Minnesota Reading Corps
Minnesota Assessment of Vocabulary for Reading Improvement and Comprehension (MAVRIC)
Social Innovation Fund

Year 4 (Final) Impact and Implementation Evaluation Report
2016-2017

Version Date: January 2nd, 2018

David Parker, PhD
Vice President—Research and Development
ServeMinnesota
david@serveminnesota.org

Patrick Kaiser
Senior Evaluation Manager
ServeMinnesota
patrick@serveminnesota.org

Evaluation supervision and support provided by the Center for Applied Research and Educational Improvement (CAREI) at the University of Minnesota.

Minnesota Assessment of Vocabulary for Reading Improvement and Comprehension (MAVRIC)

Year 4 (Final) Evaluation Report
2016-2017

David Parker
Patrick Kaiser

ServeMinnesota | Reading Corps

Center for Applied Research and Educational Improvement (CAREI)

University of Minnesota

This material is based upon work supported by the Corporation for National and Community Service (CNCS) under Grant No. 12SIHMN001, and Generation Next and Greater Twin Cities United Way. Opinions or points of view expressed in this document are those of the authors and do not necessarily reflect the official position of, or a position that is endorsed by, the Corporation, Generation Next, Greater Twin Cities United Way or the Social Innovation Fund.

Contents

EXECUTIVE SUMMARY

About MAVRIC

The Minnesota Assessment of Vocabulary for Reading Improvement and Comprehension (MAVRIC) program is a project of ServeMinnesota that is delivered through existing Reading Corps infrastructure and focuses on improving student vocabulary outcomes.   Vocabulary is a direct proxy for students' understanding of concepts and ideas in their environments (Stahl & Nagy, 2006), and as such, is a significant predictor of reading comprehension throughout school (Cunningham & Stanovich, 1997). Reading Corps trains and deploys AmeriCorps members to deliver reading interventions during the school day for students in prekindergarten through grade 3, and prior evaluations found the Reading Corps positively impacts foundational reading skills (Markovitz, Hernandez, Hedberg, & Silberglitt, 2014; 2015), including the promotion of stronger vocabulary outcomes in prekindergarten (Markovitz et al., 2015). However, the vocabulary outcomes of *at-risk* prekindergarten students, as well as the potential for Reading Corps to positively impact vocabulary skills in kindergarten and first grade are largely unknown. MAVRIC is intended to fill these gaps.

The theory of change for MAVRIC is premised largely on two functional elements: data-driven decision-making for vocabulary and optimized vocabulary interventions. For MAVRIC, data-driven decision-making has meant establishing procedures for using defensible data to identity struggling students and monitor their progress while receiving additional support (Hamilton et al., 2009). Optimizing vocabulary interventions has included creating intervention protocols for grade levels that previously lacked them (e.g., first grade), enhancing materials for standardization and quality content (e.g., improving vocabulary cards and definitions), and ensuring all materials met expectations for cultural inclusion and relevance—all of which must be consistent with evidence-based interventions for vocabulary (Beck & McKeown, 2010).

About this Report

This evaluation report is a Final Report of the MAVRIC project (spanning activities from 2016-2017), which is intended to fulfill the Social Innovation Fund requirements to determine at least a moderate level of evidence for funded projects. It follows three years of concentrated scaling efforts as well as one previous impact evaluation designed to identify a moderate level of evidence. This report also includes information on Year 4 implementation and pre-experimental findings, and thus also serves as an updated summary of the MAVRIC project evidence and learning for the ServeMinnesota intermediary, The Greater Twin Cities United Way.

About the Social Innovation Fund

The Social Innovation Fund (SIF) was a program that received funding from 2010 to 2016 from the Corporation for National and Community Service, a federal agency that engages millions of Americans in service through its AmeriCorps, Senior Corps, and Volunteer Generation Fund programs, and leads the nation's volunteer and service efforts. Using public and private resources to find and grow community-based nonprofits with evidence of results, SIF intermediaries received funding to award subgrants that focus on overcoming challenges in economic opportunity, healthy futures, and youth development. Although CNCS made its last SIF intermediary awards in fiscal year 2016, SIF intermediaries will continue to administer their subgrant programs until their federal funding is exhausted.

Prior Research and Targeted Evidence

The development of MAVRIC intervention components was driven by extant research showing convincing, positive effects for vocabulary interventions—known as Repeated Read Alouds—on vocabulary outcomes (Marulis & Neuman, 2010). During three previous years of MAVRIC implementation, annual evaluation activities indicated (a) descriptive results were promising (i.e., there was preliminary evidence) that student participants improved vocabulary skills, and (b) tutors could accurately implement the content (i.e., interventions and assessments) that was developed for the program. Of particular note, an initial impact evaluation in Year 2 targeted a moderate level of evidence using a regression discontinuity design, but found equivocal results due in part to unexpected high proportions of eligible students relative to intervention resources. Although the Year 2 evaluation did not meet the targeted evidence level, the observation about each school's pool of eligible students established the feasibility of conducting a randomized controlled trial (RCT) in the current evaluation.

This report describes impact (confirmatory and exploratory) and implementation evaluation activities that occurred during Year 4 of MAVRIC implementation, which spanned the 2016-2017 school year. Within these activities, an RCT evaluation design was conducted to inform the degree to which MAVRIC produces a moderate level of evidence as per SIF requirements for Final evaluation reports.

Evaluation Overview

The confirmatory impact evaluation in Year 4 leveraged several standard MAVRIC practices that made it feasible to conduct the RCT. Specifically, prior to receiving MAVRIC, all potential students are assessed using grade-specific vocabulary measures that have defensible technical characteristics. For prekindergarten and kindergarten the measure used was the Individual Growth and Development Indicators 2.0 (Wackerle Hollman & Bradfield, 2010), and in first grade the measure used was the 4,000 Word Listening Test (Graves & Sales, 2009). It is also standard practice in MAVRIC to use cut scores on these measures to determine whether students are eligible (or not eligible). In addition, given the observed numbers of eligible students in prior years, there were sufficient students to ethically implement a randomization procedure for determining treatment and control groups using each school's pool of eligible students. These factors permitted student level randomization into treatment and control groups based on pretest scores. Posttest scores were collected during winter for both groups, but the requirement to maintain the randomized groups was withdrawn after winter posttest to permit schools greater flexibility for which students participated in MAVRIC.

Various analytic approaches were used that corresponded to the evaluation components being addressed. For instance, all exploratory impact and implementation evaluation components were answered using descriptive approaches that provided mean and standard deviation values. The confirmatory impact evaluation followed a pre-determined model-fitting procedure for single- and multi-level regression that identified a parsimonious single-level model that included only pre-test scores as a covariate. The resulting model was applied to a final analytic sample representing 10% attrition. Attrition was not found to be related to any substantive variables (e.g., differential across groups).

Research Questions

Primary impact evaluation activities were organized by the following confirmatory impact research question:

> What is the impact on vocabulary improvement for at-risk prekindergarten, kindergarten, and 1st grade students who participate in MAVRIC compared to similar students who do not receive MAVRIC interventions?

In addition to the primary impact evaluation using an RCT, exploratory impact and implementation evaluation components were included in Year 4. The following research questions organized all additional evaluation activities.

Regarding exploratory impact:

1.) How much improvement in vocabulary skills is observed for prekindergarten, kindergarten, and 1st grade students who participate in MAVRIC?
2.) What number and percentage of prekindergarten, kindergarten, and 1st grade students who participate in MAVRIC are no longer considered at-risk for poor vocabulary outcomes after participation in MAVRIC?
3.) What is the performance for participating prekindergarten, kindergarten, and 1st grade students on vocabulary progress assessments?

Regarding implementation:

1.) To what degree are MAVRIC interventions for prekindergarten, kindergarten, and 1st grade students implemented as intended?
2.) To what degree are MAVRIC assessments of prekindergarten, kindergarten, and 1st grade students administered as intended?
3.) What is the overall inter-scorer reliability for a randomly-selected subsample of MAVRIC assessments administered to prekindergarten, kindergarten, and 1st grade student?
4.) What is the average, standard deviation, and range for number of minutes of MAVRIC intervention received each week for students in prekindergarten, kindergarten, and 1st grade?

Findings and Next Steps

In total, 886 students from 58 schools participated in the MAVRIC program during Year 4, of which 605 (including 282 non-treated control) students from 25 schools participated in the confirmatory impact evaluation that employed an RCT design. Exploratory impact evaluation results indicated all students, including treatment and control, made comparable gains across pretest and posttest data collection periods. These results were consistent with the confirmatory impact evaluation results showing no statistically significant impact of MAVRIC on student vocabulary skills. Consistent with prior evaluation findings, implementation evaluation results showed all aspects of implementation (with the exception of prekindergarten intervention delivery) were implemented with high fidelity (>90% accuracy of key steps).

Although the confirmatory impact evaluation results were unexpected, several factors—including select methodological limitations—were potentially related to the nonsignificant findings. The evaluation team identified three primary considerations: (1) the potential for different vocabulary assessments to better measure differential growth between MAVRIC participants and non-participants; (2) the possible benefit of providing additional time in intervention for MAVRIC participants; and (3) an opportunity to address implementation factors (e.g., student group size; tutor training; material refinements) that are plausibly related to vocabulary outcomes. Investigating and addressing these changes will be the focus of development and evaluation work in Year 5 (2017-18 school year) of the MAVRIC project.

Minnesota Assessment of Vocabulary for Reading Improvement and Comprehension (MAVRIC)

Year 4 (Final) Evaluation Report
2016-2017

This material is based upon work supported by the Corporation for National and Community Service (CNCS) under Grant No. 12SIHMN001, and Generation Next and Greater Twin Cities United Way. Opinions or points of view expressed in this document are those of the authors and do not necessarily reflect the official position of, or a position that is endorsed by, the Corporation, Generation Next, Greater Twin Cities United Way or the Social Innovation Fund.

INTRODUCTION

Background of Reading Corps

Reading Corps is an AmeriCorps program that provides schools literacy tutors to support reading development for students in prekindergarten through grade 3. ServeMinnesota, Minnesota's AmeriCorps Commission, oversees the Reading Corps program, which involves training tutors, providing coaching and support to tutors, and conducting ongoing development and evaluation of the program content. The central goal for Reading Corps is to successfully implement evidence-based literacy instruction and assessment protocols within school settings. The Minnesota Assessment of Vocabulary for Reading Improvement and Comprehension (MAVRIC) program was developed to be delivered through existing Reading Corps infrastructure to help schools ensure students' vocabulary skills are successful. In this Introduction section, we describe the existing Reading Corps program, the MAVRIC program, and the research questions driving the current evaluation.

