AARHUS UNIVERSITY

Coversheet

This is the accepted manuscript (post-print version) of the article.

Contentwise, the accepted manuscript version is identical to the final published version, but there may be differences in typography and layout.

How to cite this publication

Please cite the final published version:

Bleses, D. , Højen, A. , Justice, L. M., Dale, P. S., Dybdal, L. , Piasta, S. B., Markussen-Brown, J. , Clausen, M. and Haghish, E. F. (2017), The Effectiveness of a Large-Scale Language and Preliteracy Intervention: The SPELL Randomized Controlled Trial in Denmark. Child Dev. .
doi:[10.1111/cdev.12859](https://doi.org/10.1111/cdev.12859)

Publication metadata

Title:	The Effectiveness of a Large-Scale Language and Preliteracy Intervention: The SPELL Randomized Controlled Trial in Denmark
Author(s):	Dorthe Bleses, Anders Højen, Laura M. Justice, Philip S. Dale, Line Dybdal, Shayne B. Piasta, Justin Markussen-Brown, Marit Clausen & E. F. Haghish
Journal:	Child Development
DOI/Link:	https://doi.org/10.1111/cdev.12859
Document version:	Accepted manuscript (post-print)

This is the peer reviewed version of the following article: Bleses, D. , Højen, A. , Justice, L. M., Dale, P. S., Dybdal, L. , Piasta, S. B., Markussen-Brown, J. , Clausen, M. and Haghish, E. F. (2017), The Effectiveness of a Large-Scale Language and Preliteracy Intervention: The SPELL Randomized Controlled Trial in Denmark. Child Dev. . doi:[10.1111/cdev.12859](https://doi.org/10.1111/cdev.12859), which has been published in final form at <https://doi.org/10.1111/cdev.12859>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

General Rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

If the document is published under a Creative Commons license, this applies instead of the general rights.

Abstract

In the present study, we extend efficacy research on language and pre-literacy interventions for children by estimating effects in a real-world effectiveness trial in Denmark with 6,483 3-to-6-year-olds. Children in 144 childcares were assigned to a control condition or one of three planned variations of a 20-week storybook-based intervention: the basic intervention or two enhanced versions featuring extended professional development or a home-based, caregiver-administered supplemental program. Pre-posttest comparisons revealed a significant impact of all three interventions for pre-literacy skills ($ES = 0.21-0.27$) but not language skills ($ES = 0.04-0.16$). There was little evidence for differences among intervention conditions. Fidelity, indexed by number of lessons delivered, was a significant predictor of most outcomes. Implications for real-world research and practice are considered.

.

The Effectiveness of a Large-Scale Early Childhood Language and Pre-literacy Intervention:

The *SPELL* Randomized-Controlled-Trial in Denmark

The reading skill of adolescents and adults is highly correlated with their literacy skills earlier in life (Verhoeven & Van Leeuwe, 2008). Children who arrive to formal reading instruction with well-developed language and pre-literacy skills have better academic trajectories than children with lags in these areas (National Early Literacy Panel, 2008, Chapter 2). Thus, there is increasing interest in identifying children who exhibit lags in language and pre-literacy development during the preschool years and providing them with educational interventions to improve these skills. To this end, there have been many studies of interventions intended to boost children's early language and pre-literacy skills in an effort to better prepare them for the transition to formal reading instruction and future reading development (see National Early Literacy Panel, 2008). Many of these interventions have provided benefits to young children that are considered to be of practical as well as statistical significance, contributing to longer-term benefits in reading achievement (National Early Literacy Panel, 2008, Chapter 3).

The present study represents an important complement to the relatively large corpus of efficacy studies conducted in this area, by providing an estimate of effects when a systematic, explicit language and pre-literacy intervention was implemented at scale in an unselected heterogeneous sample and in real-world conditions. The study was implemented in a Danish context in which childcare programs have typically provided little systematic focus on developing children's language and pre-literacy skills. It is a particular strength of the study that the potential differential benefits of the same intervention are investigated for selected subsamples, including children from low-socioeconomic status (SES) backgrounds and dual language learners (DLLs, here children with non-native Danish background). The design enables comparisons of different intervention conditions and subsamples without the confounding which occurs when these

comparisons are made across separate studies with diverse samples and outcome measures.

Implementing Systematic Language and Pre-literacy Interventions for Young Children

The most critical element identified in research on effective early childhood language and pre-literacy interventions is the provision of systematically organized and explicit instruction that targets children's development of skills that are predictive of future reading achievement (National Early Literacy Panel, 2008, Chapter 3). Systematic instruction utilizes learning sequences that are guided by a rigorous understanding of how children learn, whereas explicit instruction makes it clear to children what they have to attend to within an activity by orienting them towards the goals of the activity, thus facilitating their engagement in the learning process. Targeted skills often include phonological awareness, print awareness, and oral language skills, including vocabulary. Manualized language and pre-literacy interventions that yield positive effects on children's skills typically feature these elements in combination (Justice, Kaderavek, Fan, Sofka, and Hunt, 2009; Whitehurst et al., 1994). Systematic, explicit interventions focused on improving the language and pre-literacy skills of children tend to yield effects of about 0.25 to 0.87 standard deviation units, which is considered of practical significance in the educational literature

Typically, the effects of early educational interventions are focused on children exhibiting risk; however, improving the language and pre-literacy skills for the entire population of young children is also important for several reasons. First, our ability to identify children who will exhibit poor reading in their future is too imperfect to identify all children in need for early intervention (e.g., Bornstein et al., 2014). Second, the growing demand for creating new technologies and innovation has led to a global competition for highly skilled and highly literate employees. Finally, within Denmark, there are concerns about the relatively modest levels of reading achievement among students despite high levels of public investment in schooling (OECD, 2010). The great majority of early language and literacy intervention studies have focused on determining whether

benefits are derived for children from low-SES backgrounds and/or those who are dual language learners. These children consistently exhibit lower average levels of language and pre-literacy skill during the preschool years than their more advantaged peers (e.g., Hoff, 2013). However, relatively few studies have examined whether there are differential intervention effects for children as a function of SES and language status (Bierman et al., 2006; Justice et al., 2010).

With respect to intervention effects as a function of SES, evidence is mixed. In a meta-analysis of vocabulary interventions, Marulis and Neuman (2013) found that higher-SES children benefited significantly more from the interventions than low-SES children. Likewise, the meta-analysis of the National Early Literacy Panel (2008, Chapter 3) often found larger effect sizes for higher-SES than low-SES children, but for reading outcomes, low-SES children benefited more. Even fewer studies systematically examined language interventions benefits for native vs. non-native (DLL) children. The meta-analysis of the National Early Literacy Panel (2008) showed higher effect sizes for Latino than Caucasian or African American children. However, the differences were not significant, possibly because of too few studies with a clear background classification. A recent systematic review evaluated the impact of early education programs and practices on 3-to-5-year old DLL children's development (Buysse, Peisner-Feinberg, Pérez, Hammer, and Knowles, 2014). The review identified 25 studies, but only four studies involved educators delivering a language and pre-literacy intervention using English as the language of instruction, and only one of them was conducted as a randomized control trial (Collins, 2010). Across studies, some empirical support was provided for positive effects of DLL children's language and pre-literacy development as a result of interventions, as is also seen in European-based studies involving DLLs acquiring languages like Arabic, Somali and Turkish as their first language and either Dutch or English as their second language (Dockrell, Stuart, & King, 2010; Van Tuijl, Leseman, & Rispen, 2001; Verhallen, & Bus, 2010,). Importantly, none of the American or

European studies directly compared intervention effects in non-DLL vs. DLL children. One reason it is difficult to draw a firm conclusion from the current literature is that in addition to methodological problems with small sample sizes and inconsistencies at all levels, few studies have examined the effects of the same intervention in both high- and low-SES children or native and DLL children, thus we are limited to comparisons across studies. Therefore, more research is needed to investigate differential effects in a larger and more heterogeneous sample of children.

Some research has focused on how the effects of systematic explicit instruction on children's outcomes can be enhanced via professional development (henceforth PD) or companion programs conducted by parents. In relation to PD, many interventions are based on the rationale that improving educators knowledge about language development and use of language facilitation strategies, sometimes combined with coaching of educators' practice during intervention, will lead to improved fidelity of the intervention and thereby improved child outcomes. Whereas many studies have investigated effects of different forms of PD (see Sheridan, Edwards, Marvin, and Knoch, 2009), few studies have compared the effect of a curriculum by itself with the added value of additional PD on children's outcome. Lonigan, Farver, Phillips, and Clancy-Menchetti (2011) evaluated the effect of the Literacy Express Preschool Curriculum and two additional PD components, a workshop and a mentor condition (Lonigan et al., 2011). In spite of more consistent and nominally higher effects, there was only one significant difference in child outcomes favoring the more intensive PD model. Similarly, Assel, Landry, Swank, and Gunnewig (2007) found some evidence for an effect of mentoring in combination with a curriculum, but the effect was not found across all skills and program sites. However, randomized-controlled-trials indicate that incorporating elements of practice-focused PD in coursework can enhance the quality of systematic explicit interventions through the provision of specific knowledge of effective practice (Hamre et al., 2012; Wasik and Hindman, 2011), training teachers to detect effective adult-child interactions

via video analyses (Hamre et al., 2012), increasing teacher knowledge of children's language and pre-literacy skills (Al Otaiba et al., 2011; Landry, Anthony, Swank, and Monseque-Bailey, 2009; Landry, Swank, Anthony, and Assel, 2011), and offering opportunities for collegial discussion (Buysse, Sparkman, and Wesley, 2003). The effect of such elements has only been evaluated in at-risk children and never as an add-on component to a curriculum.

