Nordic Journal of Literacy Research
Vol. 7 | No. 2 | 2021 | pp. 1–18

Peer-reviewed article

Improvement in oral language interventions: Differences and relation between effects on treatment-inherent measures and effects on standardized tests

Kristin Rogde,* Åste Mjelve Hagen, Arne Lervåg & Monica Melby-Lervåg
University of Oslo

Abstract
Whether the effects of an oral-language intervention is tested with measures of trained vocabulary (treatment-inherent tests) or standardized measures (treatment-independent tests) can have consequences for the mean effect size in meta-analyses. Moreover, based on a theory of transfer effects, effects on the trained words could serve as an index of how much benefit is gained by children from the intervention. We present a meta-analysis that assesses the differences and relation between the intervention effects of these two types of outcomes, trained vocabulary and standardized vocabulary tests.

The results show large effects on trained vocabulary, limited effects on standardized measures, and no clear relation between the two. The moderator analysis indicates that less instruction time is associated with larger effect sizes on trained vocabulary but that trained vocabulary is not a predictor of either standardized expressive or receptive vocabulary. Thus, in interventions and meta-analyses, it is important to distinguish between effects on trained vocabulary and standardized tests, and trained vocabulary effects does not necessarily transfer to standardized measures. This indicates that effects on trained vocabulary outcomes provide limited information when evaluating language interventions.

Keywords: *language intervention; meta-analysis; vocabulary*

Responsible editor: Michael Tengberg

Received: January, 2021; Accepted: June, 2021; Published: September, 2021

The importance of vocabulary in social interaction, understanding oral information, and reading is profound (Hjetland et al., 2017), and several meta-analyses have summarized how intervention programs can increase children's vocabulary development

*Correspondence: Kristin Rogde, e-mail: kristin.rogde@iped.uio.no

Citation: K. Rogde, Å. M. Hagen, A. Lervåg & M. Melby-Lervåg. "Improvement in oral language interventions: Differences and relation between effects on treatment-inherent measures and effects on standardized tests" Nordic Journal of Literacy Research, Vol. 7(2), 2021, pp. 1–18. http://dx.doi.org/10.23865/njlr.v7.2814

(Elleman et al., 2009; Marulis & Neuman, 2010, 2013; Mol et al., 2009; Rogde et al., 2019; Swanson et al., 2011). These meta-analyses have used different methodological approaches. In particular, the reviews vary in whether they merge researcher-developed treatment-inherent measures (typically vocabulary or listening comprehension tests with trained words embedded) and standardized measures. Despite their variation, all these meta-analyses show to some extent that oral language interventions are effective, yet little is known what actually improves and whether there is a relation between the size of improvement on the trained words and on the standardized measures.

In this paper, we present a meta-analysis that examines the extent to which the two measure types provide different mean effect sizes and whether there is a relation between the gains in researched inherent measures and in standardized tests. We also investigate the moderators of the mean effect sizes and the relation between these two measure types. The meta-analysis that we present in this paper is based on a subset of studies in a previous published Campbell systematic review (Rogde et al., 2019) and corresponding protocol (Rogde et al., 2016).

Studies of oral language interventions typically include researcher-created tests of trained vocabulary and/or standardized measures of vocabulary knowledge. These two outcome types are inherently different. Trained vocabulary outcomes are based directly on the treatment because the test measures knowledge of words that are trained in the instructional program (Slavin & Madden, 2011). In contrast, standardized outcomes are typically standardized tests that are not created for the specific intervention. Slavin and Madden (2011) refer to this type of outcome as a treatment-independent measure. These outcomes are important when researchers and practitioners want to know if learning transfers to a child's general vocabulary knowledge. Importantly, the two outcome types differ regarding the effect sizes that can be expected from vocabulary intervention. Since trained vocabulary measures test the understanding of instructed words to which only the treatment group has been exposed, the effects are obviously expected to be positive when compared to those in a control group (see Slavin & Madden, 2011 for a discussion). Conversely, the expectations of gaining effects on standardized tests are based on the theory that some components of an intervention program will lead to transfer effects on these vocabulary outcomes.

Reviews in the educational field present a confusing picture as they treat these types of outcomes differently in their syntheses and analyses. Some reviews have excluded treatment-inherent measures (e.g. Rogde et al., 2019; Slavin et al., 2011), others have included both types of measures yet made separate analyses of the outcomes (e.g. Elleman et al., 2009), and several have synthesized a mean treatment effect based on both types of outcomes (e.g. Swanson et al., 2011). Slavin and Madden (2011) find that the What Works Clearinghouse (2008a, 2008b) reading and math reviews averaged effect sizes from measures that clearly produced different estimates. Several examples of published meta-analyses on vocabulary intervention programs have also averaged the effect sizes from both trained vocabulary and standardized vocabulary

outcomes in their analyses. For instance, Marulis and Neuman (2010, 2013) report an overall effect size of d = 0.88 and d = 0.87, respectively, nearly one standard deviation (SD) on vocabulary measures in both studies. Mol et al. (2009) show an overall effect size of d = 0.62 for expressive vocabulary and d = 0.45 for receptive vocabulary, and Swanson et al. (2011) report d = 1.02 for vocabulary outcomes based on both trained and standardized tests. A recent review by Rogde et al. (2019) averages the effect sizes from 43 trials examining the effects of vocabulary instruction in educational settings. The overall effect size for vocabulary outcomes is solely based on standardized vocabulary measures, displaying an overall modest effect size of g = 0.13. Although these reviews are quite different in their inclusion criteria for eligible studies, the large differences in the synthesized overall effects are likely (at least partly) explained by their varied approaches to including trained vocabulary outcomes in their analyses of effects. In addition, previous meta-analyses of vocabulary instruction have included studies without an appropriate control group. In contrast, the current review solely include randomized controlled trials (RCTs) and quasi-experiments (QEs) with control groups and measures of baseline differences. Thus, prior reviews in the field have thus far not clearly contrasted the differing effects of these outcome measures.

Relation between treatment-inherent and standardized outcomes

While it is known that interventions targeting oral language can be effective, little is known about what drives these effects. An important question is whether gains in transfer measures (i.e. standardized tests) are related to the gains observed in specific words that are trained in the intervention. There are several theoretical reasons for expecting a relation between gains in the trained words and gains in the standardized measures. One reason why learning trained words can relate to the effects on the standardized measures is provided by the primitive elements theory (Taatgen, 2013). This suggests that transfer can happen because the intervention may improve children's ability to explain not only the specific words in which they are explicitly taught but also words in general. Thus, transfer occurs when the set of procedures learned with the trained words can also be utilized for untrained words (Taatgen, 2013).