Effective Service Delivery

The Reading Corps model aligns with Response-to-Intervention (RTI) and Multi-Tier Systems of Support (MTSS), which are two descriptions of a framework for delivering educational services effectively and efficiently (Burns, Deno, & Jimerson, 2007). The key aspects of alignment include the following:

- Clear literacy targets at each level from prekindergarten through grade 3
- Benchmark assessment three times a year to identify students eligible for individualized interventions
- Evidence-based literacy interventions
- Frequent progress monitoring during intervention delivery
- High quality training in program procedures, multi-level coaching, and observations to support fidelity of implementation

In the RTI and MTSS framework, data play the key roles of screening student eligibility for additional services and monitoring student progress towards achieving academic goals (i.e., benchmarks). Reading Corps screens students for program eligibility three times a year (i.e., fall, winter, spring) using empirically-derived grade- and content-specific performance benchmarks. Eligible students (defined as students scoring below target scores) are considered potential candidates to receive supplemental Reading Corps support, often referred to as additional "Tiers" of intervention within an RTI/MTSS framework.

Reading Corps is focused on intervention in the “Big Five Ideas in Literacy” as identified by the National Reading Panel, including phonological awareness, phonics, fluency, vocabulary, and comprehension. Full-time tutors at the prekindergarten level work within classrooms implementing literacy-rich practices for all students and more intensive interventions for students who need them. Tutors in kindergarten through grade 3 work with approximately 15-18 at-risk students for 20 minutes each day. The tutoring interventions align with primary literacy targets for prekindergarten and elementary-aged students (Shanahan et al., 2008; Snow, Burns, & Griffin, 1998) and are supplemental to the core reading instruction provided at each school. The goal of tutoring is to raise individual students’ literacy levels so that they are on track to meet or exceed the next program-specified literacy benchmark.

Coaching and Support

Reading Corps provides multiple layers of supervision to ensure tutors maintain the integrity of implementation. Site-specific Internal Coaches, who are typically staff literacy specialists, teachers, or curriculum directors, serve as immediate on-site supervisors, mentors, and advocates for tutors. The Internal Coach’s role is to monitor tutors and provide guidance in the implementation of Reading Corps’ assessments, and interventions, as well as the literacy rich schedule (in prekindergarten only). As the front-line supervisor, the Internal Coach is a critical component of the supervisory structure. The external, or Master Coach, is a literacy expert who provides site staff (i.e., Internal Coaches and AmeriCorps tutors) with expert consultation on literacy instruction and ensures implementation integrity of Reading Corps program elements. In addition to these two coaching layers, a third layer consisting of AmeriCorps program support helps ensure a successful year of AmeriCorps service. Program support staff consists of Reading Corps employees who provide administrative oversight for program implementation to sites participating in Reading Corps.

Training

Prior to the start of each school year, Reading Corps hosts a three-day Summer Institute to train returning and new Master Coaches, Internal Coaches, and AmeriCorps tutors. This intensive experience provides tutors and coaches the needed foundational training in the research-based literacy assessments and interventions employed by Reading Corps. During Summer Institute, tutors learn the skills, knowledge, and tools needed to serve as literacy interventionists. Tutors are provided with detailed literacy manuals as well as online resources that mirror and supplement the contents of the manual (e.g., videos of model interventions and best practices). Both the manuals and online resources are intended to provide tutors with timely support and opportunities for continued professional development and skill refinement. Additional training and coaching sessions are provided throughout the tutors’ year of service.

Background of MAVRIC Intervention

The Minnesota Assessment of Vocabulary for Reading Improvement and Comprehension (MAVRIC) is designed to leverage the Reading Corps model to improve vocabulary skills of young students from prekindergarten through first grade who are at risk of poor reading outcomes. It was developed to function within the broader Reading Corps program. This project was supported by the Social Innovation Fund and the Greater Twin Cities United Way. The SIF program received funding from 2010 to 2016 from the Corporation for National and Community Service to find and grow community-based nonprofits with evidence of results. SIF intermediaries such as the Greater Twin Cities United Way received funding to award subgrants that focus on overcoming challenges in economic opportunity, healthy futures, and youth development. For MAVRIC, this support provided resources to improve, scale, and evaluate student vocabulary outcomes in order to determine the future role and effectiveness of MAVRIC within the broader Reading Corps program. In Year 4, 886 students from 58 schools participated in the MAVRIC project.

MAVRIC Vocabulary Innovation Overview

Seminal research shows that students enter school with differences in their vocabulary skills (Hart & Risley, 1995), which is problematic because vocabulary represents and facilitates students' understanding of concepts and ideas in their environments (Stahl & Nagy, 2006). Accordingly, vocabulary is a well-established and significant predictor of reading comprehension throughout school (Cunningham & Stanovich, 1997).

In order to improve vocabulary outcomes for young students, MAVRIC uses two research-based concepts that are known to improve educational outcomes. These two concepts are the functional core of the MAVRIC Theory of Change (see Figure 1). First, data-driven decision-making (Hamilton et al., 2009) contributes to accurate, efficient identification of students who need additional support. Data are also used to monitor the effectiveness of intervention for individual students, so that instruction can be maximally responsive to students' needs (Fuchs & Deno, 1991). In vocabulary, these uses of data tend to be less well-developed as compared to other aspects needed for proficient reading (Snow et al. 2000). MAVRIC seeks to contribute to advancing this understanding.

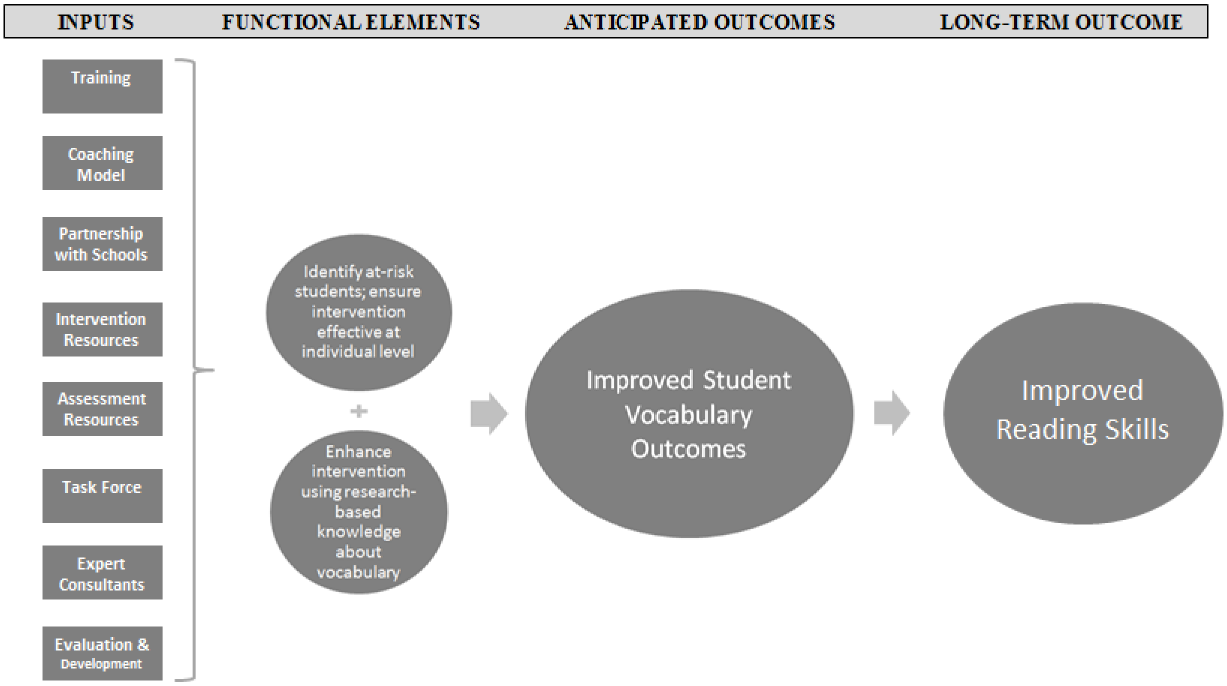


Figure 1. Illustration of MAVRIC Theory of Change

The second concept is empirically-validated intervention strategies (Beck & McKeown, 2007). The vocabulary interventions used in MAVRIC are all age-appropriate variants of the "Repeated Read Aloud" (RRA), which has been supported by meta-analytic research that showed positive outcomes across multiple potential moderating factors (Marulis & Neuman, 2010). In MAVRIC, RRA intervention procedures are enhanced by a focus on teaching high-frequency vocabulary words that are known to be important for young readers (Graves & Sales, 2009). All MAVRIC vocabulary words are selected from high-frequency vocabulary word lists, and instruction of the words is strategically sequenced across the three age groups (i.e., with simpler, more frequent words being targeted with younger students).

The MAVRIC RRAs are delivered daily in approximately 15-20 minute sessions (depending on grade level), by one adult tutor, to groups of four students in a quiet space within the school. Each tutor is asked to provide MAVRIC interventions to 12 students in groups of four in three 20 min sessions scheduled daily for at least one school semester, except where smaller numbers of students were permitted due to small school enrollment (e.g., most prekindergarten settings). The tutors use a structured script to ensure key steps are delivered correctly and consistently, (e.g., explicit definitions of words; deep processing activities). Vocabulary words taught in MAVRIC are identified as high-frequency content words on research-based word lists (e.g., Graves & Sales, 2009), meaning they make up a high percentage of content words found in prekindergarten and elementary texts.

The enrollment/tutoring process follows these steps: (1) Students are assessed for eligibility using screening vocabulary assessments, which each have target scores that correspond to at-risk (or not-at-risk) levels of performance. (2) Coaches and teachers examine the list of eligible students to determine which students will receive MAVRIC interventions (in the Year 4 Impact Evaluation, this step was replaced with a randomization procedure). (3) Students begin receiving tutoring using MAVRIC interventions. While selected students receive MAVRIC interventions, students not receiving interventions engage in independent activities or other teacher-directed tasks that are *not* directly part of core curricular literacy instruction. See Appendix A for more detail on all components of the MAVRIC Logic Model.

Evidence to Support the Innovation

Prior to Year 1 of the MAVRIC project, the RRAs that would be refined and enhanced during the project were tested in a small-scale pilot using random assignment of students to either receive the vocabulary intervention or no intervention. Results showed that students receiving the vocabulary intervention demonstrated significant gains on vocabulary skills compared with students who did not receive the intervention, and data from subsequent pilots continue to indicate students make gains in the intervention. These results, although not causal, are consistent with meta-analytic research showing vocabulary interventions using similar procedures to those employed by MAVRIC produce at least moderate effect sizes (g = 0.88; Marulis & Neuman, 2010).

In Year 1 (2013-2014) of the MAVRIC project, descriptive results from four participating schools produced preliminary evidence that students who received the intervention made greater gains than students who did not. Similar descriptive results were observed in Years 2 and 3 (2014-2015 and 2015-2016) of the project, but a more rigorous impact evaluation in Year 2 that used a regression discontinuity design did not identify a significant effect. In that impact evaluation, no effect—positive, negative, or otherwise—could be determined due to a methodological issue related to bias in student selection among eligible students, which in turn occurred due to unanticipated high numbers of eligible students. Approximately twice the number of students were eligible compared to the number of students who could be provided MAVRIC given the available tutoring resources. This finding ultimately established the feasibility for the impact evaluation in the current year (Year 4; 2016-2017) to conduct an evaluation using randomization.