There is evidence that interventions implemented in parallel in preschools and home settings can boost the effects over what is seen when interventions are implemented in preschools only. The rationale underlying such approaches is that the higher dosage of the intervention through aligned exposure in both childcare and home settings will improve children's outcomes. van Steensel, McElvany, Kurvers, and Herppich (2010) found minor improvement of effects (mean effect size of 0.17 for home-based only vs. 0.24 for home-based + child-care interventions). In contrast, Manz, Hughes, Barnabas, Bracaliello, and Ginsburg-Block (2010) found that child outcomes were reduced substantially when programs were implemented in both childcare and home settings (effect size 0.13 for childcare and home settings vs. 0.47 for home only). The vast majority of this research has been conducted in the US, predominantly in at-risk families and minority families with Spanish as the home language. More research in heterogeneous samples in non-US contexts is needed.

Taking Language and Pre-literacy Interventions to Scale

Efficacy trials, also known as explanatory trials, are distinct from effectiveness trials, also known as pragmatic trials (Schwartz, & Lellouch', 1967). Efficacy trials test the effect of an intervention in an ideal situation, whereas effectiveness studies seek to examine intervention benefits as they occur in real-world conditions; the latter type of study seeks to produce evidence of practical importance that is relevant to all individuals (e.g. Gartlehner, Hansen, Nissman, Lohr, and Carey, 2006,). It should be noted, however, that the efficacy-effectiveness distinction is best viewed as continuous rather than dichotomous, with efficacy trials focusing on homogeneity to obtain high

internal validity and controlling for as many factors as possible to avoid bias, whereas effectiveness trials are focused on maximizing heterogeneity in all aspects to obtain high external validity.

Typically, efficacy studies take place in a resource-intensive ‘ideal’ setting, use relatively small and homogenous participant samples, have strong requirements for provider expertise, evaluate highly standardized and strictly implemented interventions, employ intensive resources in order to maximize compliance, and utilize triple-blinded designs. Effectiveness trials typically take place in a real-world everyday setting, use large and heterogeneous samples and representative usual providers, implement interventions with no or limited reinforcement or compliance support beyond normal practice, and accept that blinded designs are often not possible (Singal, Higgins, and Waljee, 2014). Often, effectiveness studies have high levels of participant attrition, given their implementation in real-world settings (see O’Brien et al., 2012); to ensure that effects are unbiased even when attrition is high, effectiveness trials often include an intent-to-treat framework, in which outcomes for all participants, even those who attrite, are included in the assessment of treatment outcomes. An additional key characteristic of effectiveness studies is that implementers’ fidelity to the intervention tends to be lessened relative to the more controlled efficacy trials. Therefore measuring fidelity of an intervention is an essential component of effectiveness trials, and analyses of the relations between implementation fidelity and treatment outcomes can be helpful in exploring key mechanisms of the intervention when implemented broadly. As effects are often diluted due to lower fidelity (Hulleman and Cordray, 2009), effectiveness trials must be large enough to increase power to detect small effects and to address issues of attrition.

To date, the vast majority of rigorously conducted studies of language and pre-literacy interventions have been efficacy studies with extensive involvement of researchers in the intervention, relatively small numbers of children from mainly low-SES homes, and a focus on demonstrations of impact under those circumstances (National Early Literacy Panel, 2008, Chapter

3). The effectiveness of such interventions as implemented under real-life circumstances and at a large scale is not clear. A literature review identified only six large-scale ($n > 500$ children, a somewhat arbitrary criterion selected to ensure reasonable statistical power) randomized-control-trials of childcare-based language and pre-literacy interventions with some attempts to maximize heterogeneity and generalizability of the results (the interventions were all carried out in a real-life setting by representative usual providers with no or limited reinforcement of fidelity carried out, and with larger samples of children to increase power; but note though that studies including a PD component had some level of reinforcement of the intervention due to training and coaching during the intervention). Two studies evaluated the effects of language and/or pre-literacy curricula (Fischel et al., 2007; Neuman, Newman, and Dwyer, 2011), one study evaluated a curriculum in combination with PD (Lonigan et al., 2011), and three studies evaluated various PD approaches (Landry et al., 2009; Powell, Diamond, Burchinal, and Koehler, 2010; Wasik and Hindman, 2011). The sample sizes varied from 507 to 1,789 (see Appendix for details including effect sizes).

As expected, effect sizes of those effectiveness trials were noticeably lower than those seen in comparable efficacy trials (National Early Literacy Panel, 2008, Chapter 3) and comparable to those reported by the NICHD ECCRN (2006). The effect on phonological awareness decreased from an average of 0.87 to being significant in only half of the studies that included such measure ($k = 4$) with effect sizes at 0.22 (Powell et al., 2010 and Wasik and Hindman, 2011). Letter Identification effects decreased from 0.66 to a range that included non-significant effects ($k = 1$) to significant effect sizes between 0.15-0.47 ($k = 3$). Except for one study (Lonigan et al., 2011), the effects on vocabulary were non-significant ($k = 6$). The role of fidelity for these lower effect sizes is not clear because although measures of fidelity were provided in five studies, differential effects for children as a function of fidelity were only investigated in one study (Wasik and Hindman, 2011). In that study fidelity varied substantially across teachers, and fidelity explained 28% of the variation in

child outcomes for phonological sensitivity and 29% for vocabulary. In sum, even though these studies did increase the external validity, all seven studies only included at risk-children. It is therefore unknown how the results generalize to the entire population of preschoolers.

The Present Study

The present study is an attempt to fill a gap in the literature on the effects of early language and pre-literacy interventions by reporting the results of a large-scale randomized control effectiveness trial conducted across Denmark under real-life conditions and implemented with 6,483 unselected children. As an effectiveness trial conducted outside of the North American context, the present study promotes the generalizability of results of such work on several levels.

First, in the U.S., early learning standards are in place and childcares and preschools offer early language and pre-literacy curricula as part of their educational practice. In Denmark, in contrast, there are no learning standards applied to young children, and there is seldom a systematic instructional focus on either language or pre-literacy skills during the preschool years (hence the preferred use of the term ‘childcare’ rather than ‘preschool’ in Denmark). Therefore it may offer a different counterfactual than that seen in the numerous North American studies.

Second, this is the first large-scale effectiveness trial using an unselected heterogeneous sample of preschoolers, which makes it possible to ask whether results from earlier efficacy studies of systematic explicit language and pre-literacy interventions are generalizable to all children. The present study includes all children in the enrolled childcare centers and not only those who are in programs serving high-risk children. Importantly, the overwhelming majority of 3-5-year-olds in Denmark (approximately 97%) are enrolled in childcare (Danish Ministry for Social Affairs, 2015). The universal childcare tradition and a large sample size makes it possible to investigate potential differential benefits of an intervention for different groups of children, e.g., low-SES children vs. the population in general. Moreover, DLL children in Denmark are quite different from DLL

children in the U.S.; the largest groups of children of immigrants originate from Turkey, Lebanon, Pakistan, Iraq, and Somalia (Statistics Denmark 2014). The term DLL in the present study is used to refer to children of immigrant parents in Denmark, for whom Danish is a second or later language. However, despite many differences between Denmark and the U.S., concerns about the language and pre-literacy achievement of low-SES children and DLL children are shared. Low-SES children in Denmark are performing 0.33 of a standard deviation below their more advantaged peers, and one-half of DLL children with a non-western background perform more than one standard deviation below the mean of monolingual children (Bleses, Højen, Jørgensen, Jensen, and Vach, 2010). Interestingly, at the other end of the distribution, concerns also exist, as shared in many OECD (Organization for Economic Cooperation and Development) countries. Denmark has a lower percentage of readers at the top proficiency level compared to the OECD average (5% vs. 8%; range 0-24%), and in fact the proportion of top performers in reading decreased from 2000 to 2009 (OECD, 2010). In the light of increasing needs for advanced literacy, academic, and cognitive skills in technological societies, it has become a political concern in Denmark that all children in daycares should improve their early language and literacy skills to improve eventual educational attainment in school. It is an open question, not yet empirically investigated, whether interventions designed for at-risk children, which are typically remediation-oriented, can also have a skill acceleration effect for children in the average and above-average range.

Third, the study involves planned variations of an existing intervention, *Read It Again* (RIA) previously studied in several efficacy trials in the U.S. (e.g., Justice et al., 2010) to determine the generalizability of obtained results when taking to scale and whether enhancing options, viz. extended PD or parallel home-based activities, offer more impact to children. RIA was adapted for this Danish effectiveness trial and renamed *SPELL* (*Structured Preschool Efforts in Language and Literacy*). Although other previously-tested interventions warranted consideration for scale-up, RIA

has several desirable characteristics: first, it is a wholly manualized and soft-scripted intervention that requires no specialized materials for implementation aside from commercially available storybooks; second, it can be used with relatively high levels of fidelity by educators working in diverse settings; and third, it is available at no cost and has the potential to achieve a large number of end-users at a very low cost. The original version features a fourfold scope of instruction (vocabulary, narrative, print knowledge, phonological awareness) and 23 sequenced objectives organized into sixty 25-minute lessons implemented in a large group setting, each of which features explicit instruction embedded into a shared storybook reading session. To fill gaps in the existing research, we designed two alternatives to the SPELL-BASIC intervention (henceforth SPELL), each with an additional component: a SPELL+Extended Professional Development (henceforth SPELL+EPD) version based on a practice-focused approach to PD with the overall aim of improving fidelity of the intervention, and a SPELL+home-based companion program (henceforth SPELL+HOME) version with the overall aim of providing more learning opportunities to children.

The present study was designed to address four research questions in a population-representative sample including at-risk and non-risk children: (1) What is the effectiveness of an early language and pre-literacy intervention (SPELL) relative to business-as-usual childcare within Denmark? (2) To what extent does additional PD or inclusion of a home-based companion program enhance the effects of SPELL? (3) To what extent do child-level risk factors moderate the SPELL impacts, specifically children's SES and language status? (4) To what extent does intervention fidelity (exposure) predict SPELL impacts for children?

Method

Participants

The multisite, multi cohort study was designed based on a priori power analysis to enroll 144 childcares, with random assignment at the level of the childcare resulting in 36 childcares per

each of four conditions. A total of 144 childcares with 7,120 children were recruited over two consecutive cohorts, but the number of childcares per condition varied slightly at pretest (Business-as-usual, BAU, $n = 36$; SPELL, $n = 38$; SPELL+EPD, $n = 34$; SPELL+HOME, $n = 34$). Data for all randomized childcares, educators, and children for whom background data could be obtained ($n = 7,076$) are shown in Table 1. All classrooms included all ages 3-6.