Another theoretical reason for why transfer between trained and untrained words can occur is based on the vector semantics theory (Jurafsky & Martin, 2014). In line with this, transfer might occur because learning new words provides children with an improved understanding of the words that they already know. In line with the predictions from vector semantics, if a new word is similar either in syntax or semantics to a word already known by the child, this increases the probability that the child will learn the new word (Jurafsky & Martin, 2014). For example, if the child knows the word *damp*, it will be easier to learn the word *moist* since the two are semantically related. Similarly, learning the word *moist* may also offer the child a more nuanced understanding of *damp*.

Importantly, the opposite could also be the case, that is, it can also be deduced from the theory of broader transfer that no relation exists between the effects on trained words and on general language tests. In line with the theories of Bransford and Schwartz (1999) and Detterman (1993), it is not possible to detect transfer by training in one skill and testing whether it is directly applicable to another skill. This way of evaluating transfer is too restrictive because transfer occurs on a more general level and affects broader skills, such as critical thinking and meta-cognition (Bransford & Schwartz, 1999; Detterman, 1993; Lee, 1998).

A recent study that has examined the relation between trained words and standardized measures has found a relation between them on expressive language measures but not on receptive ones (Melby-Lervåg et al., 2020). The finding that the effects of training and the transfer effects are solely related to expressive measures could indicate that the primitive elements theory explains this transfer (Taatgen, 2013). Thus, one aspect that seems to improve and transfer to the untrained words is children's ability to develop procedures to provide better explanations of words. The primitive elements theory predicts that transfer is possible between tasks that share the same basic underlying structure and similar operators or procedures.

The current study

Our current study has two main aims. Our first objective is to examine what size of difference exists between gains in trained vocabulary and standardized vocabulary outcomes of oral language interventions. The hypothesis is that there would be a large difference in effects between trained vocabulary and standardized vocabulary outcomes. We also aim to perform a moderator analysis of the size of this difference in effects and to examine whether the duration of the instruction would relate differently to the size of the outcome effects of the two different outcomes. It could be that intervention programs of short duration would be associated with larger effect sizes than programs with longer duration for the trained vocabulary outcomes, while an opposite pattern would probably be the case for standardized outcomes. As for the trained words, the closer in time the instruction and testing occur, the more likely children are to remember the meanings of these words. In the longer time frame for instruction, more words are probably trained, and the test is likely to be based on a random selection of words for the entire period of instruction. In contrast, for the effect of standardized vocabulary outcomes, the longer duration of the intervention is assumed to be associated with higher effect sizes. The hypothesis is therefore that moderators related to the duration of the instruction would be differentially associated with the effects on trained vocabulary and general vocabulary outcomes.

Second, we aim to examine the relation between gains in trained vocabulary outcomes and in standardized outcomes. As earlier noted, large effects can be expected on trained vocabulary outcomes that relate to the direct instruction on word meanings in the programs. In contrast, gaining effects on standardized outcomes also depends

on whether the instruction has succeeded in providing children with knowledge that has enhanced their disposition to learn new words. If this is based on the transfer of knowledge, we could expect the studies with large gains in trained words to demonstrate the largest gains in the standardized measures. This would be in line with a recent study's results showing that effects on standardized measures are mediated by effects on trained words (Melby-Lervåg et al., 2020). We also aim to conduct a moderator analysis concerning this relationship, that is, whether the relation between gains in trained words and in standardized measures would be stronger in studies using expressive rather than receptive outcome measures. Because the transfer effects in the study by Melby-Lervåg et al. (2020) have been generated through expressive (not receptive) measures, we would expect to find a stronger relation between trained words and expressive standardized measures.

Our review aims to answer the following questions

1) What difference exists between gains in trained vocabulary (treatment-inherent tests) and in standardized vocabulary outcomes (treatment-independent tests) in oral language interventions?
2) Does the amount of the instruction relate differently to the effects of trained vocabulary tests and standardized vocabulary tests? Does longer treatment contribute to standardized outcomes but not trained vocabulary outcomes?
3) Is there a relation between gains in trained words and gains in standardized measures?
4) If so, is this relation stronger for expressive standardized tests than for receptive ones?

Method

Data collection: Search strategy and screening

The included studies for this review were based on a two-step process, as follows: 1) The first step refers to the strategy used in the paper by Rogde et al. (2019). In Rogde et al. (2019), a comprehensive search was conducted to assess RCTs and QEs conducted in educational contexts with the goal of improving children's language skills. This study synthesized the effect of language instruction on solely standardized language outcomes. The included studies in the current paper are based on the same search strings and terms that can be found in Appendix 1. 2) The second step of data collection refers to the inclusion of studies for the current meta-analysis reported in this paper. This involved screening for papers in Rogde et al. (2019) for further analyses. At this step, included studies had to report *both* vocabulary outcomes measured by standardized tests and trained vocabulary outcomes. Thus, the sample of studies in the current review is a subsample of the studies included in the meta-analysis by Rogde et al. (2019). In Rogde et al. (2019), trained vocabulary outcomes reported from the studies were not analyzed.

Criteria for considering studies for this review

RCTs and QEs with a pre–post controlled designs were eligible for inclusion. In addition, QEs with non-random assignment provided evidence that there were no baseline differences judged to be of substantial importance. Still, QEs represent weaker designs with more threats to the validity of causal inferences than RCTs. Imbalances between groups on variables not measured could still exist. The decision to still include QEs was made to be sure we would end up with a sufficient number of studies. The intervention programs had to be conducted in preschool or school up to the end of secondary school. Intervention programs implemented by parents or other persons in the children's home environment were excluded. The sample of participants could include typically achieving children, second-language learners, children with language weaknesses, or children from low socioeconomic backgrounds. The samples of children with special diagnoses, such as autism and other mental or sensory disabilities, were ineligible for inclusion. To be included in the current review, the studies had to report outcomes of trained vocabulary and standardized vocabulary measured at the same time point. Distinguishing between trained vocabulary and standardized vocabulary would raise the question about whether items in standardized tests could include words trained in an intervention. If this was reported in a trial, the study was excluded. Thus, for studies to be included in the review, an intervention effect on *both* the following outcome variables had to be reported: *trained vocabulary outcomes* (researcher-created tests designed to examine the knowledge of directly trained words) and *standardized vocabulary outcomes* (tests that excluded items explicitly trained in an instructional program).