Research Questions

With respect to the Year 4 impact evaluation, confirmatory and exploratory approaches were used, with the primary focus on using randomization as part of a design to produce a moderate level of evidence regarding the degree to which MAVRIC positively impacts student vocabulary outcomes. The following <u>confirmatory</u> impact question was addressed:

1.) What is the impact on vocabulary improvement for at-risk prekindergarten, kindergarten, and 1st grade students who participate in MAVRIC compared to similar students who do not receive MAVRIC interventions?

The following exploratory impact questions were also asked:

1.) How much improvement in vocabulary skills is observed for prekindergarten, kindergarten, and 1st grade students who participate in MAVRIC?
2.) What number and percentage of prekindergarten, kindergarten, and 1st grade students who participate in MAVRIC are no longer considered at-risk for poor vocabulary outcomes after participation in MAVRIC?
3.) What is the performance for participating prekindergarten, kindergarten, and 1st grade students on vocabulary progress assessments?

In addition, the following implementation questions were addressed:

1.) To what degree are MAVRIC interventions for prekindergarten, kindergarten, and 1st grade students implemented as intended?
2.) To what degree are MAVRIC assessments of prekindergarten, kindergarten, and 1st grade students administered as intended?
3.) What is the overall inter-scorer reliability for a randomly-selected subsample of MAVRIC assessments administered to prekindergarten, kindergarten, and 1st grade student?
4.) What is the average, standard deviation, and range for number of minutes of MAVRIC intervention received each week for students in prekindergarten, kindergarten, and 1st grade?

YEAR 4 EVALUATION COMPONENTS

Year 4 Evaluation Design

The Year 4 confirmatory impact evaluation design was experimental in that results were obtained from groups that were formed using randomization. This design was intended to achieve a moderate level of evidence according to the Social Innovation Fund Evaluation Plan Guidance recommendations. In this design internal validity is considered strong due to the randomization procedure distributing observed and unobserved variables related to vocabulary performance evenly between treatment (MAVRIC program recipients) and control groups. Given the evaluation created groups based on a pretest measure of vocabulary, any group differences on the posttest measure of vocabulary should reflect high internal validity with respect to the impact of the MAVRIC program. Although the design used a relatively large sample size (n = 886), external validity was not considered high because the evaluation occurred in a single school district in an urban setting; thus, the degree to which results generalized to other populations and settings is mostly unknown.

The randomization process was conducted at the student level within each participating school, separate for each grade. Only students with eligible scores below pre-established criteria on measures of the construct of interest (i.e., vocabulary assessments) were included in the randomization pool. Randomization was conducted by researchers unfamiliar with the students or schools. Specifically, school staff first identified the eligible pool of candidates based on pretest scores, and then sent the list of eligible candidates to the research team who used a simple spreadsheet program and randomization function to identify (1) students to receive treatment (i.e., the MAVRIC program), (2) students in the control group, and (3) a small group of students to serve as “backup” students in case treatment students left school for unexpected reasons. The third group was necessary to ensure the impact evaluation did not negatively impact the number of students served at the school, a major concern for participating schools, but the third group was not included in any analyses. The probability of assignment to these respective groups was approximately 55:40, with a small remaining proportion assigned as “backup”. No blocking, matching, or stratification procedures were used.

Criteria for inclusion in the randomization pool were established using pretest measures. Prekindergarten and kindergarten students were assessed at pretest using the Individual Growth and Development Indicators (IGDIs) Picture Naming (screening version). Scores of 5 and below indicated eligibility for assignment to treatment in prekindergarten and scores of 10 and below indicated eligibility for assignment in kindergarten. In first grade, the 4,000 Word Listening Test was used and scores 25 and below indicated eligibility for assignment to treatment. In prekindergarten and kindergarten, the selection of these scores corresponded to greater risk for poor reading outcomes, which was determined via criterion-related validity analyses conducted by the assessment developers. In first grade the scores for eligibility were based on previous use of the measure and normative analyses showing student scores below 26 represented approximately the 40th percentile of performance. Tutors and coaches were instructed to not provide MAVRIC intervention to students with scores at or above these targets.

Randomization occurred prior to any students beginning MAVRIC interventions. Following randomization, students assigned to treatment (i.e., the MAVRIC program) participated in MAVRIC programming according to standard procedures, while students assigned to control conditions were prevented from participating in MAVRIC until after Winter Benchmarking (although they could receive other school-based support).

For the Year 4 confirmatory impact evaluation, the primary posttest period was winter. This was based on previous research suggesting vocabulary interventions have the potential to significantly improve vocabulary skills within one semester (Marulis & Neuman, 2010), as well as on the ethical rationale to prevent students from being in control—and thus not receiving a potentially helpful intervention—for a full year. Thus, the randomized controlled trial (RCT) design that was used for the confirmatory impact evaluation used pretest data from the Fall Benchmarking period and posttest data from the Winter Benchmarking period. However, data from the Spring Benchmarking period were also used for sensitivity analyses to determine if effects were consistent across various conditions (see Data Analysis for Confirmatory Impact Evaluation section below).

Posttest assessment data were collected using the same assessments as used during pretesting in January, and again in the spring of the school year (late April). These data were collected for students who participated in the MAVRIC interventions and control students. The resulting data were used for the confirmatory impact analyses (i.e., RCT design), sensitivity analyses, and to produce descriptive data for the first two exploratory impact evaluation research questions concerning vocabulary skill growth and at-risk status of students who received MAVRIC interventions and those who did not.

In addition, the exploratory evaluation in Year 4 provided results from ongoing progress assessment data collection (i.e., exploratory impact evaluation research question 3). These data were entirely descriptive in that they were collected exclusively from students receiving the MAVRIC interventions. In prekindergarten, they consisted of the IGDIs Picture Naming (progress assessment version), given on a monthly basis. In kindergarten and first grade, progress assessments consisted of weekly researcher-made assessments that are consistent with previous research (e.g., two-question vocabulary assessment, Coyne, McCoach, & Kapp, 2007). Evaluation results report data describing student performance on these assessments (e.g., mean and standard deviation).

Participants

Year 4 activities occurred in 58 Minneapolis and St. Paul city schools and prekindergarten centers, the same urban school districts that participated in prior years. All confirmatory impact evaluation activities were conducted under Institutional Review Board oversight (and approval) via the University of Minnesota. Activities were focused within Minneapolis

Public Schools, because St. Paul Public Schools had only one previous year of implementing MAVRIC programming in any capacity, and had no prior experience implementing first grade MAVRIC programming. Thus, St. Paul Public Schools focused on integrating the MAVRIC program components with other educational programming and did not participate in confirmatory impact evaluation activities.

Across all schools, 886 students participated in MAVRIC programming during the 2016-2017 school year (note: control students are not included in this number, but treatment students are included). Some schools decided not to participate in the confirmatory impact evaluation. Participation in the confirmatory evaluation included 25 schools and 605 students, of which a total of 340 were assigned to treatment (i.e., to receive MAVRIC programming as typically implemented), and 265 were assigned to control. Table 1 shows demographic data (collected from school records) for student participants as well as district-wide demographics for each school district. **See Data Analysis for Confirmatory Impact Evaluation section for information on attrition and missing data and how the impact of missing data was assessed.**

By grade, 381 prekindergarten, and 281 kindergarten, and 224 first grade students participated in the MAVRIC intervention program over the course of the school year. After including control students,143 prekindergarten, 227 kindergarten, and 235 first grade students participated in the confirmatory impact evaluation. Parents of student participants in the impact evaluation were provided consent forms prior to randomization, and data from students from non-consenting parents were removed prior to group assignment.

Table 1

Demographic Variable	Participants			Minneapolis Public Schools	St. Paul Public Schools
	All MAVRIC n=886	ITT*[,a]			
		Treatment n=340	Control n=265		
Gender (% Female)	50%	51%	45%	n/a	n/a
Ethnicity					
American Indian	2%	2%	2%	3%	1%
Asian	21%	9%	13%	6%	32%
Black	50%	65%	60%	36%	27%
White	11%	12%	13%	34%	21%
Hispanic	12%	9%	9%	18%	14%
English Learner	39%	29%	29%	23%	34%

*ITT=Intent to Treat, which was the full sample of students included in the confirmatory impact evaluation.

[a] Baseline equivalence on demographic variables was evaluated using Chi-squared tests. Group equivalence was established on all demographic variables, with the exception of first grade English Learner status. A Welch two-sample t-test for pre-test scores showed nonsignificant baseline differences in vocabulary scores in each grade. Final analyses included demographic covariates and baseline scores.

Measures and Data Collection

Table 2 summarizes information about the measures used to collect data for answering the exploratory and confirmatory impact questions, with the top panel describing progress assessment, and the bottom panel describing the pre- and posttest assessments. All measures are commercially available with published technical characteristics (see the right-hand column of Table 2 for a brief overview), except for the Two-Question and Receptive/Expressive mastery progress assessments (which were constructed as part of the MAVRIC project). All measures were developed to be administered in English. Data from the pre- and posttest assessments were collected three times annually for all students, at the beginning (Sept-Oct), middle (Jan), and end (April-May) of the school year. At these time points, treatment and control group students were assessed using the same procedures with the assessment that corresponded to their grade level. The administrators for these data were project staff for first grade in the fall and winter (due to classwide administration) and tutors for kindergarten and prekindergarten at each time point, in addition to spring posttest periods for first grade. Tutors scored all completed student assessments using scoring keys. Progress assessment data were collected monthly (prekindergarten) or weekly (kindergarten and first grade) only for participating students.

The Individual Growth and Development Indicators (IGDIs) Picture Naming (screening version) measures used in prekindergarten and kindergarten for pre- and posttest consist of 15 items developed using item response theory that are designed to provide maximum information regarding students' risk status within fall, winter, and spring time periods. Target scores for the IGDIs Picture Naming (screening measure) were 6 for prekindergarten and 11 for kindergarten. Administration is completed in a quiet one-on-one setting and takes approximately 1 minute. The assessment includes standardized administration directions and alternative forms for use at each assessment period.

The 4,000 Word Listening Test was used for pre- and posttest assessments for first grade students. The measure consists of 40 items selected from high-frequency word banks for students in grades 1-4. The target score for 4,000 Word Listening Test was 26. Administration is completed in the whole-class setting and takes about 30 minutes. During administration, one administrator delivered instructions while at least two other adults supported students in following the instructions. The assessment uses standardized administration directions.