Table 1

Not all daycares, educators and children is shown in Table 1 made it to the final sample. Pretests were obtained only for 6,483 children (by 612 educators) due to illnesses or other absences. In addition, 8 childcares dropped out during the study between pretest and posttest (BAU, $n = 1$; SPELL, $n = 2$; SPELL+EPD, $n = 3$; SPELL+HOME, $n = 2$). Because the child-level assessments were implemented by the children's childcare educators, the study was not designed to pursue posttest assessments with children even if the childcare attrited, as would be ideal in a true intent-to-treat study. Of the eight childcares that left after randomization, six left due to leadership or staff transitions and difficulties; for the seventh, no information could be acquired. In an eighth childcare (BAU), posttest data were reportedly obtained but not received by the project management.

We evaluated differential attrition rates from randomization to posttest following the guidelines of the U.S. Institute of Education Sciences What Works Clearinghouse, version 3.0 (http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf) for each of the three intervention conditions compared to the BAU, calculating an “overall” mean attrition rate for each pairwise comparison (e.g., SPELL+HOME vs. BAU). At the childcare level, all three attrition rate comparisons were “low” even under conservative boundary assumptions (endogenous attrition reasons). At the child level, differential attrition rates were “low” under liberal assumptions for SPELL and SPELL+EPD, and “low” under conservative assumptions for SPELL+HOME. Differential attrition rates from pretest to posttest were “low” under

conservative assumptions for all comparisons at both childcare and child level. Further details on randomized, pretested and posttested children and attrition rates are available in the online material.

All children who had both pre- and posttest scores were included in analyses. The number of children tested on each subscale varied due to 1) age-related administration of subscales, and 2) failure of a few children to complete all subscales. As standard software for multiple imputation does not allow a five-level nested design, we included in each of our analyses all children for whom information on all variables involved in the analysis were available.

Procedure

SPELL. Children whose childcare was assigned to the three variations of the SPELL intervention were exposed to a 20-week intervention which provided an explicit scope and sequence of language and pre-literacy instruction over 40 soft-scripted 30-minute lessons implemented by their educator twice per week. All children in a classroom participated in the intervention, but each lesson was implemented in a small-group format. Each classroom had two educators and each educator in the classroom had two groups. The small-group format was chosen for reasons of face validity and acceptability among educators based on feedback from educators in the pilot daycares. Danish childcare educators have experience with small-group activities, but not large-group structured activities. Ten high-quality storybooks were carefully chosen for use in the SPELL lessons and each was read twice during the intervention. For the purposes of this study, the SPELL lessons were provided to teachers on an iPad with an app featuring lesson plans and a logging tool (implementation notes). Each lesson involved two components: 1) the SPELL lesson plan, which soft-scripted a sequence of step-by-step instruction featuring a before-, during-, and after-reading activity as well as suggested language that educators could use to support children's learning during each activity, and 2) a Learners' Ladder, which offered specific scaffolding strategies to

differentiate instruction (e.g., make easier or make more challenging) within the lesson on the basis of the children's individual response to the lesson (for more details, see Justice and McGinty, 2012).

To support self-reflection on the implementation of each lesson, educators maintained obligatory implementation notes on the iPad targeting critical elements in the intervention, child attendance, engagement and benefit of the intervention, and their usage of specific scaffolding strategies towards individual children (in fact, starting a new lesson required completion of previous implementation notes on the iPad). Educators also tracked children's individual progress towards each SPELL objective within each of the four learning domains addressed in SPELL on the iPad three times during the 20-week-intervention period (after lesson 6, 20 and 36) using an informal assessment indication of whether the child 'never', 'sometimes' or 'often' demonstrated the skill). Both implementation notes and progress notes were automatically wired to a database.

Prior to starting delivery of the intervention, the educators received all of the intervention materials (e.g., iPad, storybooks) and participated in a two-day training workshop (14 hours). The workshop taught about the learning domains and objectives, how to reference these explicitly during shared book reading, and how to use the Learner's Ladder.

SPELL+EPD. Educators in SPELL+EPD implemented all SPELL components, but received extra PD in two additional intensive 2-day trainings in intervention weeks 3 and 6. This training was designed to increase the quality of the intervention delivery by focusing on enhancing differentiation of instruction not least for the benefit of at-risk children. Educators were provided with a language profile of each child and received further training in practicing differentiation strategies by watching others' and their own video recordings of SPELL lessons. In addition, educators partook in a weekly refresher session in which educator pairs conducted small assignments to support implementation of the intervention.

SPELL+HOME. In SPELL+HOME, educators implemented all SPELL components. In addition, parents implemented SPELL activities twice a week at home to complement the SPELL intervention that their children received in their classrooms. Parents received companion materials (the same 10 books, 20 lesson plans, translated into seven languages) for each book reading that addressed the same learning domains and learning objectives as in SPELL and a reading calendar with stickers to mark each reading session. Besides access to model videos, the parents were not offered any specific training. The home activities were aligned in time with the SPELL lessons in the childcares (see also, Bleses et al., 2014). Parent fidelity was measured via a weekly paper checklist, however, parent provision of the checklist back to educators was very low and/or educators didn't transfer the parents responses to an on-line checklist so there was a great deal of missing data. Use of these data was therefore not informative to the reporting of the trial.

Intervention Fidelity

Intervention fidelity in terms of number of completed lessons was obtained from the educators' obligatory completion of implementation notes after each lesson on the iPad. Out of 40 possible lessons, educators in SPELL, SPELL+EPD, and SPELL+HOME completed 30.1 lessons ($SD = 10.1$), 27.6 lessons ($SD = 9.0$), and 33.3 lessons ($SD = 9.5$), respectively. The 10% of educators with the lowest implementation fidelity completed 15, 18, and 13 lessons in each condition, respectively, and the 10% of educators with the highest implementation fidelity in all three conditions completed all 40 lessons. There was a trend for the SPELL+EPD condition to complete fewer lesson, $\chi^2(2, n = 833) = 5.96, p = .051$. As each child's attendance was variable, the exposure to the intervention was somewhat lower on an individual child basis. Children in SPELL, SPELL+EPD, and SPELL+HOME were exposed to 25.6 lessons ($SD = 9.48$), 22.6 lessons ($SD = 9.28$), and 27.6 lessons ($SD = 9.26$), respectively. The 10% children with the lowest exposure received 12, 10, and 13 lessons, and the 10% children with the highest exposure received 36, 35,

and 37 lessons. The exposure of children to the intervention did not significantly differ among condition, $\chi^2(2, n = 3,992) = 4.66, p = .097$. The adherence to intervention elements was based on teachers' reports in implementation notes. In SPELL, SPELL+EPD, and SPELL+HOME educators reported implementing all critical elements of the intervention (material, activities, learning objectives) in 76% ($SD = 28$), 71% ($SD = 27$), and 75% ($SD = 26$) of the lessons, respectively. Additionally, research staff observed and rated adherence to the intervention on the basis of a checklist developed for this specific purpose via video recordings of lessons submitted by the educators at the beginning, mid and end of the 20 week program. The reliability was determined by double-coding 10 % of the video-recordings. Agreement between a master-coder and research staff was between 77-100%. Based on observational rating of 234 lessons, educators obtained adherence ratings of 94% (SPELL) and 99% (SPELL+EPD, SPELL+HOME) in terms of how well they followed structural elements in the lesson plan, whereas the extent to which educators addressed the learning objectives was somewhat lower (65% in SPELL, 75% in SPELL+EPD, and 60% in SPELL+HOME), supporting the educators' self-reporting of the adherence to the intervention. In sum, there were no significant differences in terms of implementation of the intervention across the three different SPELL conditions.

Measures

Measures used in this study were of two types: measures of children's language and pre-literacy skills, and background information for children and parents.

The language and pre-literacy skills of children were assessed pre- and post-intervention by their educators using an already published assessment instrument that was adapted for the purpose, the *Language assessment of children: 3-6* (LA; Bleses, Vach, Jørgensen, & Worm, 2010), that is administered by educators in the vast majority of Danish municipalities as part of a national screening program. LA-3-6 was used for national screening purposes in all but one of the SPELL

municipalities (in that municipality educators received specific local training for the LA). Two age dependent versions of the assessment battery were administered. Four subscales were administered to 3-year-old children: Sound Discrimination (of initial phonemes in words, max score 16), Vocabulary (expressive, max score 40), Comprehension (of words and complex sentences, max score 12), and Communication (questionnaire with communicative strategies, max score 40). Six subscales were administered to 4-6-year-old-children: Rhyme Detection (max score 17), Deletion (max score 20), Letter identification (max score 12), Vocabulary (expressive, max score 40), Comprehension (of words and complex sentences, max score 27), and Communication (questionnaire with communicative strategies, max score 76). With the exception of Communication children were directly assessed via picture identification or picture elicitation tasks.

For the present purposes, two composite measures were derived from the seven subscales: a Pre-Literacy Composite based on averaging standardized scores for Sound Discrimination, Rhyme Detection, Deletion, and Letter Identification, and a Language Composite based on averaging standardized scores for Vocabulary, Comprehension, and Communication. Standardized scores were based on age and gender specific norms derived from the pretest measurements. In general the results showed acceptable data characteristics according to the criteria for the Rasch model for the subscales. Internal consistency coefficients (Cronbach's alpha) for subscales were between .75-.91; correlations between subscales were between .25-.70. In a small unpublished validity study, the concurrent the concurrent correlations of the language subscales with The Peabody Picture Vocabulary Test-4 (PPVT-4, Dunn and Dunn, 2012) were .55 for Vocabulary and .57 for Comprehension; correlations with the Expressive Vocabulary Test 2 (EVT-2, Williams, 2007) were .42 for Vocabulary and .39 for Comprehension. For the pre-literacy subscales, the correlations with PPVT-4 were .33 for Rhyme detection, .49 for Deletion, and .39 for Letter Identification; the correlations with EVT-2 were .18 for rhyme, .29 for Deletion, and .33 Letter Identification.