Studies were excluded for the following three main reasons: The study did not report any outcome of trained vocabulary. The study only reported outcomes that included a mix of target words and untrained words. The study only reported effect sizes of trained vocabulary outcomes as 'unit tests' with different assessment time points than those of the standardized vocabulary outcomes.

Data extraction

Measures of treatment effect and training duration. The first and second authors coded the information of interest from all the studies. This included effect sizes for taught vocabulary, effect sizes for standardized vocabulary and training duration. Questions related to the coding of information were discussed within the research team. Details of outcomes and effect sizes for each study are provided in Appendix 2.

Risk of bias assessment. Risk of bias was assessed for each study, coded independently by two of the authors and decided by consensus. The studies were judged as high risk, unclear risk or low risk according to the following four type of biases: selection bias, performance bias, detection bias, attrition bias and reporting bias. This classification is recommended by Higgins et al. (2011). Details of risk of bias assessment and judgement are provided in Appendix 3. For more information about

the type of biases and a broader explanation of what the judgements were based on, see Rogde et al. (2019).

Publication bias. Publication bias occurs when a mean effect size is upwardly biased because only studies with large or significant effects are published (i.e. the file-drawer problem with entire studies) or because authors only report data on variables that show effects (often referred to as p-hacking, or the file-drawer problem for parts of studies; see Simmons et al., 2011, 2014). The studies in the current meta-analysis were included in a p-curve analysis for standardized outcomes in the systematic review by Rogde et al. (2019). Information about the p-curve analysis, transparency of data extraction and presentation of the result can be found in Rogde et al. (2019). No evidence of publication bias was detected.

Data synthesis

The data were entered into the comprehensive meta-analysis program by Borenstein et al. (2014). The effect sizes were calculated by dividing the differences in gains between pretests and posttests in the treatment group and the control group by the pooled SD for each group in the pretest, a method recommended by Morris (2008). When the effect size was positive, the group receiving vocabulary instruction made greater pretest-posttest gains than the control group. We adjusted the effect sizes for small samples using Hedges' g (Hedges & Olkin, 1985), and *d* could be converted to Hedges' g by using the correction factor J, corresponding to the following formula: $J = 1 - (3/(4\,df - 1)$ (Borenstein et al., 2014). The overall effect sizes were estimated by calculating a weighted average of individual effect sizes using a random effect model at 95% confidence intervals (CIs). Since the intervention studies were likely to differ in terms of sample characteristics, instructional features, and implementation of the programs, we selected a random effect model for estimating the effect. In the random effect model, the weighted average takes into account that the studies are associated with variations. Using this model is recommended by Borenstein et al. (2009).

Analyses of primary outcomes. To examine the difference in effect sizes between trained vocabulary and standardized vocabulary outcomes, we estimated two separate overall mean effect size – one for each outcome.

Multiple outcome reporting. When a study reported multiple indicators for the same type of outcome (e.g. multiple trained vocabulary outcomes or standardized vocabulary outcomes), the mean of the indicators was computed.

Multiple group comparisons. In one case (Silverman et al., 2013), several treatment groups compared with the same control group were reported. In this case, we computed the mean effect size from the study to avoid treating them as separate effects in the analyses.

Moderator analysis. Moderator analyses of training duration (total number of hours in the instruction programs) were conducted for the analyses of trained vocabulary and standardized vocabulary. The variable 'training duration' was originally planned to work as a continuous variable (Rogde et al., 2016); however, the variable

was not normally distributed. For the analyses, the studies were therefore divided into those that included less than 30 hours of instruction and those that reported instruction for 30 hours or more.

Results

Included studies

The screening resulted in 17 included studies that reported both outcomes of trained vocabulary and standardized vocabulary. The studies involving preschool and school-aged children were included. An overview of the study characteristics is provided in Appendix 4. In all 26 papers were excluded. This mainly entailed papers that did not report taught vocabulary outcomes. In addition, studies that made use of several unit tests of taught vocabulary during the trial were excluded from this review to ensure that the two outcomes (taught vocabulary and standardized vocabulary) were measured at the same time.

Synthesis of results

Results for research question 1. To examine the size difference between gains in trained vocabulary and in standardized vocabulary, we computed two separate overall mean effect sizes of trained vocabulary and standardized vocabulary outcomes.

Figure 1 shows the 17 effect sizes comparing treatment and control groups on trained vocabulary outcomes (*N* treatment groups = 7492, mean sample size = 441, *N* controls = 5862, mean = 345). The mean effect size was large, $g = 1.28$, 95% CI = [0.95, 1.61], $p = 0.0001$. The heterogeneity among the studies was significant,

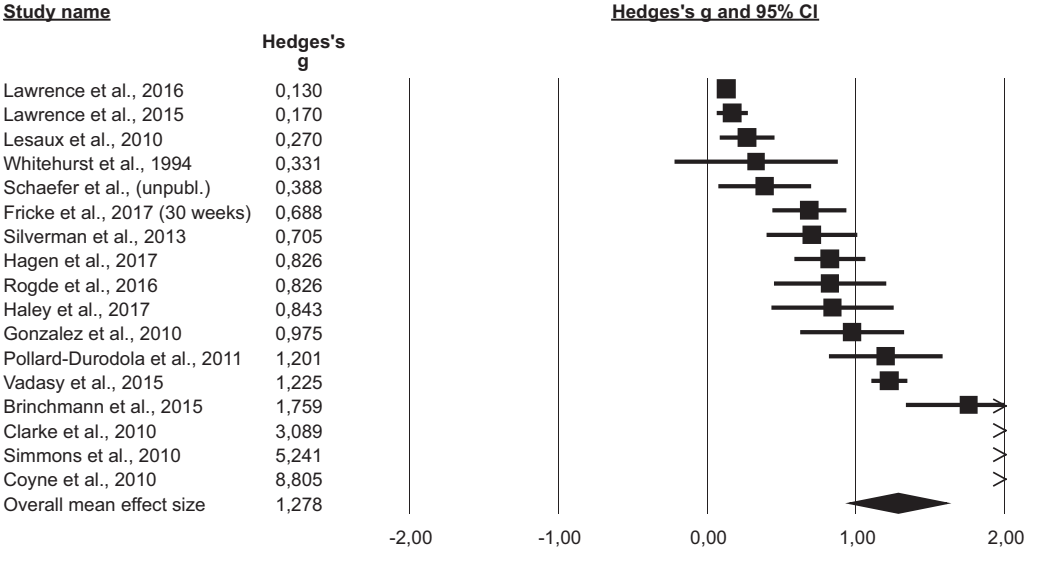

Figure 1. Effect of vocabulary instruction on taught vocabulary outcomes.