Table 2. Description of Impact Evaluation Measures for MAVRIC

	Progress Assessment (Exploratory Impact Questions)				
	Measure	Score Range	Administration	Description	Research Base
Prekindergarten	Individual Growth and Development Indicators, Picture Naming Progress Assessment	0 - 99+	Approximately 1 minute (one-on-one)	General Outcome Measure that assesses growth in students' proficient expressive vocabulary skills	Reliability coefficients .44 < *r* < .78 (McConnell et al., 2002); Validity coefficients .56 < *r* < .81 (Missall & McConnell, 2004)
Kindergarten	Receptive/Expressive Mastery Test	0 – 12	Approximately 1 minute (one-on-one)	Mastery measure that assesses students' maintained learning of words taught in vocabulary intervention	Recommended by National Reading Panel (2000); Increasing use in peer-reviewed research (e.g., Coyne, McCoach, & Kapp, 2007)
First Grade	Two-Question Assessment	0 - 10	Approximately 10 minutes (group)	Mastery measure that assesses students' maintained learning of words taught in vocabulary intervention	Recommended by National Reading Panel (2000); Increasing use in peer-reviewed research (e.g., Coyne, McCoach, & Kapp, 2007)

	Pre- and Post-Assessment (Exploratory and Confirmatory Impact Questions)				
	Measure	Score Range	Administration	Description	Research Base
Prekindergarten & Kindergarten	Individual Growth and Development Indicators, Picture Naming Screening Assessment	0 - 15	Approximately 1 minute (one-on-one)	General Outcome Measure that assesses students' risk for expressive vocabulary skills	Reliability coefficients *r* > .80 (McConnell & Greenwood, in press); Validity coefficients .54 < *r* < .75 (Wackerle-Holman & Bradfield, 2010)
First Grade	4,000 Word Listening Test	0 - 40	Approximately 20 minutes (group)	Standardized vocabulary measure that assesses student knowledge of 4,000 most frequent words.	Reliability coefficients *r* = .79 (Graves & Sales, 2009); Validity coefficient *r* = .64 (Graves & Sales, 2009)

Data Analysis for Impact Evaluation Questions

All exploratory impact questions were addressed with descriptive analyses, calculated using pre and posttest mean values (standard deviations), the range of lowest and highest scores, and the percentage meeting grade-specific cut scores. These descriptive statistics were calculated for each seasonal benchmark and include the sample size at each time period. In addition, descriptive analyses examined changes in vocabulary scores by determining the growth in scores from Fall to Winter and Spring benchmark periods.

The confirmatory impact question "What is the impact on vocabulary improvement for at-risk prekindergarten, kindergarten, and 1st grade students who participate in MAVRIC compared to similar students who do not receive MAVRIC interventions?" was answered by fitting a series of single and multilevel regression models to data from the final analytic sample obtained from the initial sample of 605 students. The final analytic sample included students who had pre and posttest scores on the outcome measures (~97% of excluded cases) and had pre and posttest scores within valid date ranges (<3% of cases). The latter exclusionary criterion represented a small number of cases that were grossly inconsistent (by >2 months) with the MAVRIC benchmark periods, primarily due to extenuating circumstances (e.g., intervening life events for tutor). Of the original 605 students, 62 students (10% of the initial sample) met exclusion criteria; the remaining students (n=543) formed the final intent to treat (ITT) analytic sample. A Welch two-sample t-test for pre-test group equivalence across treatment groups showed no statistically significant differences in vocabulary scores across any grade level, and chi-squared tests showed balance across all demographic variables (except first grade English Language Learner status).

After the analytic sample was identified, single level regression models with the effect of schools treated as fixed effects were fit to the data for each grade using the open source statistical programming language R (R Core Team, 2017). Initially, single level models estimating the treatment effect (i.e. effect of MAVRIC) were fit controlling for pre-test score in addition to the demographic variables ethnicity, gender, English language learner status, and the student's home language. These models suggested that students who received MAVRIC interventions performed no differently in their final test scores than students in the control group (i.e. there was a non-significant treatment effect at $\alpha = 0.05$ for all MAVRIC programs).

Fitting models for the final impact evaluation analyses continued by testing multilevel models for all grades while controlling for the same set of variables included in the single level models. For these models, the lme4 package (Bates, 2017) in R was used. Multilevel models treated schools as random effects and allowed for the treatment effect to vary across schools. The output for the multilevel models suggested that the treatment effect of the MAVRIC interventions mirrored results of the single level models.

Because there was no substantial difference in the treatment effect estimate between the single and multilevel regression models, all subsequent analyses used the simpler single level models. Additionally, there proved to be no substantial difference in the treatment effect estimate for the single level models which controlled for a variety of variables (e.g. demographics variables ethnicity, gender, English language learner status, and student home language, as well as pre-test) compared to the single level models which controlled for only pre-test scores. For this reason, the authors report the results from the single level models that only control for pre-test scores.

Missing Data

Given the applied nature of this evaluation, the potential influence of missing data was also investigated. While there were no missing data on the control variables in the regression analyses, missing data were present for the outcome measure. Specifically, out of the 605 students initially considered for analysis, 62 (10%) met exclusion criteria and were subsequently removed from the fitted models. To understand whether and how these students differed from the sample as a whole, a logistic regression model was fit to predict whether a student had missing data for fall or winter scores using ethnicity, gender, English language learner status, student home language, tutor fidelity, and the treatment variable as predictors in the model. The results suggested there was no relationship between the likelihood of having missing data and the tested covariates ($p > 0.37$). For this reason, the authors believe the results of the analyses are generalizable to the broader sample and not just the sample with complete data.

Implementation Evaluation Design and Results

Implementation evaluations are a standard part of Reading Corps. It is routine to collect data and conduct descriptive analyses for the dosage students receive of MAVRIC programming. In addition, fidelity data are collected for all evidence-based interventions implemented throughout the school year, as well as for all assessments in order to ensure the data are accurate and valid. The MAVRIC implementation evaluation included these practices and therefore focused on descriptive analyses (i.e., means, standard deviations, and ranges) to show (a) the amount of exposure to MAVRIC interventions participating students received, and (b) the degree to which MAVRIC interventions and assessments were administered as intended.

To determine the degree to which MAVRIC assessments and interventions were administered as intended, the implementation evaluation included fidelity checklists (see Appendix B for example). The fidelity checklists described key steps for administering the interventions (e.g., vocabulary words were explicitly introduced and defined according to the script; the deep processing activities were followed) and assessments (e.g., the administrator followed standardized script instructions for assessment items), and were used during coaching observations which occurred roughly monthly and before all benchmark data collection periods. Coaches used the checklists during observations to record whether or not each key step for administering the assessments and interventions was observed. Within the coaching process, no prescriptive guidelines were provided to coaches for specific follow-up actions based on the checklist results; however, all coaches followed a general guideline to provide modeling and constructive feedback for all incorrect administration steps as well as when an observation resulted in an overall fidelity percentage that was below the target goal of 90%.

The data collected from these checklists were compiled and stored in the Reading Corps data management system. They provided an overall percentage of accurately-implemented steps for the assessment and the intervention. Approximately 15% of the assessment administration periods were randomly identified to be observed independently using the fidelity checklists, resulting in 155 fidelity checklists completed for the IGDIs 2.0 Picture Naming measure, and 69 fidelity checklists completed for the 4,000 Word Listening Test. Intervention administration fidelity data were collected for all tutors approximately monthly but not on a prescribed schedule. Thus, missing data from intervention checklists were only identifiable at an aggregate level (e.g., a coach failing to conduct and record fidelity checks), but any observed missing data resulted in follow-up reminders from project leads. Intervention fidelity observations were completed for 116 first grade intervention sessions, 118 kindergarten intervention sessions, and 212 prekindergarten intervention sessions. Data from fidelity measures were not included in the main impact analyses, but were included in follow-up sensitivity analyses; however, missing fidelity data was not addressed from a statistical or methodological approach.

Tables 3 and 4 below show the fidelity for administration of assessments and interventions. Table 3 indicates very high levels of fidelity were observed for administration of the assessments (i.e., >95% mean fidelity levels). In Table 4, a similarly high fidelity level is shown for the kindergarten and first grade RRA interventions (=>90%), which indicated the intervention was delivered as intended. The mean fidelity for the Repeated Read Aloud intervention in prekindergarten was below the conventional goal of 90%, suggesting the potential for additional tailored training and coaching.

In addition to ensuring the administration of the assessments and interventions was accurate, the implementation evaluation for MAVRIC also examined the degree to which assessments were reliably scored. For this purpose, a random selection of assessment administrations was co-scored by MAVRIC project evaluation staff. As shown in Table 5, approximately 16% of IGDIs Picture Naming screening assessments and 19% of the 4,000 Words Listening Test results were scored by a second, independent scorer. The two scores were in agreement on over 99% of the assessment items scored, suggesting that the assessment results were reliably scored.

Table 3: Fidelity of Assessment Data

Measure	Total Complete Fidelity Checks	Fidelity Range Reported	Mean (Standard Deviation) Percent Fidelity Reported
IGDIs 2.0 Picture Naming Test	155	86-100%	98.9% (.03%)
First 4,000 Words Listening Vocabulary Test	69	83-100%	98.5% (.03%)

Table 4: Fidelity of Intervention Data

Measure	Total Complete Fidelity Checks	Fidelity Range Reported	Mean (Standard Deviation) Percent Fidelity Reported
Prekindergarten **Repeated Read Aloud**	212	29-100%	86.1% (.15%)
Kindergarten **Repeated Read Aloud**	118	24-100%	90.0% (.13%)
First Grade **Repeated Read Aloud**	116	68-100%	93.6% (.06%)

Table 5: Reliability of Assessment Data

Measure	Total Number of Assessments Scored	Number of Assessments Independently Co-Scored	Percent of Assessment Data Co-Scored	Inter-Scorer Reliability
IGDIs 2.0 Picture Naming Test	1,561	255	16.3%	99.2%
First 4,000 Words Listening Vocabulary Test	1025	196	19.1%	99.6%

Table 6 summarizes participation for students with MAVRIC interventions. The table displays average weekly minutes of tutoring, average days of tutoring in a week, as well as the variability in those metrics. These data show that students received approximately 3 to 4 sessions of tutoring each week, on average. That resulted in an average of approximately 31 minutes per week for prekindergarten, 67 minutes per week for kindergarten, and 70 minutes per week for 1st grade. Although specific benchmarks for determining whether these participation results are within expectations for similar interventions are difficult to determine, effect sizes for various dosage characteristics remained consistently high for similar interventions reported for prekindergarten and kindergarten students. Specifically, durations ranging from less than 1 week to more than 42 days across session lengths either greater or lesser than 20 min per session produced effect sizes ranging with a relatively narrow range .85 to 1.12 (Marulis & Neuman, 2010). Thus, when compared to peer-reviewed studies of researcher-led studies, dosage was within expected ranges.

Table 6: Participation by Grade

Grade	Weeks Participating		Minutes per Week		Sessions per Week	
	Average	Range	Average	Range	Average	Range
Prekindergarten	10.4	1-26	31.2 (Goal = 40)	6-75	3.2	1-5
Kindergarten	15.4	1-29	66.8 (Goal = 70)	20-92	3.4	1-5
First Grade	16.9	1-30	70.2 (Goal = 70)	43-100	3.6	2-5

YEAR 4 EVALUATION RESULTS

Improvement and Performance Relative to Target Scores (Exploratory Impact Research Questions 1 and 2)

Prekindergarten

Table 7 displays prekindergarten results for exploratory impact questions 1 and 2. Mean, standard deviation, range, and percent of students meeting the target score are shown in the top three sections of the table, organized by benchmark period (Fall, Winter, Spring). Columns correspond to (a) all students who participated in the MAVRIC intervention program, (b) students randomly assigned to participate in the MAVRIC intervention program in the first semester, and (c) students randomly assigned to not participate in the MAVRIC intervention program in the first semester (but were permitted to participate in the program in the second semester).