Background information for children and parents with respect to SES and language status (DLL or not) was also collected. The former was indexed based on about maternal education (see Table 1); for the latter, parental immigration status served as index. This information was obtained from Statistics Denmark, using the Danish Central Personal Number System.

Analytic Strategy

The data collected in this effectiveness trial were nested in a complex way, with children nested in small groups (in which intervention was received), educators (who delivered the intervention), classrooms, and center. To address this nested data structure, the research questions were addressed using 5-level hierarchical linear models (HLM; Raudenbush & Bryk, 2002), with child represented as Level 1, group as Level 2, educator as Level 3, classroom as Level 4, and child-care center as Level 5 in the analyses. The Level 4 (classroom) and 5 (center) variables were treated as random effects for all three conditions; for the Level 2 (group) and 3 (educator) variables, however, they were treated as random effects for the two intervention conditions only, as children in the BAU condition were not assigned to groups or to educators.

Covariates were included in the models based on the specific research question being addressed. Specifically, for all four questions, the pretest value corresponding to the posttest value of interest was always included as a covariate. When the Comprehension scale was the outcome variable of interest, the version of LA was included as a covariate to take into account the different number of items in the two different age-dependent versions. When maternal education was entered into a model as a covariate, being a DLL child was considered as an additional category, such that only the effect within the sample of Danish children was estimated; this was due to a lack of information about maternal education for many immigrants in Statistics Denmark; Statistics Denmark has data on maternal education only for those who complete their schooling in the country.

Cohen's *d* effect sizes were calculated by dividing the effect estimates from the HLM analyses by the standard deviation of the individual change scores observed. *P*-values for all six pairwise comparisons were based on the Kramer-Tukey method correcting for multiplicity. All analyses were conducted using the STATA mixed command except that Kramer-Tukey *p*-values were calculated using SAS PROC MIXED.

Results

Preliminary Analyses

Prior to addressing the research questions, preliminary analyses were conducted to determine the extent to which educators, classrooms, and children were initially equivalent across the four study conditions. Table 1 presents descriptive statistics for those variables investigated in preliminary analyses. HLM analyses or logistic regression with robust standard errors indicated that there were no significant differences among the four conditions on any continuous or binary characteristic shown in Table 1 (all $p > .20$). Similarly, an analysis of pretest values showed no significant differences (all $p > .24$). In interpreting the effects of the intervention in the following sections, it should be noted that some of the subscales exhibited ceiling effects already at pretest. For 3-year-olds the median score on Sound Discrimination was 16, equal to the maximum score on the scale; children at the 25th percentile obtained 14 items correct. The Comprehension subscale also exhibited ceiling effects for 3-year-olds (median = 16, maximum = 20). No scale exhibited ceiling effect for 4-year-olds, but for 5-year-olds, there were ceiling effects for Letter Knowledge (score on 75th percentile = 11, maximum = 12), Rhyme (score on 90th percentile = 16, maximum = 17), and a less pronounced ceiling effect for vocabulary (score on 90th percentile = 69, maximum = 76). Given these ceiling effects on the pretests, a proportion of the children (10-75% depending on subscale) had no possibility of showing improvement on posttests. It is worth noting that, if anything, this would limit estimated effect sizes (i.e., effect estimates are conservative).

Main Effects

To address the first and second research questions, we examined the extent to which three variations of the SPELL intervention had positive effects on children as compared to BAU instruction. Table 2 in Supplementary Online Material provides pre- and posttest data for the seven subtests of the LA and the two composites. We estimated nine separate HLMs, one for each outcome variable and the two composites, with the following fixed effect covariates included at each level: at level 1 (child-level), the corresponding pretest score, age, subtest version (depending on age; only relevant for comprehension subtest), and gender; and at level 5 (childcare-level), the condition to which the childcare was assigned, the cohort in which the childcare participated, and the municipality of the childcare. Results for the nine HLM analyses appear in Table 3.

Looking first at the contrasts specific to SPELL and BAU, the results show that children who experienced SPELL outperformed those receiving BAU on three subtests (Rhyme Detection, $d = 0.20$; Deletion, $d = 0.27$; Letter Knowledge, $d = 0.19$) in addition to the Pre-literacy Composite ($d = 0.22$). Children in SPELL also scored higher than those in BAU on all other indices, with the exception of Sound Discrimination ($d = -0.05$ - -0.07), with effect-size contrasts ranging from .08 (Communication) to .10 (Comprehension), although the contrasts were not statistically significant. For the SPELL+EPD and BAU contrast, results show that children in the SPELL+EPD condition outperformed those receiving BAU on the Rhyme Detection ($d = 0.25$), Deletion ($d = 0.31$) and Letter Knowledge ($d = 0.17$) subtests and on the Pre-literacy Composite ($d = 0.27$). For the SPELL+HOME and BAU contrast, children in the former condition outperformed those in the latter on the Deletion ($d = 0.16$) and Letter Knowledge ($d = 0.21$) subtests and on the Pre-literacy Composite ($d = 0.21$); no other comparisons were significant. With the exception of Sound Discrimination, which generally showed no effect, there was not a clear relationship between the existence of ceiling effects and the effect size for measures.

When comparing the three SPELL conditions to one another using pairwise comparisons, only one reached statistical significance, namely the Sound Discrimination measure. For this comparison, the SPELL+EPD condition was favored over the SPELL condition ($ds = 0.07$ and -0.04 , respectively, $p = .044$). All other comparisons were not statistically significant. Therefore, although the SPELL+EPD condition had the highest effect sizes relative to BAU, it did not distinguish itself from the other two SPELL conditions with respect to impacts on children's language and literacy skills. Thus, findings suggest that the SPELL+HOME and SPELL+EPD supplements did not have practical relevance for enhancing effects across the whole sample.

Effect Moderation by Child Characteristics

The third research question sought to determine whether non at-risk and at-risk children (indexed here as low maternal education, see also Table 1) and DLL children, would benefit from the intervention equally well. To address this question, two HLMs (one per each of the two composite outcomes [Pre-literacy Composite, Language Composite]) were estimated in which child-level covariates of interest (DLL status, maternal education) were included both as a main effect and as interactions with condition. The continuous variables were grand mean centered. For the sake of parsimony and power for analysis of interactions, we collapsed all three SPELL conditions into a single intervention condition for these analyses, given that the three interventions did not distinguish themselves from one another with respect to child outcomes. Additional covariates included in the models included children's pretest scores, age, gender, and intervention.

Table 4 provides the unstandardized coefficients (B), standard errors, and p -values for each of the predictor variables and interaction terms. Note that in these models, the estimated effect of SPELL is the difference between the BAU condition and the pooled intervention group for a non-DLL boy with average pretest score, and average maternal education. For the interaction terms, the interactions including pretest, age, gender, DLL status, and maternal education describe how the

slopes of these variables change if we move from BAU to the intervention conditions. For the interaction including the DLL variable, the interaction can be best interpreted as the change of the SPELL effect when moving from a non-immigrant to an immigrant child.

Among the variables included in the first model, for which the Pre-literacy Composite served as the outcome variable, children's pretest scores ($p < .001$) and maternal education ($p < .001$) served as unique, significant predictors of the outcome variable; in addition, intervention served as a unique, significant predictor of the outcome ($p < .001$) as well as the pretest scores ($p = .044$).

Figure 1 in Supplementary Online Material depicts the nature of the interaction between pretest scores and intervention effect (which was the only significant interaction) showing that children with lower pretest scores gained more from the intervention than children with higher pretest scores

Variables included in the second model, with Language Composite as the outcome variable, both pretest ($p < 0.001$), age ($p < .001$), DLL status ($p < .001$) and maternal education ($p = .001$) served as independent, significant predictors of the outcome variable. Neither intervention ($p = .224$) nor any of the other variables interacted significantly with outcome. In sum, neither maternal education level (SES) nor DLL background significantly moderated the SPELL treatment effects.

Relations between Implementation Fidelity and Child Outcomes

The fourth research question considered the role of implementation fidelity, particularly children's exposure to the intervention (number of lessons in which the child participated), as a potential variable of influence on children's outcomes. To examine the relations between children's intervention exposure and intervention outcomes, two additional HLMs were estimated, one for each of the posttest composites (Pre-literacy Composite Outcome, Language Outcome). For these models, only children within the three SPELL conditions were included ($n = 3,982$ for the Pre-literacy Composite outcome, $n = 3,986$ for the Language Composite outcome). Predictors included in the base model were the corresponding pretest, age, gender, DLL status, and maternal education

as well as the interaction terms for intervention exposure (number of lessons completed at the child level) X DLL status and intervention exposure X maternal education. The interactions allowed us to consider whether increased exposure to SPELL for either all or a subsample of children was associated with improved outcomes. The interaction terms describe how this number changes if we consider a non-DLL versus DLL child (intervention exposure X DLL status) or if we increase maternal education by one level (intervention exposure by maternal education); that is, if the effect of increased exposure varies depending on immigrant status or maternal education.

Results, as presented in Table 5, showed that among base model variables included in the first model, for which the Pre-literacy Composite served as the outcome variable, children's pretest scores ($p < .001$), gender ($p = .006$), DLL status ($p < .001$), maternal education ($p < .001$) and intervention exposure ($p < .001$) served as a unique, significant predictors of the outcome variable. Neither DLL status ($p = .882$) nor maternal education ($p = .403$) interacted significantly with intervention exposure to predict the outcome variable.