$Q(16) = 731.09$, $p = 0.0001$, $I^2 = 97.81$, $T^2 = 0.42$. After removing two outliers (Coyne et al., 2010; Simmons et al., 2010), the mean effect was $g = 0.85$, ($k = 15$), 95% CI = [0.57, 1.13], $p = 0.0001$. The heterogeneity among the studies was significant, $Q(14) = 481.04$, $p = 0.0001$, $I^2 = 97.09$, $T^2 = 0.27$. In the protocol for the review by Rogde et al. (2019), outliers larger than three SDs from the mean should be excluded. It can be noted that the study by Clarke et al. (2010) also yielded a high effect size. Still, this study was closer to the mean effect size and judged to be at low risk of several biases on the quality assessment (see Appendix 4). It was therefore kept in the analyses.

Figure 2 shows the 17 effect sizes comparing treatment and control groups on standardized vocabulary outcomes (N treatment groups = 7492, mean sample size = 440, N controls = 5862, mean = 345). The mean effect size was negligible, $g = 0.01$, 95% CI = [−0.03, 0.04], $p = 0.62$, and there was no overlap between this CI and that for trained vocabulary. The heterogeneity among the studies was not significant, $Q(16) = 14.21$, $p = 0.58$, $I^2 = 0.00$, $T^2 = 0.00$. These results indicated a mean difference of $g = 1.27$ between standardized vocabulary outcomes and trained vocabulary outcomes. Taking into account the two outliers for the trained vocabulary, the difference between the two outcomes was still large, showing $g = 0.84$.

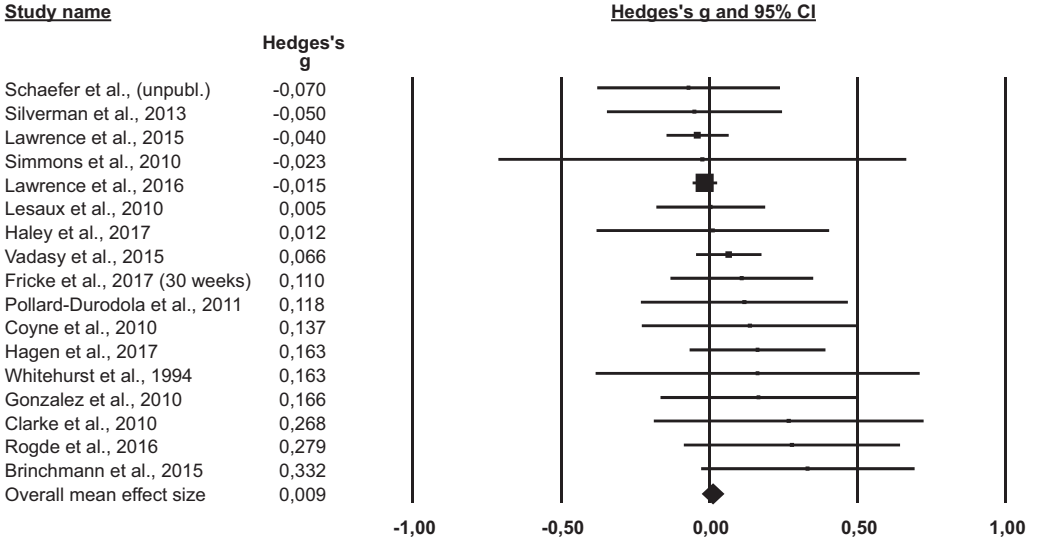

Figure 2. Effect of vocabulary instruction on standardized vocabulary outcomes.

Results for research question 2: Moderator analysis of training duration. For the trained vocabulary outcomes, the effect sizes for the treatment groups with less than 30 hours of instruction [$g = 3.17$, 95% CI = 1.28 to 5.06, $k = 5$] were significantly larger than for the groups that received 30 hours of instruction or more [$g = 0.86$, 95% CI = 0.55 to 1.17, $k = 12$]. This indicated a pattern where less instruction

time was associated with larger effect sizes. When removing the two outliers with very high effect sizes on trained vocabulary from the analysis (Coyne et al., 2010; Simmons et al., 2010), only three studies were left in the group with less than 30 hours of instruction, and there was no difference in the effect sizes related to the duration of instruction.

The overall effect on standardized vocabulary was close to zero, and the moderator analysis of training duration was not significant for this outcome.

Results for research question 3: Is there a relation between gains in trained words and in standardized measures?

We conducted a meta-regression analysis to examine whether standardized vocabulary skills could be mediated by trained vocabulary skills. The results showed that trained vocabulary was not a significant predictor of standardized vocabulary skills ($\beta = 0.04$, $R^2 = 0.00$, $k = 17$).

Results for research question 4: Is this relation stronger for expressive standardized tests than for receptive ones? Figure 3 shows the 10 reported expressive standardized tests. The results indicated a small mean effect size, $g = 0.131$, 95% CI = [0.03, 0.23], $p = 0.01$ for these studies. The heterogeneity among the studies was not significant, $Q(9) = 8.998$, $p = 0.44$, $I^2 = 0.00$, $T^2 = 0.00$. All eight studies that reported receptive standardized tests (Figure 4) showed a small mean effect size, $g = 0.138$, 95% CI = [0.03, 0.25], $p = 0.02$. The heterogeneity among the studies was not significant, $Q(7) = 3.862$, $p = 0.80$, $I^2 = 0.00$, $T^2 = 0.00$.

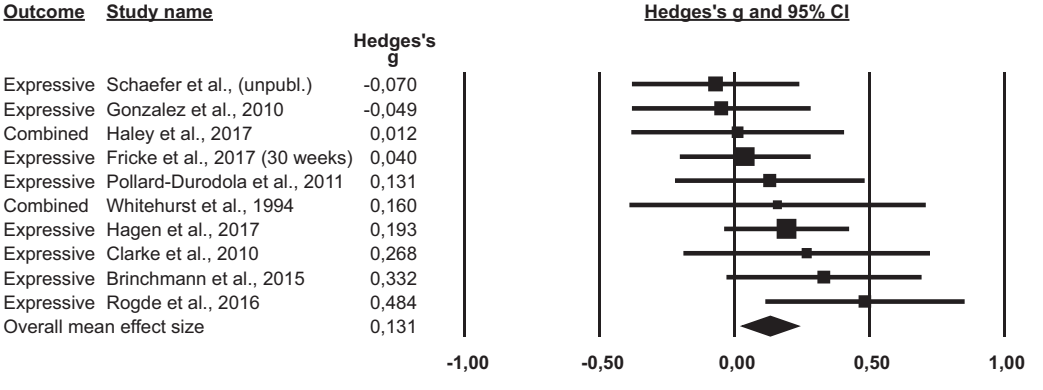

Figure 3. Effect of vocabulary instruction on standardised expressive vocabulary outcomes.