Average vocabulary scores improved across the three benchmark periods for all groups. In the Fall, the average score for all students was slightly above three correctly identified words, and for the impact evaluation groups was slightly below three. By Winter, the intervention group had the highest average of 5.26, but all students—including the control group—had made gains toward or above five correctly identified words. By Spring, average scores had increased further to approximately 5.5 correctly identified words for all students, and over 6.5 for both impact evaluation subgroups (after Winter, control students were able to access MAVRIC interventions).

The percent of students meeting the target score increased substantially over the course of the evaluation period. Intervention group students made the most pronounced growth, with approximately 43.50% reaching target scores by winter and 68.25% reaching target by Spring. However, control group students reached target scores at nearly the same percentage (41.5% in winter and 66.00% in spring), and nearly 50% of all students reached target scores by Spring. It should be noted that although no students from the impact evaluation subset had scores at target in the Fall, and were thus appropriately identified as eligible, approximately 17.5% of control group students *exceeded* target scores in the Fall but were provided MAVRIC support after Winter due to scores falling below target after the passing of the first semester.

The bottom section of Table 7 also reports mean growth for students with available data between fall and winter and fall and spring time periods. Overall, growth was greatest for intervention students from Fall to Winter, although all students improved by more than two correctly identified words from Fall to Winter. From Fall to Spring, growth was approximately four correctly identified words for both groups of students in the impact evaluation subset, as opposed to three for all students.

Table 7. Prekindergarten Benchmark Performance and Vocabulary Average Growth

	Students Served – All Sites		Impact Evaluation Subset			
			Intervention Students		Control Students	
Fall (Pretest) Benchmark						
Number of students with pretest	216		78		65	
Range	-1 – 11		-1 – 5		-1 – 5	
Average score (Standard Deviation)	3.24 (2.70)		2.51 (1.84)		2.92 (1.86)	
Percent meeting target score	17.59%		--		--	
Winter (Posttest) Benchmark						
Number of students with posttest**	306		69		53	
Range	-1 – 15		-1 – 13		0 – 11	
Average score (Standard Deviation)	4.37 (2.84)		5.26 (2.68)		5.06 (2.42)	
Percent meeting target score	31.70%		43.48%		41.51%	
Spring Benchmark						
Number of students with posttest**	284		63		50	
Range	-1 – 13		0 – 13		-1 – 13	
Average score (Standard Deviation)	5.52 (2.89)		6.63 (2.76)		6.70 (2.89)	
Percent meeting target score	48.94%		68.25%		66.00%	
Growth	**F to W**	**F to S**	**F to W**	**F to S**	**F to W**	**F to S**
Number of students with scores in both windows	136	173	69	63	53	50
Range of growth*	4 pt loss to 10 pt gain	5 pt loss to 10 pt gain	2 pt loss to 9 pt gain	1 pt loss to 10 pt gain	1 pt loss to 7 pt gain	3 pt loss to 10 pt gain
Average growth (Standard deviation)	2.07 (2.55)	2.95 (2.89)	2.70 (2.30)	4.03 (2.30)	2.36 (2.08)	4.04 (2.98)

*Note: Range of growth represents the lowest growth (reflected as a loss, where applicable) and the most growth across all students in a given time period.
**Number of students at Winter and Spring benchmark periods reflects attrition throughout the year (at All Sites, also reflects students who started intervention on or after Winter Benchmark period).

Kindergarten

Table 8 displays kindergarten results for exploratory impact questions 1 and 2. Mean, standard deviation, range, and percent of students meeting the target score are shown in the top three sections of the table, organized by benchmark period (Fall, Winter, Spring). Columns correspond to (a) all students who participated in the MAVRIC intervention program, (b) students randomly assigned to participate in the MAVRIC intervention program in the first semester, and (c) students randomly assigned to not participate in the MAVRIC intervention program in the first semester (but were permitted to participate in the program in the second semester).

Average vocabulary scores generally improved across the three benchmark periods, with the exception between Winter and Spring for the impact evaluation subset of students. Decreased average scores for these students likely reflect two factors. First, the difficulty of items on the IGDI

2.0 vocabulary measure increases across the year. As a result, absolute values may show lower scores even though actual vocabulary knowledge increased. Second, for intervention students the decrease from Winter to Spring reflects a substantial proportion of students no longer receiving intervention after the Winter benchmark period and therefore may have plateaued in their growth, as suggested by nearly 43% obtaining the Winter target score (those students were no longer eligible for MAVRIC intervention supports).

In the Fall, the average score for all students was approximately seven correctly identified words, although it was marginally higher for control students. By Winter, the control group had the highest average of approximately 7.5, but intervention students made a comparable increase given their lower starting point. By Spring, average scores had continued to increase for students in the entire sample, but both groups of students in the impact subset had decreased by a similar margin.

The percent of students meeting the target score increased substantially from Fall to Winter benchmark periods, with the highest percentage of students reaching Winter target scores in the intervention group, at approximately 43%. However, nearly 40% of the control group students reached the Winter target and 30% of students at all sites reached the Winter target. From Winter to Spring, the percentage of students reaching target scores decreased substantially, likely due to the reasons noted above.

The bottom section of Table 8 also reports mean growth for students with available data between fall and winter and fall and spring time periods. Growth was comparable for intervention and control students from Fall to Winter, with both improving slightly more than students at all sites. From Fall to Spring, growth decreased relative to the improvement between Fall and Winter, and was relatively similar across all groups (ranging from 1.73 to 1.84).

Table 8: Kindergarten Benchmark Performance and Vocabulary Average Growth

	Students Served – All Sites		Impact Evaluation Subset			
			Intervention Students		Control Students	
Fall (Pretest) Benchmark						
Number of students with pretest	210		127		100	
Range	0 – 12		1 – 10		1 – 10	
Average score (Standard Deviation)	7.02 (2.62)		7.17 (2.50)		7.31 (2.30)	
Percent meeting target score	--		--		--	
Winter (Posttest) Benchmark						
Number of students with posttest**	253		112		89	
Range	1 – 15		1 – 15		4 – 14	
Average score (Standard Deviation)	8.87 (2.90)		9.30 (3.12)		9.53 (2.49)	
Percent meeting target score	30.43%		42.86%		39.33%	
Spring Benchmark						
Number of students with posttest**	246		110		85	
Range	0 – 15		3 – 14		2 – 14	
Average score (Standard Deviation)	8.99 (2.67)		8.82 (2.62)		9.05 (2.40)	
Percent meeting target score	30.08%		21.82%		27.06%	
Growth	**F to W**	**F to S**	**F to W**	**F to S**	**F to W**	**F to S**
Number of students with scores in both windows	162	183	112	110	89	85
Range of growth*	3 pt loss to 10 pt gain	6 pt loss to 10 pt gain	3 pt loss to 10 pt gain	5 pt loss to 10 pt gain	3 pt loss to 9 pt gain	3 pt loss to 8 pt gain
Average growth (Standard deviation)	2.09 (2.14)	1.84 (2.44)	2.23 (2.29)	1.73 (2.46)	2.31 (2.26)	1.80 (2.44)

*Note: Range of growth represents the lowest growth (reflected as a loss, where applicable) and the most growth across all students in a given time period.
**Number of students at Winter and Spring benchmark periods reflects attrition throughout the year (at All Sites, this number also reflects the fact that some students started intervention on or after Winter Benchmark period).

First Grade

Table 9 displays first grade results for exploratory impact questions 1 and 2. Mean, standard deviation, range, and percent of students meeting the target score are shown in the top three sections of the table, organized by benchmark period (Fall, Winter, Spring). Columns correspond to (a) all students who participated in the MAVRIC intervention program, (b) students randomly assigned to participate in the MAVRIC intervention program in the first semester, and (c) students randomly assigned to not participate in the MAVRIC intervention program in the first semester (but were permitted to participate in the program in the second semester).

Average vocabulary scores generally improved across the three benchmark periods. In the Fall, the average score for all students was approximately ranged between approximately 20.50 and

20.75. By Winter, average scores had increased by approximately five additional correct items on the 4,000 Word Listening assessment. By Spring, average scores had continued to increase for students in the entire sample, though not by as much as from Fall to Winter.

The percent of students meeting the target score increased substantially from Fall to Winter benchmark periods, with the highest percentage of students reaching Winter target scores in the control group, at approximately 56%. From Winter to Spring, the percentage of students reaching target scores continued to increase, with the most substantial increase occurring for intervention group students, who with a 21.5% increase exceeded the approximately 15% increase for control group students and 12% increase overall.

The bottom section of Table 9 also reports mean growth for students with available data between fall and winter and fall and spring time periods. Growth was greater for control students than for intervention students at both Winter and Spring, while intervention students grew more than the entire sample.

Table 9: First Grade Benchmark Performance and Vocabulary Average Growth

	Students Served – All Sites		Impact Evaluation Subset			
			Intervention Students		Control Students	
Fall (Pretest) Benchmark						
Number of students with pretest	202		135		100	
Range	10 – 26		10 – 25		9 – 25	
Average score (Standard Deviation)	20.73 (3.62)		20.47 (3.67)		20.64 (3.95)	
Percent meeting target score	1.49%		--		--	
Winter (Posttest) Benchmark						
Number of students with posttest**	191		124		94	
Range	14 – 33		14 – 33		17 – 34	
Average score (Standard Deviation)	24.60 (3.99)		25.35 (4.16)		25.70 (4.21)	
Percent meeting target score	39.79%		51.61%		56.38%	
Spring Benchmark						
Number of students with posttest**	196		119		93	
Range	17 – 34		17 – 34		17 – 35	
Average score (Standard Deviation)	26.63 (3.55)		27.12 (3.60)		27.37 (4.10)	
Percent meeting target score	64.31%		73.11%		69.89%	
Growth	**F to W**	**F to S**	**F to W**	**F to S**	**F to W**	**F to S**
Number of students with scores in both windows	152	177	124	119	94	93
Range of growth*	5 pt loss to 18 pt gain	5 pt loss to 18 pt gain	5 pt loss to 18 pt gain	2 pt loss to 18 pt gain	4 pt loss to 18 pt gain	2 pt loss to 17 pt gain
Average growth (Standard deviation)	4.32 (4.29)	6.02 (3.96)	4.69 (4.38)	6.45 (3.84)	5.11 (4.15)	6.80 (4.16)

*Note: Range of growth represents the lowest growth (reflected as a loss, where applicable) and the most growth across all students in a given time period.

**Number of students at Winter and Spring benchmark periods reflects attrition throughout the year (at All Sites, also reflects students who started intervention on or after Winter Benchmark period).