Among base model variables included in the second model, for which the Language Composite served as the outcome variable, children's pretest scores ($p < .001$), age ($p < .001$), DLL status ($p < .001$), maternal education ($p < .001$) and intervention exposure ($p < .001$) served as a unique, significant predictor of the outcome variable. DLL status interacted significantly with intervention exposure ($p = .04$) to predict the outcome, whereas maternal education did not ($p = .66$). Figure 2 in Supplementary Online Material illustrates the larger effect of exposure for DLL children's Language Composite scores than for non-DLL children. Fidelity accounted for 9% of the variation in outcomes in Rhyme, 14% in deletion, 17% in Letter identification and 19% of the variation in outcome scores in Vocabulary. These results indicate that the strength of the intervention was reduced by the lower implementation rate. We found significant effects on Language Composite scores for children who were exposed to more than 20 lessons corresponding

to 10 hours of intervention (SPELL condition [$d = 0.24$] and the SPELL+EPD condition [$d = 0.19$] for the 21+ subsample) which further suggests that had the dosage of intervention been higher, we might have seen higher, and significant, overall effects for language measures. Correlational analyses indicate that child background characteristics predict amount of exposure, in particular DLL status showing that DLL children were exposed to fewer lessons than non-DLL children ($r = -.15$) while at the same time the effect of exposure was stronger for DLL children than other children. It should be acknowledged that all results for fidelity are correlational, not experimentally manipulated.

Discussion

The present study is distinctive in that it was designed to investigate the extent to which language and pre-literacy interventions positively affect young children when taken to scale under real-world circumstances with an unselected sample in a European setting. This investigation was conducted throughout the country of Denmark, which distinguishes it from prior efficacy and effectiveness trials largely situated within the United States. Several contrasts to the United States' context warrant reference in understanding the contribution of this work. First, Danish childcare centers are publicly funded and available at-scale; to this end, nearly all preschool-aged children attend, thus allowing us to examine effects for a general population of youngsters instead of children attending programs with targeted enrollment (e.g., Head Start in the US). Second, these centers are situated in mostly residential areas across Denmark, thus their socio-economic composition is solely determined by the composition of the neighborhood. Thus Danish daycares typically serve a socioeconomically diverse group of children, even though centers with a high proportion of at-risk children also exist. In such a context, we were able to investigate intervention effects for all children within these daycare settings, while also exploring whether certain groups of children, as a function of family and child characteristics, may derive particular benefit. Finally, the

explicit language and pre-literacy instructional approach utilized in the SPELL intervention is not one typically utilized in the Danish context, which tends to use child-oriented pedagogies and emphasize social development with little to no attention to academically oriented outcomes. The trial results are compelling in this regard, as it helps us to understand whether interventions developed in one context (the US) can be applied to a novel and distinct context.

The principal finding of this effectiveness study is that the SPELL intervention had significant impacts on two fundamental pre-literacy skills, phonological awareness and letter knowledge, both of which were explicitly targeted by the intervention. Children in the intervention classrooms outscored children in the BAU classrooms by about 0.25 of a standard deviation unit, on average for these measures. Though children in the three SPELL conditions also scored higher ($d = 0.05 - 0.10$) than those in the BAU condition on vocabulary, also targeted by the intervention, none of these contrasts reached significance. These effects are generally comparable to results recently reported for a larger-scale effectiveness trial for Read It Again (the intervention from which SPELL was adapted) involving preschoolers from low-SES backgrounds ($n = 406$) in Appalachian communities (Mashburn, Justice, McGinty, & Slocum, 2015). Effect-size estimates indicated that exposure to RIA improved children's pre-literacy skills on measures of alphabet knowledge and print concepts on the magnitude of 0.18 and 0.25 standard deviation units, respectively, as compared to a control condition (Justice et al., 2010). Similar to the present study, no positive effects were seen on measures of early language skill (i.e., vocabulary). A difference between the results of the SPELL trial and the Mashburn study, however, concerns effects on phonological awareness. Danish children exposed to SPELL showed significant improvements in phonological awareness relative to controls, but no such effect was found for American children exposed to Read It Again. It may be that SPELL was particularly beneficial to children's development of phonological awareness because it is seldom taught within business-as-usual classrooms in Denmark. In contrast, within the

US, many preschool standards and curricula explicitly emphasize the importance of teaching phonological awareness to preschool-aged children.

While positive effects on the pre-literacy skills were observed in the present study for children exposed to SPELL, the effect sizes observed were considerably lower than those obtained in prior efficacy trials, which range from 0.25 for language outcomes to 0.87 for phonological awareness (see National Early Literacy Panel, 2008, Chapter 3). Similarly-sized effects were seen in an early efficacy study involving Read It Again (e.g., 0.35 for grammar, 0.44 for print concepts; see Justice et al., 2010). The present findings suggest, as would be expected, that taking systematic explicit language and pre-literacy interventions to scale is likely to decrease the size of the effects observed, but also that favorable effects are still observed in real-world contexts. Indeed, benchmarking the effects against the seven randomized controlled effectiveness trials identified (see Introduction and Appendix), the effect sizes are quite comparable, even though the SPELL intervention had a duration of just 20 weeks (vs. school year for the other effectiveness studies).

An additional and important result derived from this study, and one that stands in contrast to previous research (see e.g. National Early Literacy Panel, 2008, Chapter 3), is that neither SES, indexed as maternal education, nor DLL background significantly moderated the intervention effect pooled across the three intervention groups. Unlike many prior studies, the current effectiveness trial was sufficiently powered to assess moderating variables, with more than 6,000 children participating. Therefore, the lack of moderation of intervention effects by children's SES or language status is an important contribution to the literature, as it suggests that low-SES children and DLL benefited to the same extent as other, more advantaged children. Stated conversely, the advantaged children in the sample benefited similarly to the more at-risk participants. In emphasizing this finding, however, we must caution that such effects may not generalize outside of the Danish context, as children who are considered low-SES and DLLs in other contexts may differ

in important ways from Danish children who are low-SES and/or are DLLs. Nonetheless, given that intervention effects were not contingent on key child-level characteristics, such as socioeconomic and linguistic background, this adds to the practical relevance of the effects. As argued in NICHD (2006), a phenomenon with a small-to-modest effect on many individuals may have as large an impact collectively as a phenomenon with a large effect on a few individuals.

A third finding of interest in this effectiveness study concerns issues related to implementation. Across the three groups of educators using the experimental intervention, low levels of implementation fidelity were observed. This was not unexpected for several reasons. First, from a design perspective, effectiveness studies are designed in such a way so as to reduce the researchers' level of control over implementation of the intervention. With relatively low levels of researcher involvement in and control over implementation, the primary goal of an effectiveness trial is to determine whether effects seen in tightly controlled investigations are maintained. Second, from a contextual perspective, Danish early educators are not accustomed to using explicit, systematic interventions, and implementation of structured curricula in these settings is atypical. Therefore, the educators may be unaccustomed to use explicit and systematic practices. The organizational structure in many Danish daycare centers is not supportive of implementing structured interventions like SPELL, which may utilize small-group activities and structured routines. When adopting novel language- and pre-literacy instructional interventions, it can take time for early educators to feel comfortable applying the approach, and fidelity may be reduced as a result (Pence, Justice, & Wiggins, 2008). More than this, however, the early educational context within Denmark emphasizes democratic values and social development, with academic foci treated warily. Therefore, some early educators are reluctant to engage in explicit instruction that is "school-like." Thus, it is not completely surprising that children in this study were exposed to an average of only 25 of the 40 planned 30-minute SPELL lessons, corresponding to about 12.5 hours

of intervention in total. These relatively low levels of fidelity may attenuate effects of the intervention on targeted child outcomes, but also represent the level of implementation that may be realistically expected in every circumstances.

Although low levels of fidelity were observed overall, there was also considerable variation across educators in this respect, resulting in some children receiving relatively higher exposure to SPELL than other children. In turn, this level of exposure was found to be related to outcomes: children appeared to benefit more from higher levels of exposure to the intervention with respect to both language and pre-literacy development. This result, which was not anticipated a priori, suggests the importance of exploring more carefully those organizational-, educator-, and classroom-level factors that may affect children's exposure to the intervention. Indeed, qualitative feedback provided by educators in this study suggested that planning and lack of support from the local leadership may have played a substantial role for the low implementation.

Implementation-science research can provide an important mechanism for exploring the extent to which these factors may present barriers to educators' implementation of novel language- and pre-literacy interventions. The Theoretical Domains Framework (see French et al., 2012) is one approach used within the field of implementation science to identify determinants of behavior that negatively affect intervention implementation. For instance, educators in the SPELL trial may have perceived that use of an explicit, systematic intervention with young children may have negative effects on their development, or may detract from educational focus on other important instructional objectives. The Theoretical Domains Framework can be used to both identify behavioral determinants that detract from implementation, but also empirically informed approaches that work to affect these determinants. As early interventions are increasingly taken to scale outside of the context in which they were developed, efforts to understand and improve implementation processes should be a central focus (see Fixsen et al., 2005).

Several additional findings also warrant discussion. In this study, we also tested the extent to which two additional components improved the effectiveness of SPELL. Neither of the two components resulted in significantly improved outcomes in children. Even though there were nominally higher effect sizes for phonological awareness subscales in SPELL+EPD, the group differences reached significance in just a single case (SPELL vs. SPELL+EPD for Sound Discrimination). Similarly to the results of Lonigan et al. (2011) and Assel et al. (2007) our study therefore only found limited evidence for an effect of extended PD in combination with a curriculum. However, inspecting effect sizes across all subgroups show a consistent pattern of nominally higher effect sizes among all children, though highest in children from lowest-SES homes, perhaps indicating that providing educators with a language profile of children and specific training in recognizing and practicing differentiation strategies may assist educators in meeting the needs of children, in particular those with at-risk backgrounds. Given that children in the SPELL+EPD condition were exposed to the fewest number of lessons, it is possible that the nominal advantage of the SPELL+EPD group, in particular for children from low-SES homes, would have been larger and significant, had they received the same number of lessons as the other two intervention groups. This hypothesis is further supported by the fact that even though the teachers completed fewer lessons, the quality of the completed lessons was somewhat higher, teachers used more scaffolding strategies and addressed learning objectives more frequently in this condition. Together, these results suggest that combining a cost-effective PD component to a systematic explicit curriculum, based on elements from a practice-based approach (Hamre et al., 2012; Landry et al., 2009; 2001) is potentially useful at scale and merits further investigation. However, research on professional development and implementation also suggests that outcomes can be improved by on-going implementation support (e.g. Fixen et al., 2005; Zaslow et al., 2010).