Two meta-regression analyses were conducted to examine whether standardized expressive or receptive vocabulary skills could be mediated by trained vocabulary skills. The results showed that trained vocabulary predicted neither standardized expressive vocabulary skills ($\beta = 0.11$, $R^2 = 0.00$, $k = 10$) nor standardized receptive vocabulary skills ($\beta = -0.001$, $R^2 = 0.00$, $k = 8$).

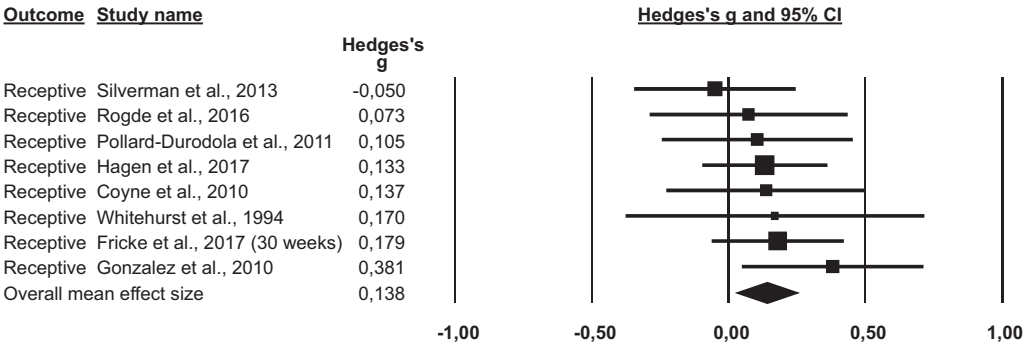

Figure 4. Effect of vocabulary instruction on standardised receptive vocabulary outcomes.

Risk of bias in the included studies

The risk of bias assessment (Appendix 3) showed that six studies were judged to be at high risk and eleven studies at low risk of selection bias. All studies represented a risk of performance bias, as blinding of personnel or participants are not possible in these type of trials. In all, seven studies reported blinding of the outcome assessment and were judged as low risk for detection bias; the remaining ten studies did not report whether or not the assessments were blinded and were categorized as being unclear. The assessment of attrition bias resulted in thirteen studies judged to be at low risk, one was judged unclear, and three were judged as high risk. None of the studies showed indications of reporting bias. No threshold was defined to exclude studies related to specific criteria for high risk of bias. This implies that all studies were included without any stratification incorporated into the conducted analyses. Thus, the risk of bias assessment was not incorporated in the mean analyses reported.

Discussion

In this paper, our main aim is to examine the size of the difference between trained vocabulary and standardized vocabulary outcomes and how the duration of the instruction is associated with the two different outcome measures. The results support the first hypothesis in which the effect of vocabulary programs on trained vocabulary outcomes shows considerably larger effect sizes than on standardized vocabulary outcomes. The mean difference between the two types of outcome tests is more than 1 SD (g = 1.27). These results are in line with Slavin and Madden's (2011) findings from other educational reviews that much larger positive effect sizes are associated with treatment-inherent measures in contrast to treatment-independent measures.

The results partly support the hypothesis that the training duration is likely to be differentially associated with trained vocabulary and standardized vocabulary outcomes. The effect sizes on trained vocabulary from treatment groups characterized by the fewest hours of training are associated with larger effects than effect sizes derived from treatment groups with more hours and more sessions of instruction. However,

when two outliers showing very high values on the trained vocabulary outcomes are excluded from the analysis, no difference in effect sizes among the studies in relation to the hours of instruction is found. Due to the small number of studies and the use of categorical moderator variable analyses, this finding is in general not straightforward in its interpretation. It is also clear that the effect on trained vocabulary outcomes may be influenced by the number of trained words. A higher number of trained words could possibly be associated with smaller effect sizes. A limited number of trained words could reflect more repetition work or more elaborate learning strategies when the words are trained, thus leading to larger effect sizes. However, several studies do not report the total amount of trained words, and we have been unable to examine this issue further. As for the standardized vocabulary outcomes, it is not possible to detect any pattern of training duration as a moderator of effect sizes because the mean overall effect size is negligible. In conclusion, we can therefore not dismiss the possibility that these types of vocabulary outcomes may be differentially moderated by the duration of the instruction program.

Our second aim is to examine whether standardized outcomes would be mediated by the effects on trained words, as well as whether expressive and receptive outcomes would show similar or different associations with trained vocabulary. Contrary to the primitive elements theory (Taatgen, 2013) and the vector semantics theory (Jurafsky & Martin, 2014), we do not find any evidence of the relations between the effects on trained vocabulary and on standardized vocabulary measures. As opposed to earlier research showing the transfer of effects between trained vocabulary and standardized expressive measures (Melby-Lervåg et al., 2020), we do not find that trained vocabulary predicts effects on either receptive or expressive standardized measures of vocabulary. Thus, our findings support the theories by Bransford and Schwartz (1999) and Detterman (1993), suggesting that the method of testing the transfer is too restrictive to detect a possible relation. Alternatively, the results are in line with the findings of Singley and Anderson (1989) and Thorndike and Woodworth (1901) that this kind of transfer is quite rare and usually mainly occurs if the tasks are highly similar.

Why then do we not find any relation here between trained vocabulary and standardized measures (or more specifically, expressive standardized measures), as in the study of Melby-Lervåg et al. (2020)? It should be noted that the studies in this review vary in the kind of intervention programs used. Some studies focus mainly on the trained words and not on broader language skills; other studies have a broader approach and a longer duration. As outlined earlier, this moderator also seems to have at least a weak relation to the size of the effects on trained vocabulary versus standardized measures. In the study by Melby-Lervåg et al. (2020), which finds a relation between the sizes of the gains in trained vocabulary and in standardized expressive tests, the broader oral training is by far the largest part of the intervention. The intervention strength is also considerable, with 5 x 6 weeks over 1.5 years. Due to the few studies, it is not possible to enter the effects of expressive language, the effects on trained words, and training duration in one regression model. However,

to examine more closely what actually improves in oral language interventions, these variables are important to consider in future studies.