Performance of Participating Students on Vocabulary Progress Assessments (Exploratory Impact Research Question 3)

To answer this research question, the average performance of each student was calculated on progress assessments. For prekindergarten this was produced from the monthly progress assessment data collected using the IGDIs Picture Naming progress assessment. The overall average for prekindergarten performance on this measure was 14.2 (SD = 7.7), meaning students correctly identified over 14 words on average each month. For kindergarten, average performance was produced from the weekly progress assessments collected using the assessment developed as part of the Year 1 activities for this project. The overall average performance for kindergarten was 10.7 (SD = 1.5). This means students received an average score of approximately 11 out of a maximum score of 12 on the weekly assessments. (NOTE: the kindergarten progress assessment included both receptive/identification and expressive/labeling components for assessing the taught words from a given week). For first grade, average performance was produced using the weekly progress assessments. The overall average performance for first grade was 6.6 (SD = 2.5). This means students correctly identified approximately 6/10 words that were taught each week. For both kindergarten and first grade, the average scores on the progress assessments reflect notable improvements from previous year average scores.

Confirmatory Impact Evaluation—Main Analysis Results

The main analysis approach for the confirmatory impact evaluation followed a procedure in which several regression models were tested to determine the optimal analytic model (see p. 21 for details). In addition to interpreting results from the parsimonious single-level regression models, a density plot figure showing the distribution of post-test scores for both MAVRIC and control group students was produced for each grade. These figures illustrate graphically the observed effect and serve to complement the results from the inferential analyses.

Prekindergarten (IGDI 2.0)

Figure 2 below presents a plot of post-test scores for prekindergarten students on the IGDI 2.0 by whether students received the MAVRIC intervention. The vertical dashed line represents the IGDI 2.0 post-test score mean for each group. Table 10 below contains the treatment effect estimate based on the parsimonious single level regression model controlling for pre-test score. There was no statistically significant treatment effect for prekindergarten students ($\beta = 0.35, p = 0.37$). The results were consistent when fitting a more complex single level model additionally controlling for demographic variables as well as when fitting a multilevel model that treats schools as a random effect.

Figure 2: Plot of post-test scores on IGDI 2.0 by MAVRIC status for prekindergarten

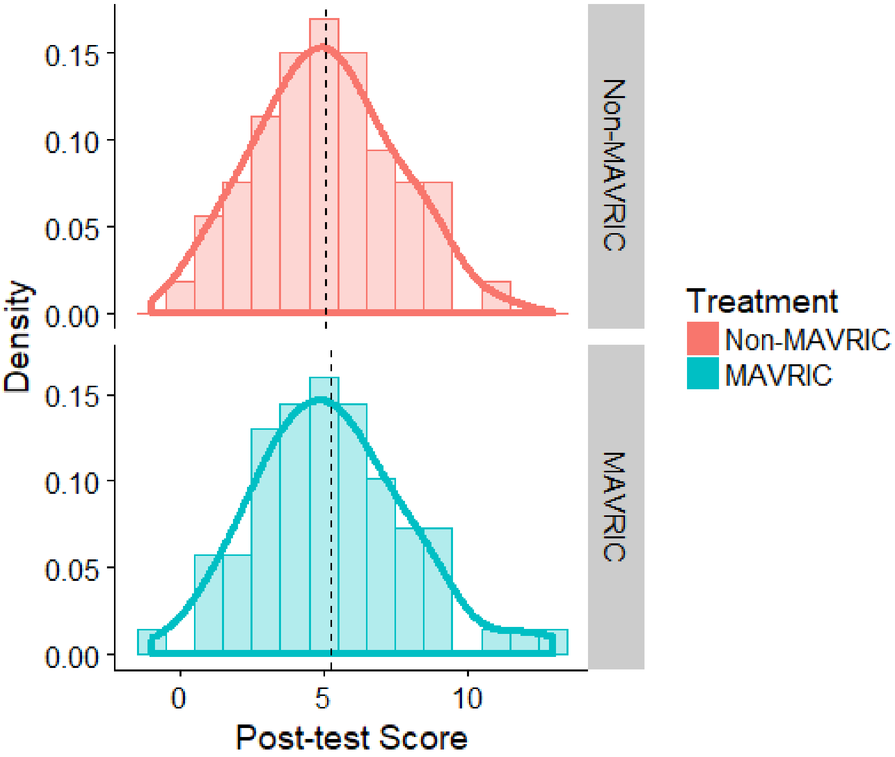


Table 10: Effect of MAVRIC on post-test IGDI 2.0 scores in prekindergarten

Type	N	Estimate	SE	p-value
Treatment Effect	120	0.35	0.39	0.37

Note: Results based on a model controlling for pre-test score and treatment assignment.

Kindergarten

Figure 3 below presents a plot of post-test scores for kindergarten students by whether they received the MAVRIC intervention. The vertical dashed line represents the post-test score mean for each group. Table 11 below contains the treatment effect estimate based on the single level regression model controlling for pre-test score. There was no statistically significant treatment effect for kindergarten students ($\beta = -0.12, p = 0.71$). The results were consistent when fitting a more complex single level model additionally controlling for demographic variables as well as when fitting a multilevel model that treats schools as a random effect.

Figure 3: Plot of post-test scores by MAVRIC status for kindergarten

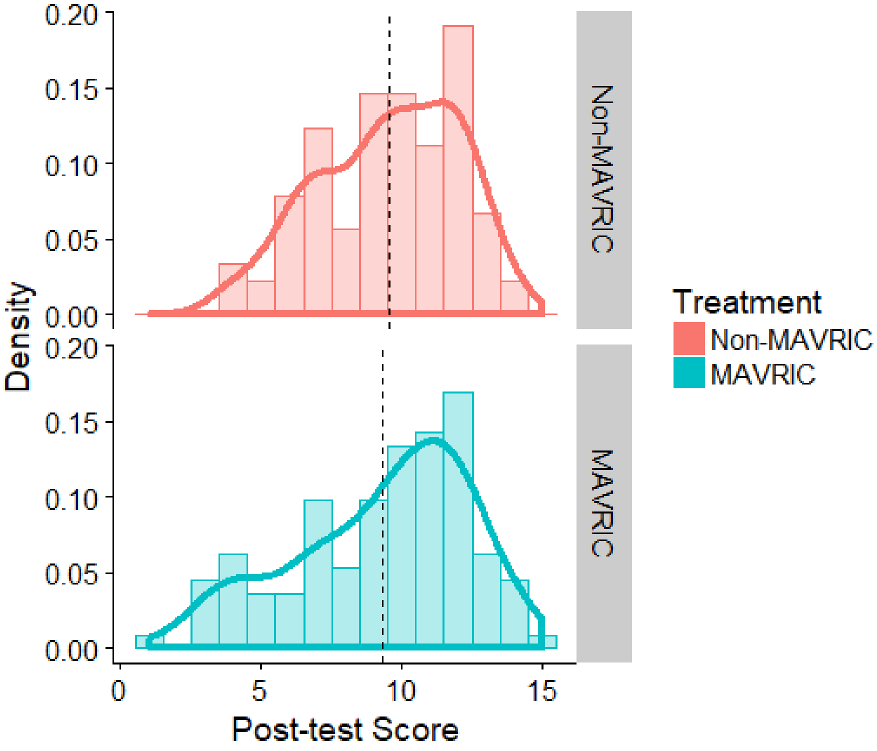


Table 11: Effect of MAVRIC on post-test scores in kindergarten

Type	N	Estimate	SE	p-value
Treatment Effect	200	-0.12	0.31	0.71

Note: Results based on a model controlling for pre-test score and treatment assignment.

First Grade

Figure 4 below presents a plot of post-test scores for 1st grade students by whether they received the MAVRIC intervention. The vertical dashed line represents the post-test score mean for the corresponding group. Table 12 below contains the treatment effect estimate based on the single level regression model controlling for pre-test score. There was no statistically significant treatment effect for kindergarten students ($\beta = -0.38, p = 0.46$). The results were consistent when fitting a more complex single level model additionally controlling for demographic variables as well as when fitting a multilevel model that treats schools as a random effect.

Figure 4: Plot of post-test scores by MAVRIC status for 1st grade

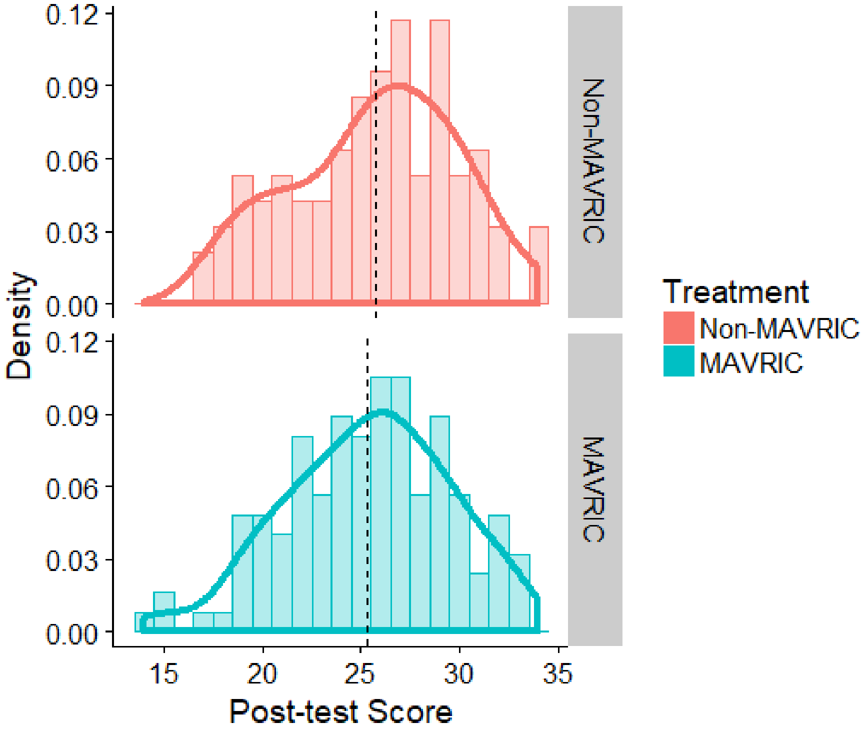


Table 12: Effect of MAVRIC on post-test scores in 1st grade

Type	N	Estimate	SE	p-value
Treatment Effect	217	-0.38	0.52	0.46

Note: Results based on a model controlling for pre-test score and treatment assignment.

Confirmatory Impact Evaluation—Results from Sensitivity Analyses

In addition to the planned impact evaluation analyses, data collected from the MAVRIC program in Year 4 also permitted several "sensitivity analyses" designed to explore additional factors that were potentially related to student vocabulary outcomes. This included considerations such as (a) the nature of the outcome measure for prekindergarten, (b) the dosage of MAVRIC intervention students received, and (c) the level of fidelity tutors demonstrated in delivering the MAVRIC interventions. The results of these sensitivity analyses are described below.

