, S. Q., Kilday, C. R., Knighton, K., & Huffman, G. (2010). Language and literacy curriculum supplement for preschoolers who are academically at risk: a feasibility study. *Language, Speech, and Hearing Services in Schools, 41*, 161–178. https://doi.org/10.1044/0161-1461(2009/08-0058).

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49*, 241–253. https://doi.org/10.3102/0013189X20912798.

Landry, S. H., Zucker, T. A., Taylor, H. B., Swank, P. R., Williams, J. M., Assel, M., & Lonigan, C. J. (2014). Enhancing early childcare quality and learning for toddlers at risk: the responsive early childhood program. *Developmental psychology, 50*, 526–541. https://doi.org/10.1037/a0033494.

Larson, A. L., Cycyk, L. M., Carta, J. J., Hammer, C. S., Baralt, M., Uchikoshi, Y., & Wood, C. (2020). A systematic review of language-focused interventions for young children from culturally and linguistically diverse backgrounds. *Early Childhood Research Quarterly, 50*, 157–178. https://doi.org/10.1016/j.ecresq.2019.06.001.

McClelland, M. M., Cameron, C. E., Connor, C. M., Farris, C. L., Jewkes, A. M., & Morrison, F. J. (2007). Links between behavioral regulation and preschoolers' literacy, vocabulary, and math skills. *Developmental psychology, 43*, 947–959. https://doi.org/10.1037/0012-1649.43.4.947.

Metz, A., Naoom, S., Halle, T., & Bartley, L. (2015). *An integrated stage-based framework for implementation of early childhood programs and systems (OPRE Research Brief OPRE 201548)* Washington, DC: Office of Planning. *Research and Evaluation, Administration for Children and Families, US Department of Health and Human Services*.

Ministry of Children and Social Affairs. (2018a). *Børns alder ved start i dagtilbud*.

Ministry of Children and Social Affairs. (2018b). *Velfærdspolitisk Analyse: Personale i daginstitutioner - normering og uddannelse*.

Moreno, A. J., Green, S., & Koehn, J. (2015). The effectiveness of coursework and onsite coaching at improving the quality of care in infant–toddler settings. *Early Education and Development, 260*, 66–88. https://doi.org/10.1080/10409289.2014.941260.

Nix, R. L., Bierman, K. L., Domitrovich, C. E., & Gill, S. (2013). Promoting children's social-emotional skills in preschool can enhance academic and behavioral functioning in kindergarten: findings from Head Start REDI. *Early Education & Development, 24*, 1000–1019. https://doi.org/10.1080/10409289.2013.825565.

Olswang, L. B., & Goldstein, H. (2017). Collaborating on the development and implementation of evidence-based practices: advancing science and practice. *Evidence-Based Communication Assessment and Intervention, 11*, 61–71. https://doi.org/10.1080/17489539.2017.1386404.

Phillips, D., Lipsey, M., Dodge, K., Haskins, R., Bassok, D., Burchinal, M., & Weiland, C. (2017). Puzzling it out: the current state of scientific knowledge on pre-kindergarten effects. A consensus statement. *Issues in Pre-kindergarten Programs and Policy*, 19–30.

Purpura, D. J., & Reid, E. E. (2016). Mathematics and language: Individual and group differences in mathematical language skills in young children. *Early Childhood Research Quarterly, 36*, 259–268. https://doi.org/10.1016/j.ecresq.2015.12.020.

Rosholm, M., Paul, A., Bleses, D., Højen, A. S., Dale, P., Jensen, P., & Calmar Andersen, S. (2021). Are impacts of early interventions in the Scandinavian welfare state consistent with a Heckman curve? A meta-analysis. *Journal of Economic Surveys, 35*, 106–114. https://doi.org/10.1111/joes.12400.

Schindler, H. S., McCoy, D. C., Fisher, P. A., & Shonkoff, J. P. (2019). A historical look at theories of change in early childhood education research. *Early Childhood Research Quarterly, 48*, 146–154. https://doi.org/10.1016/j.ecresq.2019.03.004.