The studies included in this review involved both RCTs and QEs with a pre-post controlled design. Based on the results of the quality assessment, the studies in the analyses represent studies with both low and high risk of selection bias which refers to processes of randomization and allocation. Since the analyses do not adjust for selection bias or other biases assessed, it is important to note that there are possible biases associated with the studies included in this review.

As indicated by our previous review (Rogde et al., 2019), there was no evidence of publication bias in the included studies. Despite the fact that missing studies always presents a possible source of biased conclusion in systematic reviews, the results from the p-curve analysis in Rogde et al. (2019) indicates that this is a true effect that is not limited by publication bias.

In conclusion, this study's results highlight the importance of distinguishing between the interpretations of effect sizes associated with researcher-created tests of trained vocabulary and standardized vocabulary outcomes. The results support Slavin and Madden's (2011) view that these types of outcomes should be differentiated when synthesizing effects in meta-analyses. Reviews that incorporate both types of outcomes but conduct separate analyses of them should also be precise in their interpretations of effects and their communication of evidence for practice relative to the types of outcomes to which they refer.

Limitations

The current paper is written based upon additional analyses of a prior review published in 2019. The paper is based on a search conducted in 2018, and has not been updated for the current findings.

Analyses of publication bias have not been conducted exclusively for this paper, but the studies involved are included in a previous p-curve analysis of the standardized outcomes in the paper by Rogde et al. (2019). Effect sizes on taught vocabulary outcomes are likely to be larger than standardized outcomes. Thus, it is more likely that studies reporting taught vocabulary outcomes solely (i.e. and no standardized outcomes) are more biased to publication reporting than studies reporting both taught vocabulary and standardized outcomes (which are included in this review). Hence, we argue that publication bias analysis would be most important for the standardized outcomes in the studies, and since the studies in this review were already included in the analysis in the paper by Rogde et al. (2019), additional tests were not conducted for the purpose of this paper.

Declaration of conflicting interest

The authors have papers that are included in the review themselves.

Author biographies

Kristin Rogde is a postdoc at the Department of Education at the University of Oslo. Her research have mainly focused on the effect of language comprehension instruction in first and second language learners.

Åste Mjelve Hagen is an associate professor at the Department of Special Needs Education at the University of Oslo. Her research interests relate to understanding child language and literacy development, special needs education, and developing interventions for preschools.

Arne Lervåg is a professor in the Department of Education at the University of Oslo. His research and academic interest include the development of reading and language skills, reading comprehension, word decoding, and latent variable and growth modeling.

Monica Melby-Lervåg is a professor at the Department of Special Needs Education, University of Oslo. She specializes in research on children's language and reading development and how to ameliorate difficulties in these areas, but are also interested in math and cognitive development. She has done a number of longitudinal studies and randomized controlled trials, and she is also acknowledged for her work involving meta-analyses. Melby-Lervåg started out as an educational psychologist, and a closeness to educational practice has inspired her work.

References

*Brinchmann, E. I., Hjetland, H. N., & Lyster, S. A. H. (2015). Lexical quality matters: Effects of word knowledge instruction on the language and literacy skills of third- and fourth-grade poor readers. *Reading Research Quarterly*, *51*(2), 165–180. https://doi.org/10.1002/rrq.128

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2014). *Comprehensive meta-analysis* (Version 3). [Computer software]. Guildford Press.

Bransford, J. D., & Schwartz, D. L. (1999). Chapter 3: Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, *24*(1), 61–100. https://doi.org/10.3102/0091732X024001061

*Clarke, P. J., Snowling, M. J., Truelove, E., & Hulme, C. (2010). Ameliorating children's reading-comprehension difficulties: A randomized controlled trial. *Psychological Science*, *21*(8), 1106–1116. https://doi.org/10.1177/0956797610375449

*Coyne, M. D., McCoach, D. B., Loftus, S., Zipoli Jr., R., Ruby, M., Crevecoeur, Y. C., & Kapp, S. (2010). Direct and extended vocabulary instruction in kindergarten: Investigating transfer effects. *Journal of Research on Educational Effectiveness*, *3*(2), 93–120. https://doi.org/10.1080/19345741003592410

Detterman, D. L. (1993). The case for the prosecution: Transfer as an epiphenomenon. In D. K. Detterman, & R. J. Sternberg (Eds.), *Transfer on trial: Intelligence, cognition, and instruction* (pp. 1–24). Ablex.

Elleman, A. M., Lindo, E. J., Morphy, P., & Compton, D. L. (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. *Journal of Research on Educational Effectiveness*, *2*(1), 1–44. https://doi.org/10.1080/19345740802539200

*Fricke, S., Burgoyne, K., Bowyer-Crane, C., Kyriacou, M., Zosimidou, A., Maxwell, L., … & Hulme, C. (2017). The efficacy of early language intervention in mainstream school settings: A randomized controlled trial. *Journal of Child Psychology and Psychiatry*, *58*(10), 1141–1151. https://doi.org/10.1111/jcpp.12737

*Gonzalez, J. E., Pollard-Durodola, S., Simmons, D. C., Taylor, A. B., Davis, M. J., Kim, M., & Simmons, L. (2010). Developing low-income preschoolers' social studies and science vocabulary knowledge through content-focused shared book reading. *Journal of Research on Educational Effectiveness*, *4*(1), 25–52. https://doi.org/10.1080/19345747.2010.487927

*Hagen, Å. M., Melby-Lervåg, M., Lervåg, A. (2017). Improving language comprehension in preschool children with language difficulties: A cluster randomized trial. *Journal of Child Psychology and Psychiatry*, *58*(10), 1132–1140. https://doi.org/10.1111/jcpp.12762

*Haley, A., Hulme, C., Bowyer-Crane, C., Snowling, M. J., & Fricke, S. (2017). Oral language skills intervention in pre-school—a cautionary tale. *International Journal of Language & Communication Disorders*, *52*(1), 71–79. https://doi.org/10.1111/1460-6984.12257

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Higgins, J. P. T., Altman, D. G., & Sterne, J. A. C. (Eds.) (2011). Chapter 8: Assessing risk of bias in included studies. In J. P. T. Higgins, & S. Green (Eds.). *Cochrane handbook for systematic reviews of interventions version 5.1.0* (updated March 2011). The Cochrane Collaboration, 2011 Retrieved from www.cochrane-handbook.org

Hjetland, H. N., Brinchmann, E. I., Scherer, R., & Melby-Lervåg, M. (2017). Preschool predictors of later reading comprehension ability: A systematic review. *Campbell Systematic Reviews*, *14*, 1–156. https://doi.org/10.4073/csr.2017.14

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing* (Vol. 3). Pearson.