Prekindergarten (Progress Assessment)

In prekindergarten, student performance was also collected on the IGDI Progress Assessment measure, which as described in the Measures section above provides a fluency-based assessment of student vocabulary skills. Winter post-test data were collected for all students on this measure as part of standard Reading Corps practices, and were therefore available for sensitivity analyses. Because the IGDI Progress Assessment measures student performance based on the overall number of known vocabulary concepts (assessed as a count of correctly identified pictures within a 1 minute timeframe), it may have been more sensitive to student improvement because it offered students more opportunities to demonstrate vocabulary skill growth, as opposed to constraining the number of items to which students can respond, as with the IGDI 2.0 measure.

The figure below presents a plot of post-test scores for prekindergarten students on the IGDI Progress Assessment by whether students received the MAVRIC intervention. The vertical dashed line represents the IGDI 1 Progress Assessment post-test score mean for the corresponding group. Table 13 below contains the treatment effect estimate based on the single level regression model controlling for pre-test score. There was no statistically significant treatment effect for prekindergarten students ($\beta = 0.48, p > 0.05$). The results were consistent when fitting a more complex single level model that controlled for demographic variables as well as when fitting a multilevel model that treated schools as a random effect.

Figure 5: Plot of post-test scores on IGDI Progress Assessment by MAVRIC status for prekindergarten

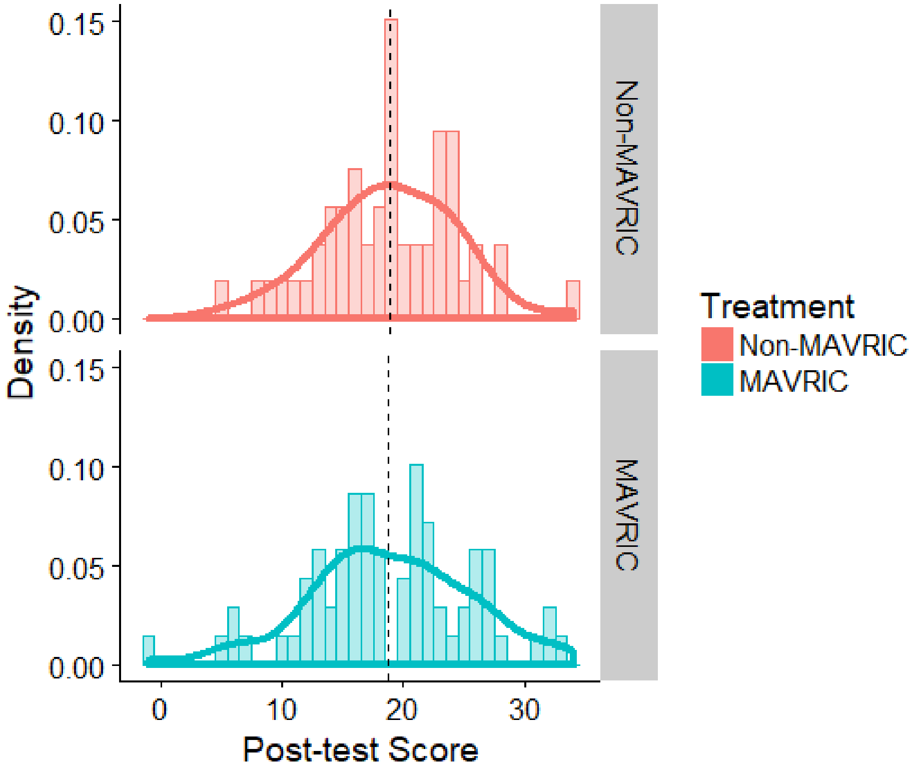


Table 13: Effect of MAVRIC on post-test IGDI Progress Assessment scores in prekindergarten

Type	N	Estimate	SE	p-value
Treatment Effect	120	0.48	0.98	0.62

Note: Results based on a model controlling for pre-test score and treatment assignment.

Dosage and Tutor Fidelity

In an effort to determine whether the treatment effect estimate was sensitive to the dosage of the MAVRIC interventions, the authors stratified students who received the MAVRIC interventions into two groups on the basis of their dosage. The first group of students is referred to as the "Optimal Dosage" (OD) group, and consisted of only students who a) received a minimum of 10 weeks of MAVRIC intervention at no less than 20 minutes per week for prekindergarten students and b) received a minimum of 10 weeks of MAVRIC intervention at no less than 60 minutes per week for kindergarten and first grade students. The second group was referred to as the "High Tutor Fidelity" (HTF) group, and it consisted of only students who received MAVRIC interventions from tutors with a fidelity rating of 90% or greater. Table 14 below provides a description of these dosage groups.

Table 14: Dosage-based MAVRIC groups

Group	Grade	Criteria
Optimal Dosage	Prekindergarten	>= 10 weeks of MAVRIC >= 20 minutes per week
Optimal Dosage	Kindergarten, 1st Grade	>= 10 weeks of MAVRIC >= 60 minutes per week
High Tutor Fidelity	Prekindergarten, Kindergarten, 1st Grade	>= 90% tutor fidelity

The criteria for the OD group removed 58 students from the analytic sample (i.e. from 543 to 485), a 10% reduction. As with the ITT analyses, the regression results suggested a non-significant treatment effect when comparing the OD group to the non-MAVRIC group for all grades and assessments. Summaries of model results are found in the table below.

Table 15: Optimal Dosage Group Treatment Effect by Grade

Type	N	Estimate	SE	p-value
Prekindergarten, IGDI 1.0 assessment	100	-0.56	1.02	0.57
Prekindergarten, IGDI 2.0 assessment	100	0.14	0.43	0.74
Kindergarten	180	-0.09	0.33	0.76
First grade	203	-0.10	0.52	0.84

Note: Results based on models controlling for pre-test score and treatment assignment.

The criteria for the HTF group proved to be more restrictive, removing 165 (30%) due to their tutor fidelity scores. Consistent with previous results, the regression results suggested a non-significant treatment effect when comparing the HTF group to the non-MAVRIC group for all grades and assessments. Summaries of model results are found in the table below.

Table 16; High Tutor Fidelity Group Treatment Effect by Grade

Type	N	Estimate	SE	p-value
Prekindergarten, IGDI 1.0 assessment	69	0.84	1.42	0.55
Prekindergarten, IGDI 2.0 assessment	69	-0.56	0.58	0.33
Kindergarten	136	-0.28	0.40	0.49
First grade	172	0.10	0.56	0.85

Note: Results based on models controlling for pre-test score and treatment assignment.

YEAR 4 EVALUATION RESULTS: INTERPRETATION AND LIMITATIONS

Interpretation: Impact and Implementation Evaluation Results

The Year 4 impact evaluation of the MAVRIC program included exploratory and confirmatory components. The exploratory component was intended to describe how vocabulary skills changed for students who did and did not receive intervention. The confirmatory component was intended to produce defensible conclusions about the effectiveness of the program. Overall, the results indicated MAVRIC does not produce a significant, positive impact on student vocabulary skills. This finding was stable across grades and potential mediating variables (e.g., tutor fidelity; dosage), and was also reflected in the descriptive analysis tables for each grade that showed relatively comparable performance at pretest and posttest for treatment and control groups. Although this finding was inconsistent with meta-analytic reviews of vocabulary programs (Marulis & Neuman, 2010) it was consistent with more recent work that found vocabulary interventions were the only approach out of twenty studied to not observe a significant impact for reading (Gersten et al., 2017).

Despite results that do not support a moderate level of evidence as per the Social Innovation Fund evidence guidelines (REF), implementation evaluation results demonstrated that the program appears successful in training tutors to administer assessments and deliver interventions. Specifically, with the exception of delivering prekindergarten interventions, which were delivered with approximately 86% accuracy, tutors administered assessments, delivered interventions, and scored tests with high accuracy (i.e., all averaged >90% accuracy). These results suggest that MAVRIC tutors, who come from various backgrounds, can learn to deliver the procedural elements of MAVRIC assessment and intervention components. Further, these results are consistent with prior year findings showing similar levels of implementation accuracy. Procedural accuracy does not encompass all aspects of effective program implementation but it is an essential component of implementation (O'Donnell, 2008), and is noteworthy for a program that leverages AmeriCorps members with varied backgrounds. Further, given the MAVRIC tutors provided intervention in Year 4 using materials (e.g., books and target word picture cards) that had been improved for cultural diversity and sensitivity in Year 3, it is noteworthy that the updated MAVRIC content was accurately delivered to student participants. Nonetheless, the fact that strong procedural implementation and improved content did not lead to significant positive effects was ultimately disappointing. The Limitations section below discusses factors that may have limited the extent to which the Year 4 impact evaluation was able to identify a significant impact on student vocabulary skills.

Limitations

In any evaluation, limitations qualify and contextualize how results should be interpreted. Limitations for this evaluation can be organized by the following topics: (a) measurement, (b) time, and (c) implementation. A fourth set of limitations covers additional methodological issues that qualify the results.

Measurement

One of the most notable limitations for this evaluation concerns measurement. Specifically, the outcome measures may have had limited sensitivity to detect differential growth between the treatment (MAVRIC) and control conditions. Both the IGDI 2.0 and the 4,000 Word Listening Test were evidently able to measure student improvement over time, as noted by increasing mean scores across testing periods in Tables 7-9, which would be expected based on developmental acquisition of vocabulary (Hart & Risley, 1995). However, in each grade, the improvement from Fall to Winter was accompanied by considerable variation as noted by standard deviation values. Perhaps more importantly, the measures had limits with regard to assessing change that are common with vocabulary measures (National Reading Panel, 2000). Neither measure was designed for frequent repeated administration, and both were distally related to the content of the specific intervention. Such measurement issues are recognized limits of data-driven decision making within the context of vocabulary skills (Coyne, Capozzoli-Oldham, Cuticelli, & Ware, 2015). Recommendations for assessing the effectiveness of vocabulary interventions have historically included researcher-developed measures that are more proximal to the intervention content (i.e., more directly measure the vocabulary words learned in intervention) (National Reading Panel, 2000). Although such recommendations come with their own limitations—notably decreased generalizability with respect to broader claims of student vocabulary improvement—they may have been relevant to this project and are likely directions of future work for MAVRIC.

Time

A second notable limitation concerns the time students spent participating in the intervention. Only first grade students met the grade-specific dosage goal of minutes per week (i.e., 70+ minutes/week), and by a narrow margin (i.e., average minutes per week of 70.2). Prekindergarten and kindergarten missed their grade-specific dosage goal for minutes per week, and in the case of prekindergarten by a relatively wide margin. A related issue was the number of weeks students participated, which was approximately 10 for prekindergarten and 15-17 for kindergarten and first grade students. Effects of vocabulary interventions were robust across various dosage levels in meta-analyses of vocabulary intervention research (Marulis & Neuman, 2010), but there is reason to question the sufficiency of the dosage received by students in MAVRIC. Dosage effects in existing meta-analyses were not disaggregated by the type of outcome measure, which were considerably larger for "author created" as opposed to standardized measures (Marulis & Neuman, 2010). For example, intervention study for kindergartners that found promising results lasted 2 weeks and measured outcomes on a proximal researcher-developed measure (Loftus, Coyne, McCoach, Zipoli, & Pullen, 2010). Given the possibility that shorter-duration studies likely used "author created" measures, a single semester may have been insufficient to detect differences in growth between treatment and control students, given the broad/standardized nature of the IGDIs 2.0 and 4,000 Word Listening Test measures.