Turning to the SPELL+HOME condition, the effect sizes were generally no larger, and even sometimes slightly smaller, in magnitude compared to the basic SPELL condition, similar to the findings of Manz et al (2010). An inspection of effect sizes in subgroups of children did, however, indicate that in the SPELL+Home condition, children from lower-SES homes had nominally higher effect sizes than children from the high-SES homes, suggesting that providing more learning opportunities to children via parallel activities in the home may benefit children from low-SES homes. This result is in contrast to results from the two meta-analyses by van Steensel et al. (2010) and Mantz et al (2010) where effect sizes were highest for non-at-risk families. To this end, a high priority for future effectiveness studies of home educational components is to investigate both their overall effect and differential effects for subgroups while holding the center-based curriculum constant. However, the results of this study already suggest that instead of universal provision of books and activities to all children, further effectiveness studies should explore how parents could be engaged differentially depending on the needs of the children. There is some anecdotal evidence for this suggestion. Some educators commented that children whose parents appeared to deliver the SPELL+HOME intervention with high fidelity seemed to lose interest in SPELL during the course of intervention, perhaps from an “over-dosage” of these learning opportunities.

Several limitations of this study warrant mention. First, as is typical of real-world studies, attrition was an issue, although an analysis based on WWC guidelines suggests that this is unlikely to have biased the results. Second, as noted previously, some of the subscales exhibited ceiling effects at pretest, such that a percentage of children had no possibility of showing improvement on posttests. This may have led to an underestimation of effects. Third, the assessments administered to children were administered by their child-care providers, and not trained research staff. While we do not believe this biased results, as the educators were blind to the various conditions in the study as well as the hypotheses being tested, confidence in these results would be boosted with

independent assessments of children. Additionally, the absence of verifiable fidelity of the Home component is also a limitation of the study. Finally, until long-term follow-up data are available, it is unknown whether the obtained effects will result in improved reading skills.

In sum, this randomized effectiveness study was motivated by the need to enhance language and pre-literacy skills during the preschool period, with a focus on those skills known to be foundational for later educational success. Although children from low-SES/DLL homes are at higher risk, poor development can occur for children throughout the population, and improved educational attainment is also possible even for children doing relatively well. For these reasons it is vital to develop effective interventions for language and pre-literacy development which are of benefit for all children. The results of this study demonstrated that even when implemented at full scale in a heterogeneous sample, a semi-scripted, sequenced, educator-managed and skill-focused curriculum supplement requiring little training prior to implementation and no costly ongoing PD and coaching elements can be an economically reasonable means for improving pre-literacy skills to the benefit of all children in childcare, even in an educational setting like Denmark where educators have little training in supporting pre-literacy skills.

While the intervention significantly impacted children's pre-literacy skills, significant impacts were found for language skills only in children who were exposed to at least half of the 40 lessons. This highlights the importance of investigating organizational, administrative, and personnel training factors, among others, in the local educational context that can support higher implementation fidelity of interventions such as SPELL. Despite the challenges of conducting rigorous effectiveness trials, this approach yields results that are highly relevant for policy makers, since these are designed to answer the most relevant questions for decision and budget makers.. There is a great need for rigorous effectiveness studies with high level of both internal and external validity to test the generalizability of well-known curricula with established efficacy.

Table 1

Baseline Characteristics of Recruited Children, Childcare Centers, Classrooms, Childcare Educators, and Groups in Four Conditions

	BAU	SPELL	SPELL+EPD	SPELL+HOME
Childcares, <i>n</i> ^a	36	38	34	34
Children, <i>n</i>	1,850	1,943	1,696	1,587
Classrooms per childcare, mean	2.6	2.7	2.7	2.6
Educators per classroom, mean	-	1.9	1.8	1.8
Groups per educator, mean	-	1.8	1.9	1.8
Children per group, mean	-	5.4	5.4	5.4
Educator characteristics				
% > 10 yr. teaching experience	62.1	58.5	54.3	54.8
% BA or higher education	78.3	74.3	77.0	66.7
Child Characteristics				
% boy	55.3	50.1	52.4	52.1
Age in months (<i>SD</i>)	53.8 (10.5)	52.6 (10.8)	53.9 (10.5)	52.9 (10.8)
Maternal education ^a				
% Low	15.5	14.7	14.7	17.5
% Low-Mid	37.0	36.2	37.4	37.3
% High-Mid	30.9	32.4	34.2	32.0
% High	16.6	16.8	13.6	13.2
Family income as mean quintile (<i>SD</i>)	2.9 (1.4)	2.9 (1.4)	2.9 (1.3)	2.8 (1.4)
% DLL background	10.7	11.5	12.9	12.1
Range of % DLL per day care	1-92	1-87	1-78	2-88

Note: Group characteristics were not significantly different (uncorrected *p*-values > .2).

^aThe unevenness was caused by 1) the formal merger around intervention start of two neighboring childcares (SPELL and SPELL+HOME), who therefore both did the SPELL condition, 2) one administrative error in the substitution of a childcare in SPELL+EPD by which the substitute childcare did the SPELL condition, and 3) the dropout of two childcares (SPELL+EPD, SPELL+HOME) following randomization and just prior to the start of intervention. ^bCf. education percentages for Denmark's population of women age 25-49 years: Low (Primary school), 20.0%; Low-Mid (High school, Vocational education), 39.0%; High-Mid (e.g. professional BA such as teacher), 25.7%; High (BA and advanced university education), 15.2%.

Table 3

Effect-Size Estimates for Three SPELL Conditions Compared to BAU

Variable (BAU <i>n</i>)	SPELL	SPELL+EPD	SPELL+HOME
Pre-literacy Composite (<i>n</i> = 1,324)	0.22** (<i>n</i> = 1,533)	0.27** (<i>n</i> = 1,317)	0.21** (<i>n</i> = 1,132)
Sound Discrimination (<i>n</i> = 407)	-0.05 (<i>n</i> = 516)	0.07 (<i>n</i> = 441)	0.01 (<i>n</i> = 378)
Rhyme (<i>n</i> = 722)	0.20* (<i>n</i> = 850)	0.25** (<i>n</i> = 749)	0.11 (<i>n</i> = 602)
Deletion (<i>n</i> = 867)	0.27** (<i>n</i> = 996)	0.31** (<i>n</i> = 839)	0.16* (<i>n</i> = 735)
Letter Identification (<i>n</i> = 812)	0.19** (<i>n</i> = 959)	0.17* (<i>n</i> = 777)	0.21** (<i>n</i> = 667)
Language Composite (<i>n</i> = 1,329)	0.16 (<i>n</i> = 1,539)	0.13 (<i>n</i> = 1,317)	0.04 (<i>n</i> = 1,130)
Vocabulary (<i>n</i> = 1,325)	0.09 (<i>n</i> = 1,535)	0.09 (<i>n</i> = 1,315)	0.05 (<i>n</i> = 1,108)
Comprehension (<i>n</i> = 1,310)	0.10 (<i>n</i> = 1,523)	0.08 (<i>n</i> = 1,304)	0.04 (<i>n</i> = 1,106)
Communication (<i>n</i> = 1,250)	0.08 (<i>n</i> = 1,511)	0.05 (<i>n</i> = 1,270)	0.01 (<i>n</i> = 1,018)

p* < .05, *p* < .01.

Table 4

HLM Results for Predicting Pre-literacy Composite and Language Outcomes from Child-Level Characteristics

	Pre-literacy Outcomes			Language Outcomes		
	B	(SE)	p-value	B	(SE)	p-value
Base model						
Pretest	0.50	(0.02)	<.001	0.72	(0.02)	<.001
Age	0.03	(0.02)	.201	-0.08	(0.02)	<.001
Gender	-0.05	(0.04)	.138	-0.04	(0.03)	0.108
DLL Status	-0.10	(0.08)	.188	-0.24	(0.06)	<.001
Maternal Education ^a	0.09	(0.02)	<.001	0.06	(0.02)	.001
Interaction model						
Intervention (SPELL)	0.19	(0.05)	<.001	0.06	(0.05)	.224
SPELL X Pretest	-0.05	(0.03)	.044	0.00	(0.00)	.929
SPELL X Age	0.00	(0.00)	.466	-0.00	(0.00)	.748
SPELL X Gender	-0.00	(0.04)	.925	-0.05	(0.03)	.118
SPELL X DLL	-0.16	(0.09)	.066	-0.05	(0.07)	.505
SPELL X Maternal Education ^a	-0.02	(0.03)	.361	-0.01	(0.02)	.649

Note. ^aMaternal Education describes the effect in non-DLL children only.

Table 5

HLM Results for Predicting Children's Intervention Outcomes from Fidelity Variables

	Pre-literacy Outcomes			Language Outcomes		
	B	(SE)	p	B	(SE)	p
Base model						
Pretest	4.5	(0.1)	<.001	7.2	(0.1)	<.001
Age	0.0	(1.2)	.118	-0.1	(0.0)	<.001
Gender	-0.5	(0.2)	.006	-0.0	(0.1)	.744
DLL status	-2.4	(0.4)	<.001	-2.6	(0.3)	<.001
Maternal education*	0.7	(0.1)	<.001	0.4	(0.1)	<.001
Intervention Exposure	0.1	(0.0)	<.001	0.1	(0.0)	<.001
Interaction model						
Intervention Exposure X DLL Status	0.0	(0.0)	.882	0.1	(0.0)	.042
Intervention Exposure X Maternal Education	-0.0	(0.0)	.403	0.0	(0.0)	.665

Note. *Maternal education describes the effect in DLL children only.