*Lawrence, J. F., Crosson, A. C., Paré-Blagoev, E. J., & Snow, C. E. (2015). Word generation randomized trial: Discussion mediates the impact of program treatment on academic word learning. *American Educational Research Journal*, *52*(4), 750–786. https://doi.org/10.3102/0002831215579485

*Lawrence, J. F., Francis, D., Paré-Blagoev, J., & Snow, C. E. (2017). The poor get richer: Heterogeneity in the efficacy of a school-level intervention for academic language. *Journal of Research on Educational Effectiveness*, *10*, 767–793. https://doi.org/10.1080/19345747.2016.1237596

Lee, A. Y. (1998). Transfer as a measure of intellectual functioning. In S. Soraci, & W. J. Mcllvane (Eds.), *Perspectives on fundamental processes in intellectual functioning: A survey of research approaches* (Vol. 1, pp. 351–366). Ablex.

*Lesaux, N. K., Kieffer, M. J., Faller, S. E., & Kelley, J. G. (2010). The effectiveness and ease of implementation of an academic vocabulary intervention for linguistically diverse students in urban middle schools. *Reading Research Quarterly*, *45*(2), 196–228. https://doi.org/10.1598/RRQ.45.2.3

Marulis, L. M., & Neuman, S. B. (2010). The effects of vocabulary intervention on young children's word learning: A meta-analysis. *Review of Educational Research*, *80*(3), 300–335. https://doi.org/10.3102/0034654310377087

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*(3), 223–262. https://doi.org/10.1080/19345747.2012.755591

Melby-Lervåg, M., Hagen, Å. M., Lervåg, L. (2020). Disentangling the far transfer of language comprehension gains using latent mediation models. *Developmental Science*, *23*(4), 1–14. https://doi.org/10.1111/desc.12929

Mol, S. E., Bus, A. G., & de Jong, M. T. (2009). Interactive book reading in early education: A tool to stimulate print knowledge as well as oral language. *Review of Educational Research*, *79*(2), 979–1007. https://doi.org/10.3102/0034654309332561

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, *11*, 364–386. https://doi.org/10.1177/1094428106291059

*Pollard-Durodola, S. D., Gonzalez, J. E., Simmons, D. C., Kwok, O., Taylor, A. B., Davis, M. J., … & Simmons, L. (2011). The effects of an intensive shared book-reading intervention for preschool children at risk for vocabulary delay. *Exceptional Children*, *77*(2), 161–183. https://doi.org/10.1177/001440291107700202

Rogde. K., Hagen, Å. M., Melby-Lervåg, M., Lervåg, A., (2016). Protocol: The effect of linguistic comprehension training on language and reading comprehension: A systematic review. *Campbell Systematic Reviews*. https://doi.org/10.1002/cl2.161

Rogde. K., Hagen, Å. M., Melby-Lervåg, M., Lervåg, A. (2019). Review: The effect of linguistic comprehension instruction on generalized language and reading comprehension skills: A systematic review. *Campbell Systematic Reviews*. https://doi.org/10.1002/cl2.1059

*Rogde. K., Melby-Lervåg, M., Lervåg, A. (2016). Improving the general language skills of second-language learners in kindergarten: A randomized controlled trial. *Journal of Research on Educational Effectiveness*, *9*(Sup1.), 150–170. https://doi.org/10.1080/19345747.2016.1171935

*Schaefer, B., Fricke, S., Bowyer-Crane, C., Millard. G., & Hulme, C. *Effects of an oral language intervention programme for children with English as an additional language and monolingual peers with language weaknesses: A randomised controlled trial* [Unpublished manuscript].

*Silverman, R., Crandell, J. D., & Carlis, L. (2013). Read alouds and beyond: The effects of read aloud extension activities on vocabulary in Head Start classrooms. *Early Education & Development*, *24*(2), 98–122. https://doi.org/10.1080/10409289.2011.649679

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill* (No. 9). Cambridge, MA: Harvard University Press.

*Simmons, D., Hairrell, A., Edmonds, M., Vaughn, S., Larsen, R., Willson, V., … & Byrns, G. (2010). A comparison of multiple-strategy methods: Effects on fourth-grade students' general and content-specific reading comprehension and vocabulary development. *Journal of Research on Educational Effectiveness*, *3*(2), 121–156. https://doi.org/10.1080/19345741003596890

Slavin, R. E., Lake, C., Davis, S., & Madden, N. A. (2011). Effective programs for struggling readers: A best-evidence synthesis. *Educational Research Review*, *6*(1), 1–26. https://doi.org/10.1016/j.edurev.2010.07.002

Slavin, R., & Madden, N. A. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness*, *4*(4), 370–380. https://doi.org/10.1080/19345747.2011.558986

Swanson, E., Vaughn, S., Wanzek, J., Petscher, Y., Heckert, J., Cavanaugh, C., … & Tackett, K. (2011). A synthesis of read-aloud interventions on early reading outcomes among preschool through third graders at risk for reading difficulties. *Journal of Learning Disabilities*, *44*(3), 258–275. https://doi.org/10.1177/0022219410378444

Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychological Review*, *120*(3), 439. https://doi.org/10.1037/a0033138

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. II. The estimation of magnitudes. *Psychological Review*, *8*(4), 384.

*Vadasy, P. F., Sanders, E. A., & Logan Herrera, B. (2015). Efficacy of rich vocabulary instruction in fourth- and fifth-grade classrooms. *Journal of Research on Educational Effectiveness*, *8*(3), 325–365. https://doi.org/10.1080/19345747.2014.933495

What Works Clearinghouse. (2008a). *Beginning reading. What Works Clearinghouse Topic Report*. Institute of Education Sciences. http://www.ies.ed.gov/ncee/wwc

What Works Clearinghouse. (2008b). *Elementary school mathematics. What Works Clearinghouse Topic Report*. Institute of Education Sciences. http://www.ies.ed.gov/ncee/wwc

*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*(5), 679–85. https://doi.org/10.1037/0012-1649.30.5.679

*Studies included in the meta-analysis

References to excluded studies

Apthorp, H. S. (2006). Effects of a supplemental vocabulary program in third-grade reading/language arts. *Journal of Educational Research, 100*(2), 67–79. https://doi.org/10.3200/JOER.100.2.67-79

Apthorp, H., Randel, B., Cherasaro, T., Clark, T., McKeown, M., & Beck, I. (2012). Effects of a supplemental vocabulary program on word knowledge and passage comprehension. *Journal of Research on Educational Effectiveness, 5*(2), 160–188. https://doi.org/10.1080/19345747.2012.660240