Sjö, N. M., Bleses, D., Dybdal, L., Nielsen, H., Sehested, K. K., Kirkeby, H., & Jensen, P. (2019). Measurement properties of the SEAM questionnaire using Rasch analysis on data from a representative Danish sample of 0-to 6-year-olds. *Journal of Psychoeducational Assessment, 37*, 320–337 pDs:O//dIo: i1.o0r.g1/107.171/077/304732482892197177446625.

Sjoe, N. M., Kiil, A., Bleses, D., Dybdal, L., Kreiner, S., & Jensen, P. (2020). Assessing strengths and difficulties in social development: a comparison of the Social Emotional Assessment Measure (SEAM) with two established developmental psychopathological questionnaires. *European Journal of Developmental Psychology, 17*, 103–122. https://doi.org/10.1080/17405629.2018.1540975.

Slot, P. L., Bleses, D., & Jensen, P. (2020). *Infants' and toddlers' language, math and social-emotional development: evidence for reciprocal relations and differential gender and age effects* Manuscript accepted in Frontiers of Psychology.

Slot, P. L., Bleses, D., Justice, L. M., Markussen-Brown, J., & Højen, A. (2018). Structural and process quality of Danish preschools: direct and indirect associations with children's growth in language and preliteracy skills. *Early Education and Development, 29*, 581–602. https://doi.org/10.1080/10409289.2018.1452494.

Slot, P. L., & von Suchodoletz, A. (2018). Bidirectionality in preschool children's executive functions and language skills: is one developing skill the better predictor of the other? *Early Childhood Research Quarterly, 42*, 205–214. https://doi.org/10.1016/j.ecresq.2017.10.005.

Son, S.-H. C., Choi, J. Y., & Kwon, K.-. A. (2019). Reciprocal associations between inhibitory control and early academic skills: evidence from a nationally representative sample of head start children. *Early Education and Development, 30*, 456–477. https://doi.org/10.1080/10409289.2019.1572382.

Squires, J. (2014). *Social-emotional assessment/evaluation measure (SEAM)*. Paul H. Brookes Publishing.

Vach, W., Bleses, D., & Jørgensen, R. N. (2010). Construction of a Danish CDI short form for language screening at the age of 36 months: Methodological considerations and results. *Clinical Linguistics & Phonetics, 24*, 602–621. https://doi.org/10.3109/02699201003710606.

Walker, D., Sepulveda, S. J., Hoff, E., Rowe, M. L., Schwartz, I. S., Dale, P. S., & Levine, S. C (2020). Language intervention research in early childhood care and education: a systematic survey of the literature. *Early Childhood Research Quarterly, 50*, 68–85. https://doi.org/ 10.1016/j.ecresq.2019.02.010.

Weiland, C., McCormick, M., Mattera, S., Maier, M., & Morris, P. (2018). Preschool curricula and professional development features for getting to high-quality implementation at scale: a comparative review across five trials. *AERA Open, 4*, Article 2332858418757735. https://doi.org/10.1177/2332858418757735.

Werner, C. D., Vermeer, H. J., Linting, M., & Van IJzendoorn, M. H. (2018). Video-feedback intervention in center-based childcare: a randomized controlled trial. *Early Childhood Research Quarterly, 42*, 93–104. https://doi.org/10.1016/j.ecresq.2017.07.005.

Bishop, G., Spence, S. H., & McDonald, C. (2003). Can parents and teachers provide a reliable and valid report of behavioral inhibition? *Child Dev., 74*, 1899–1917. https://doi.org/10.1046/j.1467-8624.2003.00645.

National Center for Education Statistics [NCES], (2002). Early Childhood Longitudinal Study, Kindergarten class of 199899 (ECLS-K): Psychometric Report for Kindergarten Through First Grade. Washington, DC: NCES.