References

- Al Otaiba, S., Connor, C. M., Folsom, J. S., Greulich, L., Meadows, J., & Li, Z. (2011). Assessment data-informed guidance to individualize kindergarten reading instruction: Findings from a cluster-randomized control field trial. *The Elementary school journal*, 111, 535-560.
- Assel, M. A., Landry, S. H., Swank, P. R., & Gunnewig, S. (2007). An evaluation of curriculum, setting, and mentoring on the performance of children enrolled in pre-kindergarten. *Reading and writing*, 20, 463-494.
- Bierman, K. L., Domitrovich, C. E., Nix, R. L., Gest, S. D., Welsh, J. A., Greenberg, M. T., . . . Gill, S. (2008). Promoting academic and social-emotional school readiness: The Head Start REDI program. *Child Development*, 79, 1802-1817.
- Bleses, D., Højen, A., Dybdal, L., Dale, P., Justice, L., Piasta, S., . . . Aktürk Ari, B. (2014). A randomized controlled trial of a large scale systematic and explicit language and literacy intervention in Danish daycares: The SPELL study. *British Journal of Education, Society & Behavioural science*, 4, 504-1524.
- Bleses, D., Højen, A., Jørgensen, R. N., Jensen, K. Ø., & Vach, W. (2010). Sprog vurdering af 3-årige (09) - karakteristika og risikofaktorer. *Center for Child Language, e-prints*, 10.
- Bleses, D., Vach, W., Jørgensen, R. N., & Worm, T. (2010). The internal validity and acceptability of the Danish SI-3: a language screening instrument for 3-year-olds. *Journal of Speech, Language, and Hearing Research*, 53, 490-507.
- Bornstein, M. H., Hahn, C. S., Putnick, D. L., & Suwalsky, J. T. (2014). Stability of core language skill from early childhood to adolescence: A latent variable approach. *Child Development*, 85, 1346-1356.

- Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly*, 25, 166-176.
- Buyse, V., Sparkman, K. L., & Wesley, P. W. (2003). Communities of practice: Connecting what we know with what we do. *Exceptional children*, 69, 263-277.
- Buyse, V., Peisner-Feinberg, E., Páez, M., Hammer, C. S., & Knowles, M. (2014). Effects of early education programs and practices on the development and learning of dual language learners: A review of the literature. *Early Childhood Research Quarterly*, 29, 765-785.
- Collins, M. F. (2010). ELL preschoolers' English vocabulary acquisition from storybook reading. *Early Childhood Research Quarterly*, 25, 84-97.
- Danish Ministry for Social Affairs. (2015). Tal og statistik om dagtilbud. Retrieved February 17, 2015, from <http://sm.dk/arbejdsmrader/dagtilbud/Tal%20og%20statistik>
- Danmarks Statistik. (2014). Indvandrere i Danmark 2014. Copenhagen.
www.dst.dk/publ/indvandreridk.
- Dockrell, J. E., Stuart, M., & King, D. (2010). Supporting early oral language skills for English language learners in inner city preschool provision. *British Journal of Educational Psychology*, 80, 497-515.
- Dunn, L. M., & Dunn, D. (2012). *Peabody Picture Vocabulary Test, (PPVT-4)*. Johannesburg: Pearson Education Inc.
- Fischel, J. E., Bracken, S. S., Fuchs-Eisenberg, A., Spira, E. G., Katz, S., & Shaller, G. (2007). Evaluation of curricular approaches to enhance preschool early literacy skills. *Journal of Literacy Research*, 39, 471-501.
- Gartlehner, G., Hansen, R. A., Nissman, D., Lohr, K. N., & Carey, T. S. (2006). Criteria for distinguishing effectiveness from efficacy trials in systematic reviews. *Technical Reviews*, 12.

- Hamre, B. K., Pianta, R. C., Burchinal, M., Field, S., LoCasale-Crouch, J., Downer, J. T., . . . Scott-Little, C. (2012). A Course on Effective Teacher-Child Interactions Effects on Teacher Beliefs, Knowledge, and Observed Practice. *American Educational Research Journal*, 49, 88-123.
- 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.
- Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, 2, 88-110.
- Justice, L. M., Kaderavek, J. N., Fan, X., Sofka, A., & Hunt, A. (2009). Accelerating preschoolers' early literacy development through classroom-based teacher-child storybook reading and explicit print referencing. *Language, Speech, and Hearing Services in Schools*, 40, 67-85.
- Justice, L. M., & McGinty, A. S. (2012). Read It Again Pre-K. *A preschool curricular supplement to promote language and literacy foundations*. Retrieved August, 18, 2012.
- 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.
- Landry, S. H., Anthony, J. L., Swank, P. R., & Monseque-Bailey, P. (2009). Effectiveness of comprehensive professional development for teachers of at-risk preschoolers. *Journal of Educational Psychology*, 101, 448-465.
- Landry, S. H., Swank, P. R., Anthony, J. L., & Assel, M. A. (2011). An experimental study evaluating professional development activities within a state funded pre-kindergarten program. *Reading and writing*, 24, 971-1010.
- Lonigan, C. J., Farver, J. M., Phillips, B. M., & Clancy-Menchetti, J. (2011). Promoting the development of preschool children's emergent literacy skills: A randomized evaluation of a

literacy-focused curriculum and two professional development models. *Reading and writing*, 24, 305-337.

Manz, P. H., Hughes, C., Barnabas, E., Bracaliello, C., & Ginsburg-Block, M. (2010). A descriptive review and meta-analysis of family-based emergent literacy interventions: To what extent is the research applicable to low-income, ethnic-minority or linguistically-diverse young children? *Early Childhood Research Quarterly*, 25, 409-431.

Marulis, L. M., & Neuman, S. B. (2013). How vocabulary interventions affect young children at risk: A meta-analytic review. *Journal of Research on Educational Effectiveness*, 6, 223-262.

National Early Literacy Panel. (2008). *Developing early literacy*. Washington, DC: National Institute for Literacy.

Neuman, S. B., Newman, E. H., & Dwyer, J. (2011). Educational effects of a vocabulary intervention on preschoolers' word knowledge and conceptual development: A cluster-randomized trial. *Reading Research Quarterly*, 46, 249-272.

NICHD Early Child Care Research Network (2006). Child-care effect sizes for the NICHD study of early child care and youth development. *American Psychological Association*, 61, 99-116.

O'Brien, R. A., et al. (2012). Mixed methods analysis of participant attrition in the nurse-family partnership. *Prevention Science*, 13, 219-228.

OECD (2010). PISA 2009 at a Glance. OECD.

Powell, D. R., Diamond, K. E., Burchinal, M. R., & Koehler, M. J. (2010). Effects of an early literacy professional development intervention on head start teachers and children. *Journal of Educational Psychology*, 102, 299-312.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1): Sage.

- Schwartz, D., & Lellouch, J. (1967). Explanatory and pragmatic attitudes in therapeutical trials. *Journal of chronic diseases*, 20, 637-648.
- Sheridan, S. M., Edwards, C. P., Marvin, C. A., & Knoche, L. L. (2009). Professional development in early childhood programs: Process issues and research needs. *Early Education and Development*, 20, 377-401.
- Singal, A. G., Higgins, P. D. R., & Waljee, A. K. (2014). A primer on effectiveness and efficacy trials. *Clinical and translational gastroenterology*, 5, e45. doi: 10.1038/ctg.2013.13
- van Steensel, R., McElvany, N., Kurvers, J., & Herppich, S. (2011). How effective are family literacy programs? Results of a meta-analysis. *Review of Educational Research*, 81, 69-96.
- Van Tuijl, C., Leseman, P. P., & Rispens, J. (2001). Efficacy of an intensive home-based educational intervention programme for 4-to 6-year-old ethnic minority children in the Netherlands. *International Journal of Behavioral Development*, 25, 148-159.
- Verhoeven, L., & Van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology*, 22, 407-423.
- Verhallen, M. J., & Bus, A. G. (2010). Low-income immigrant pupils learning vocabulary through digital picture storybooks. *Journal of Educational Psychology*, 102, 54-61.
- Wasik, B. A., & Hindman, A. H. (2011). Improving vocabulary and pre-literacy skills of at-risk preschoolers through teacher professional development. *Journal of Educational Psychology*, 103, 455-469.
- Whitehurst, G. J., Arnold, D. S., Epstein, J. N., Angell, A. L., Smith, M., & Fischel, J. E. (1994). A picture book reading intervention in day care and home for children from low-income families. *Developmental psychology*, 30, 679-689.
- Williams, K. T. (2007). *EVT-2: Expressive vocabulary test*: Pearson Assessments.

APPENDIX

Overview of Studies Reporting Randomized Controlled Trials with Effectiveness Characteristics

Study	Intervention	Sample size	ES Phonological awareness	ES Letter Identification	ES Vocabulary
Fichel (2007)	Let's Begin with the Letter people; Waterford Early Reading Program	507	-	NS	NS ⁴
Landry et al. (2009)	4 PD ¹ variants	1,789	0.14	0.15*, 0.47**	0.15
Lonigan et al., (2011)	The literacy Express Preschool Curriculum + PD	735	0.11, 0.14, 0.17	0.18, 0.32*, 0.45** ³	0.20, 0.27*, 0.32** ⁶
Neuman et al. (2011)	World of Words	604	-	-	NS
Powell et al. (2010)	PD program (coaching)	759	0.22*	0.32**	-0.03 ⁴
Wasik & Hindman (2011)	Exceptional Coaching for Early Language and Literacy	541	0.22* ²	-0.07 ²	0.08 ^{2, 4}

¹PD=professional development; ²The effect sizes are calculated by the authors; ³Broader print knowledge; ⁴PPVT; ⁵Calculated based on e- and posttests for intervention and comparison group; ⁶Oral language (PLS-IV)

APPENDIX S1**Attrition**

Due to, e.g., illnesses, absences, and attrition, pretest and posttest was obtained for 5,359 also had pretest. The final sample contributing both pre- and posttest data was for BAU, 35 childcares, 87 classrooms, 1,343 children; for SPELL, 36 daycares, 96 classrooms, 1,548 children; for SPELL+EPD 31 childcares, 87 classrooms, 1,330 children; SPELL+HOME, 32 daycares, 82 classrooms, 1,138 children. All classrooms in each daycare were required to participate.