Block, C. C., & Mangieri, J. N. (2006*). The effects of powerful vocabulary for reading success on students' reading vocabulary and comprehension achievement* (Research Report 2963–005, Institute for Literacy Enhancement). Retrieved from http://teacher.scholastic.com/products/fluencyformula/pdfs/powerfulvocab_Efficacy.pdf

Cable, A. L. (2007). *An oral narrative intervention for second graders with poor oral narrative ability*. ProQuest. Retrieved from http://hdl.handle.net/2152/3550

Crain-Thoreson, C., & Dale, P. S. (1999). Enhancing linguistic performance: Parents and teachers as book reading partners for children with language delays. *Topics in Early Childhood Special Education, 19*(1), 28–39. https://doi.org/10.1177/027112149901900103

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*(4), 497–515. https://doi.org/10.1348/000709910X493080

Farver, J. A. M., Lonigan, C. J., & Eppe, S. (2009). Effective early literacy skill development for young Spanish-speaking English language learners: An experimental study of two methods. *Child Development, 80*(3), 703–719. https://doi.org/10.1111/j.1467-8624.2009.01292.x

Fricke, S., Bowyer-Crane, C., Haley, A. J., Hulme, C., & Snowling, M. J. (2013). Efficacy of language intervention in the early years. *Journal of Child Psychology and Psychiatry, 54*(3), 280–290. https://doi.org/10.1111/jcpp.12010

Johanson, M., & Arthur, A. M. (2016). Improving the language skills of pre-kindergarten students: Preliminary impacts of the let's know! Experimental curriculum. *Child &Youth Care Forum, 45*(3) 367–392. https://doi.org/10.1007/s10566-015-9332-z

Justice, L. M., Mashburn, A., Pence, K. L., & Wiggins, A. (2008). Experimental evaluation of a preschool language curriculum: Influence on children's expressive language skills. *Journal of Speech, Language, and Hearing Research, 51*(4), 983–1001. https://doi.org/10.1044/1092-4388(2008/072)

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*(2), 161–178. https://doi.org/10.1044/0161-1461(2009/08-0058)

Kelley, E. S., Goldstein, H., Spencer, T. D., & Sherman, A. (2015). Effects of automated Tier 2 storybook intervention on vocabulary and comprehension learning in preschool children with limited oral language skills. *Early Childhood Research Quarterly, 31*, 47–61. https://doi.org/10.1016/j.ecresq.2014.12.004

Lesaux, N. K., Kieffer, M. J., Kelley, J. G., & Harris, J. R. (2014). Effects of academic vocabulary instruction for linguistically diverse adolescents: Evidence from a randomized field trial. *American Educational Research Journal, 51*(6), 1159–1194. https://doi.org/10.3102/0002831214532165

Lonigan, C. J., & Whitehurst, G. J. (1998). Relative efficacy of parent and teacher involvement in a shared-reading intervention for preschool children from low-income backgrounds. *Early Childhood Research Quarterly, 13*(2), 263–290. https://doi.org/10.1016/S0885-2006(99)80038-6

Lonigan, C. J., Anthony, J. L., Bloomfield, B. G., Dyer, S. M., & Samwel, C. S. (1999). Effects of two shared-reading interventions on emergent literacy skills of at-risk preschoolers. *Journal of Early Intervention, 22*(4), 306–322. https://doi.org/10.1177/105381519902200406

Lonigan, C. J., Purpura, D. J., Wilson, S. B., Walker, P. M., & Clancy-Menchetti, J. (2013). Evaluating the components of an emergent literacy intervention for preschool children at risk for reading difficulties. *Journal of Experimental Child Psychology, 114*(1), 111–130. https://doi.org/10.1016/j.jecp.2012.08.010

Murphy, A., Franklin, S., Breen, A., Hanlon, M., McNamara, A., Bogue, A., & James, E. (2016). A whole class teaching approach to improve the vocabulary skills of adolescents attending mainstream secondary school, in areas of socioeconomic disadvantage. *Child Language Teaching and Therapy*, 1–16. https://doi.org/10.1177/0265659016656906

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*(3), 249–272. https://doi.org/10.1598/RRQ.46.3.3

Nielsen, D. C., & Friesen, L. D. (2012). A study of the effectiveness of a small-group intervention on the vocabulary and narrative development of at-risk kindergarten children. *Reading Psychology, 33*(3), 269–299. https://doi.org/10.1080/02702711.2010.508671

Phillips, B. M., Tabulda, G., Ingrole, S. A., Burris, P. W., Sedgwick, T. K., & Chen, S. (2016). Literate language intervention with high-need prekindergarten children: A randomized trial. *Journal of Speech, Language, and Hearing Research, 59*(6), 1409–1420. https://doi.org/10.1044/2016_JSLHR-L-15-0155, https://doi.org/10.1080/19345741003596890

Proctor, C. P., Dalton, B., Uccelli, P., Biancarosa, G., Mo, E., Snow, C., & Neugebauer, S. (2011). Improving comprehension online: Effects of deep vocabulary instruction with bilingual and monolingual fifth graders. *Reading and Writing, 24*(5), 517–544. https://doi.org/10.1007/s11145-009-9218-2

Spencer, T. D., Petersen, D. B., Slocum, T. A., & Allen, M. M. (2014). Large group narrative intervention in Head Start preschools: Implications for response to intervention. *Journal of Early Childhood Research, 13*(2), 196–217. https://doi.org/10.1177/1476718X13515419

Styles, B., & Bradshaw, S. (2015). *Talk for literacy: Evaluation report and executive summary*. National Foundation for Educational Research.

Valdez-Menchaca, M. C., & Whitehurst, G. J. (1992). Accelerating language development through picture book reading: A systematic extension to Mexican day care. *Developmental Psychology, 28*(6), 1106. https://doi.org/10.1037/0012-1649.28.6.1106

van Kleeck, A., Vander Woude, J., & Hammett, L. (2006). Fostering literal and inferential language skills in Head Start preschoolers with language impairment using scripted book-sharing discussions. *American Journal of Speech-Language Pathology, 15*(1), 85–95. https://doi.org/10.1044/1058-0360(2006/009)

Wasik, B. A., & Bond, M. A. (2001). Beyond the pages of a book: Interactive book reading and language development in preschool classrooms. *Journal of Educational Psychology*, *93*(2), 243–250. https://doi.org/10.1037/0022-0663.93.2.243