Child-level attrition rates for each condition from randomization to posttest and from pretest to posttest, respectively, were BAU, 27%, 17%; SPELL, 20%, 16%; SPELL+EPD, 22%, 16%; SPELL+HOME, 28%, 21%. Chi-square tests revealed that total child-level attrition rates differed significantly from randomization to posttest ($p = .010$) but not from pretest to posttest ($p = .414$) across the four conditions. Randomization to posttest attrition rates in subgroups defined by DLL-status, maternal education (high, low) or family income (high, low) (all $p > .20$) revealed significant or marginally significant differences for “low” groups but not “high” groups or DLL children.

Table SI-1. Number of childcares, classrooms, educators and children at two stages in the study, viz. randomized and posttest for whom register data were available. In addition, attrition rates in numbers and percentages are given, both as a total, attrition due to childcare attrition (embedded attrition) and attrition within remaining childcares.

Condition	Level	Recruitment N	Posttest N	Total attrition N	Childcare embedded attrition N	Attrition in remaining childcares N	Total attrition %	Childcare embedded attrition %	Attrition in remaining childcares %
Overall	Childcares	142	134	8			6		
Overall	Classrooms	379	352	27	17	10	7	4	3
Overall	Educators	627	561	66	32	34	11	5	5
Overall	Children	7,076	5,359	1,717	319	1398	24	5	20
Control	Childcares								
Basis	Childcares	36	35	1			3		
Home	Childcares	38	36	2			5		
EPD	Childcares	34	32	2			6		
Control	Classrooms	34	31	3			9		
Basis	Classrooms	92	87	5	2	3	5	2	3
Home	Classrooms	104	96	8	6	2	8	6	2
EPD	Classrooms	90	82	8	5	3	9	6	3
Control	Educators	93	87	6	4	2	6	4	2
Basis	Educators	109	101	8	4	4	7	4	4
Home	Educators	193	174	19	11	8	10	6	4
EPD	Educators	160	137	23	10	13	14	6	8
Control	Children	165	149	16	7	9	10	4	5
Basis	Children	1,850	1343	507	38	469	27	2	25
Home	Children	1,943	1548	395	98	297	20	5	15
EPD	Children	1,587	1138	449	115	334	28	7	21

Table SI-2. Number of childcares, classrooms, educators and children at two stages in the study, viz. pretest and posttest for whom register data were available. In addition, attrition rates in numbers and percentages are given, both as a total, attrition due to childcare attrition (embedded attrition) and attrition within remaining childcares.

Condition	Level	Recruitment N	Posttest N	Total attrition N	Childcare embedded attrition N	Attrition in remaining childcares N	Total attrition %	Childcare embedded attrition %	Attrition in remaining childcares %
Overall	Childcares	142	134	8			6		
Overall	Classrooms	379	352	27	17	10	7	4	3
Overall	Educators	627	561	66	32	34	11	5	5
Overall	Children	7,076	5,359	1,717	319	1398	24	5	20
Control	Childcares	36	35	1			3		
Basis	Childcares	38	36	2			5		
Home	Childcares	34	32	2			6		
EPD	Childcares	34	31	3			9		
Control	Classrooms	92	87	5	2	3	5	2	3
Basis	Classrooms	104	96	8	6	2	8	6	2
Home	Classrooms	90	82	8	5	3	9	6	3
EPD	Classrooms	93	87	6	4	2	6	4	2
Control	Educators	109	101	8	4	4	7	4	4
Basis	Educators	193	174	19	11	8	10	6	4
Home	Educators	160	137	23	10	13	14	6	8
EPD	Educators	165	149	16	7	9	10	4	5
Control	Children	1,850	1343	507	38	469	27	2	25
Basis	Children	1,943	1548	395	98	297	20	5	15
Home	Children	1,587	1138	449	115	334	28	7	21
EPD	Children	1,696	1330	366	68	298	22	4	18

Table SI-3. Attrition in subgroups

DLL*	N recruited (% of sample)	N posttested (% of sample)
BAU	190 (10.7)	96 (7.4)
SPELL	220 (11.5)	151 (10.0)
SPELL+EPD	208 (12.9)	134 (10.6)
SPELL+HOME	185 (12.1)	117 (10.7)

* DLL, dual language learners

Low education*	N recruited (% of sample)	N posttested (% of sample)
BAU	262 (15.5)	153 (12.5)
SPELL	265 (14.7)	207 (14.4)
SPELL+EPD	226 (14.7)	155 (12.9)
SPELL+HOME	252 (17.5)	176 (17.1)

* Lowest of four categories of education level

High education*	N recruited (% of sample)	N posttested (% of sample)
BAU	282 (16.6)	209 (17.0)
SPELL	304 (16.8)	220 (15.3)
SPELL+PD	209 (13.6)	160 (13.3)
SPELL+HOME	191 (13.2)	142 (13.8)

* Highest of four categories of education level

Low income*	N recruited (% of sample)	N posttested (% of sample)
BAU	386 (21.6)	237 (18.3)
SPELL	391 (20.5)	306 (20.2)
SPELL+EPD	339 (21.0)	242 (19.2)
SPELL+HOME	346 (22.6)	233 (22.3)

* Lowest of five categories of income level

High income*	N recruited (% of sample)	N posttested (% of sample)
BAU	269 (15.0)	211 (16.3)
SPELL	326 (17.1)	245 (16.2)
SPELL+EPD	213 (13.2)	165 (15.0)
SPELL+HOME	219 (14.3)	153 (14.0)

* Highest of five categories of income level

Below we provide information on the estimated standard deviations with 95% confidence intervals for each random effect in our model for each scale, as well as estimates of the standard deviation of the residuals. We can observe that for most scales the lower bound of the 95% confidence interval is distinctly above 0, and that there is evidence for a true contribution of these random effects to the overall variance. (Exceptions appear mainly for scales measured in a limited number of children, where this may just indicate a lack of power to come to precise estimates of the standard deviations). Moreover, the estimated standard deviations of the random effects at the different levels are often of similar magnitude, which makes it impossible to suggest levels we may omit. It can, of course, be discussed, whether it is necessary to allow different standard deviations in the intervention arms compared to the control arm. However, as we have two levels we can assess and model in the intervention groups only, and as we have to expect that the interventions may increase the overall variance, we feel it is essential to allow this to avoid that neglected differences may bias the estimation of the intervention effects.

institution (control arm): sd=0.089 (0.040,0.197)

institution (intervention arms): sd=0.140 (0.080,0.244)

classroom (control arm): sd=0.171 (0.123,0.238)

classroom (intervention arms): sd=0.127 (0.075,0.215)

educator (intervention arms): sd=0.156 (0.109,0.222)

group (intervention arms): sd=0.165 (0.127,0.213)

residuals (control arm): sd=0.611 (0.587,0.635)

residuals (intervention arms) : sd=0.569 (0.555,0.583)

Table SI-4. Results for Children who Completed both Pre- and Posttest. Study Measures by Condition and Total Number of Children Tested on Each Subscale

Measure	<i>n</i>	BAU		SPELL		SPELL+EPD		SPELL+HOME	
		Pre <i>M</i> (<i>SD</i>)	Post <i>M</i> (<i>SD</i>)	Pre <i>M</i> (<i>SD</i>)	Post <i>M</i> (<i>SD</i>)	Pre <i>M</i> (<i>SD</i>)	Post <i>M</i> (<i>SD</i>)	Pre <i>M</i> (<i>SD</i>)	Post <i>M</i> (<i>SD</i>)
Pre-literacy Composite	5,306	-0.02 (0.86)	0.39 (0.83)	-0.06 (0.87)	0.54 (0.82)	-0.02 (0.88)	0.58 (0.79)	-0.10 (0.85)	0.49 (0.82)
Sound Discrimination ^a	1,742	14.9 (2.1)	15.5 (1.3)	14.6 (2.3)	15.3 (1.9)	14.8 (2.1)	15.5 (1.3)	14.6 (2.0)	15.4 (1.4)
Rhyme Detection ^b	2,923	10.4 (4.0)	11.6 (3.9)	10.0 (4.0)	12.5 (3.5)	10.4 (4.1)	12.8 (3.5)	10.5 (4.0)	12.4 (3.7)
Deletion ^b	3,432	4.2 (5.8)	6.9 (6.8)	3.6 (5.3)	7.7 (6.4)	4.1 (5.5)	8.1 (6.4)	3.1 (4.8)	6.8 (6.4)
Letter Identification ^b	3,215	7.0 (3.6)	8.2 (3.5)	6.7 (3.6)	8.7 (3.3)	7.1 (3.7)	8.8 (3.3)	6.7 (3.7)	8.6 (3.5)
Language Composite	5,315	-0.04 (0.84)	0.49 (0.86)	-0.17 (0.83)	0.48 (0.86)	-0.10 (0.84)	0.50 (0.85)	-0.18 (0.80)	0.39 (0.83)
Vocabulary ^c	5,283	43.6 (19.6)	52.8 (17.5)	40.6 (20.0)	51.3 (18.6)	42.0 (20.3)	52.2 (18.2)	40.2 (19.4)	50.2 (18.4)
Comprehension ^c	5,243	17.8 (4.2)	19.4 (3.7)	17.2 (4.4)	19.2 (4.0)	17.5 (4.3)	19.4 (3.8)	17.2 (4.4)	18.9 (4.0)
Communication ^c	5,049	29.5 (6.4)	31.3 (6.7)	28.3 (6.1)	31.0 (6.2)	29.1 (6.4)	31.4 (5.9)	28.6 (6.4)	30.7 (6.0)

Note: ^aAdministered only to 3-year-olds; ^bAdministered only to 4-6-year-olds; ^cAdministered to all children.

Figure SI-1. Posttest Pre-literacy Composite scores (predicted) for children in BAU and SPELL condition as a function of children's pretest scores

Figure SI-2. Posttest Language Composite scores (predicted) for non-DLL and DLL children in the SPELL condition as a function of intervention exposure (mean number of lessons per child)