Implementation

A third notable limitation relates to implementation challenges that could have attenuated MAVRIC effects on the treatment group. First, intervention was typically delivered in groups of 4 students. Group size was not substantively related to effects in the aforementioned meta-analytic work (Marulis & Neuman, 2010), but it is perhaps notable that small groups like those used in MAVRIC produced relatively smaller effects (compared to individual and larger groups). Further, in other reading research, group size is clearly an important consideration, with individual interventions and those with two or three students leading to stronger effects than larger groups (Vaughn & Linan-Thomson, 2003), presumably because learning is intrinsically related to the opportunities students have to practice skills (Greenwood, Delquadri, & Hall, 1984), and in larger group sizes those opportunities are clearly limited.

A second implementation variable that may have impacted results is related to the student participant sample, which came exclusively from urban schools where the proportion of students who have experienced traumatic life experiences and therefore have considerable behavioral needs is higher relative to students in other educational settings (Thompson & Rippey Massat, 2005). Students with the highest behavioral or mental health needs should not be provided MAVRIC in lieu of individualized support, but in some schools the baseline level of trauma is such that many students have experienced at least some kind of traumatic event in their lives and have corresponding behavioral needs. In some schools, therefore, students with substantial behavioral needs might have been allowed to participate in MAVRIC. The potential impact on MAVRIC outcomes is that tutors may have struggled to support small-group behavioral management and also ensure maximal engagement during intervention. It is possible that a relatively high proportion of these students might have contributed to attenuating MAVRIC effects, given that multiple risk factors, such as low socio-economic status (SES) and non-dominant ethnicity status, corresponded with decreased effects for vocabulary interventions factors (Marulis & Neuman, 2013).

The two previous factors reflect opportunities to further refine aspects of MAVRIC implementation to promote more efficacious tutor delivery of interventions. Specifically, as noted above simple procedural accuracy (i.e., following essential intervention steps) may not be sufficient in itself for producing additional vocabulary growth, particularly for groups of higher-risk students. Qualitative aspects of implementation, such as how tutors manage challenging behaviors or promote quality engagements with and among students (O'Donnell, 2008), could potentially be addressed in a way that leads to better outcomes. For example, reducing group size might not only promote more opportunities to respond (Greenwood et al., 1984), it might also minimize challenging behaviors and facilitate engagement. Further, although tutors were provided behavioral training beginning in Year 2 of MAVRIC (see Year 2 evaluation report), training focused primarily on behavioral management (e.g., using attention to reinforce good behavior) and did not explicitly address how to facilitate *high-quality* engagement around vocabulary learning (e.g., use of rich words; Beck & McKeown, 2007). Finally, ongoing work to simplify materials may facilitate improved logistical aspects of MAVRIC intervention delivery.

Other Methodological Limitations

In addition to the substantive limitations listed above, the current evaluation results should be interpreted in light of other methodological limitations. First, the results need to be interpreted as only generalizing to urban settings with high proportions of students facing multiple risk factors. This aspect of the study leaves open a possibility that MAVRIC interventions could be efficacious in other settings (e.g., suburban schools), but such hypotheses require additional research. Second, the evaluation may have been under-powered. The power analysis began by considering established effects for vocabulary interventions (*g* = 0.88; Marulis & Neuman, 2010), but then decreased those effect estimates by more than half (i.e., *m.d.e.s.* = 0.40) to account for unknown influences such as non-educators implementing the MAVRIC interventions in exclusively urban settings. It may be that a more appropriate minimal detectable effect size is smaller in the current intervention context.

Other methodological limitations are related to the sample. First, the extent to which special education eligibility was related to (a) outcomes and (b) missing data is unknown. Descriptively, special education eligibility was likely unrelated to missing data at all grade levels (i.e., with differences of no more than two for students with and without missing data), but unfortunately special education status was unavailable for 40% of the students due to lack of school reporting. A similar issue pertains to participation in MAVRIC in prior years. Given MAVRIC was implemented at a reasonably-large scale in the same urban school district beginning in Year 2, it is likely that students had participated in a previous year. Descriptively, more treatment students had previously participated in MAVRIC than had control students, particularly for first grade (e.g., 39 treatment students had previously participated in MAVRIC compared to 24 control students). Such differences could have attenuated the potential impact on treatment students, particularly if gains from MAVRIC tend to be realized in the first year students participate. This would have primarily influenced the results for first grade students, but the generally consistent finding of non-significant results across grades suggests such an interpretation is unlikely. Further, similar to special education eligibility, a considerable portion of students (>18%) of students were unable to be tracked across years. Finally, qualitative data were collected on control group student experiences, and such data indicated that in several schools control students participated in other interventions. The MAVRIC intervention could not ethically require schools to not provide other interventions—which was also not desired given an interest in testing its effects in terms of practical "effectiveness"—but such other support may have contributed to the growth of control students.

YEAR 4 EVALUATION RESULTS: CONCLUSION AND NEXT STEPS

The Year 4 evaluation of the MAVIC program was designed to provide confirmatory impact evidence with regard to the program effects on student vocabulary skills. The results indicate that MAVRIC does not improve student vocabulary outcomes. However, given the breadth of the evaluation—including implementation components and various methodological considerations—several positive directions for future work (and evaluation) with MAVRIC were noted. In particular, the limitations with respect to measurement, time, and implementation factors suggest actionable changes that could result in positive effects for future evaluations. The most notable of these include (a) identifying defensible, proximal vocabulary assessments that could measure differential growth between MAVRIC participants and non-participants; (b) a need to potentially provide additional time in intervention; and (c) making changes to group size and other intervention delivery components (e.g., tutor training; material refinements). Addressing these changes will be the focus of work in Year 5 of the MAVRIC project.

Ideally, these changes will be pilot tested with respect to identifying a preliminary level of evidence for a positive impact on student outcomes. Doing so is not only consistent with the broader mission of SIF, which is to identify and scale effective programs for solving social problems, it is aligned with the goals of ServeMinnesota as a subgrantee. If ServeMinnesota is to tangibly contribute to improving student vocabulary outcomes via MAVRIC, it is essential to understand its potential impact. Doing so will help understand the future role MAVRIC plays in ServeMinnesota's broader portfolio of programs that improve social issues.

REFERENCES

Beck, I. L., & McKeown, M. G. (2007). Increasing young low-income children's oral vocabulary repertoires through rich and focused instruction. *The Elementary School Journal*, *107*(3), 251-271.

Burns, K.M., Deno, S.L., & Jimerson, S.R. (2007). Toward a unified Response-to-Intervention model. In S.R. Jimerson, M.K. Burns, & A. VanDerHeyden (Eds.), *Handbook of Response to Intervention* (pp. 428-440). New York: Springer.

Coyne, M. D., McCoach, D. B., & Kapp, S. (2007). Vocabulary intervention for kindergarten students: Comparing extended instruction with embedded instruction and incidental exposure. *Learning Disabilities Quarterly, 30*, 70–88.

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental psychology*, *33*(6), 934.

Dimmery, Drew. (2013). rdd: Regression Discontinuity Estimation. R package version 0.56. http://CRAN.R-project.org/package=rdd.

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional children*, *57*(6), 488-500.

Graves. M. & Sales G., C. (2009). *The first 4,000 words.* Minneapolis, MN: Seward Inc.

Greenwood, C. R., Delquadri, J., & Hall, R. V. (1984). *Opportunity to respond and student academic performance.* In W. Heward, T. Heron, D. Hill, & J. Trap-Porter (Eds.), Focus on behavior analysis in education (pp. 58–88). Columbus, OH: Merrill.

Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). Using student achievement data to support instructional decision making (NCEE 2009-4067). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/publications/practiceguides/.

Hart, B., & Risley, T. R. (1995). Meaningful differences in the everyday experience of young American children. Baltimore, MD: Paul H Brookes.

Imbens, Guido, & Karthik Kalyanaraman. (2009) "Optimal Bandwidth Choice for the regression discontinuity estimator," NBER Working Paper Series. 14726. http://www.nber.org/papers/w14726.

Loftus, S., Coyne, M., McCoach, D., Zipoli, R., & Pullen, P. (2010). Effects of a supplemental vocabulary intervention on the word knowledge of kindergarten students at risk for language and literacy difficulties. *Learning Disabilities Research & Practice, 25*, 124-136.

Markovitz, C.; Hernandez, M.; Hedberg, E.; Silberglitt, B. (2014). *Impact Evaluation of the Minnesota Reading Corps K-3 Program*. NORC at the University of Chicago: Chicago, IL.

Markovitz, C.; Hernandez, M.; Hedberg, E.; Silberglitt, B. (2015). *Outcome Evaluation of the Minnesota Reading Corps PreK Program*. NORC at the University of Chicago: Chicago, IL.

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.

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.

McConnell, S. R., Priest, J. S., Davis, S. D., & McEvoy, M. A. (2002). *Best practices in measuring growth and development for preschool children.* In A. Thomas & J. Grimes (Eds.), Best Practices in School Psychology (4th ed., Vol. 2, pp. 1231 – 1246). Washington DC: National Association of School Psychologists.

Missall, K. N., & McConnell, S. R. (2004). *Technical report: Psychometric characteristics of individual growth and development indicators—Picture naming, rhyming, and alliteration.* Center for Early Education and Development. Minneapolis, MN: University of Minnesota.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (National Institute of Health Pub. No. 00-4769). Washington, DC: National Institute of Child Health and Human Development.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78*(1), 33-84.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Snow, C., Burns, M., & Griffin, P. (Eds.), (1998). *Preventing reading difficulties in young children*. Washington DC: National Academy Press

Shanahan, T.; Cunningham, A.; Escamilla, KC.; Fischel, J.; Landry, S.; Lonigan, CJ., et al. (2008). *Developing early literacy: Report of the national early literacy panel.* Washington DC: National Institute for Literacy.

Stahl, S.A., & Nagy, W. (2006). Teaching word meanings. Mahwah, NJ: Erlbaum

Thompson, T., & Massat, C. R. (2005). Experiences of violence, post-traumatic stress, academic achievement and behavior problems of urban African-American children. *Child and Adolescent Social Work Journal*, *22*, 367-393.

Vaughn, S., & Linan-Thompson, S. (2003). *Group size and time allotted to intervention: Effects for students with reading difficulties.* In B. Foorman (Ed.), Preventing and remediating reading difficulties: Bringing science to scale (pp. 299–324). Baltimore: York Press.

Wackerle-Holman, A., & Bradfield, T. (2010, October). *Developing a new set of early literacy and language IGDIs.* Kansas City, KS: Center for Response to Intervention in Early Childhood, University of Kansas. Retrieved January 14, 2012 from http://www.crtiec.org/rti_summit/2010/12-wackerlehollman-bradfield.shtml.

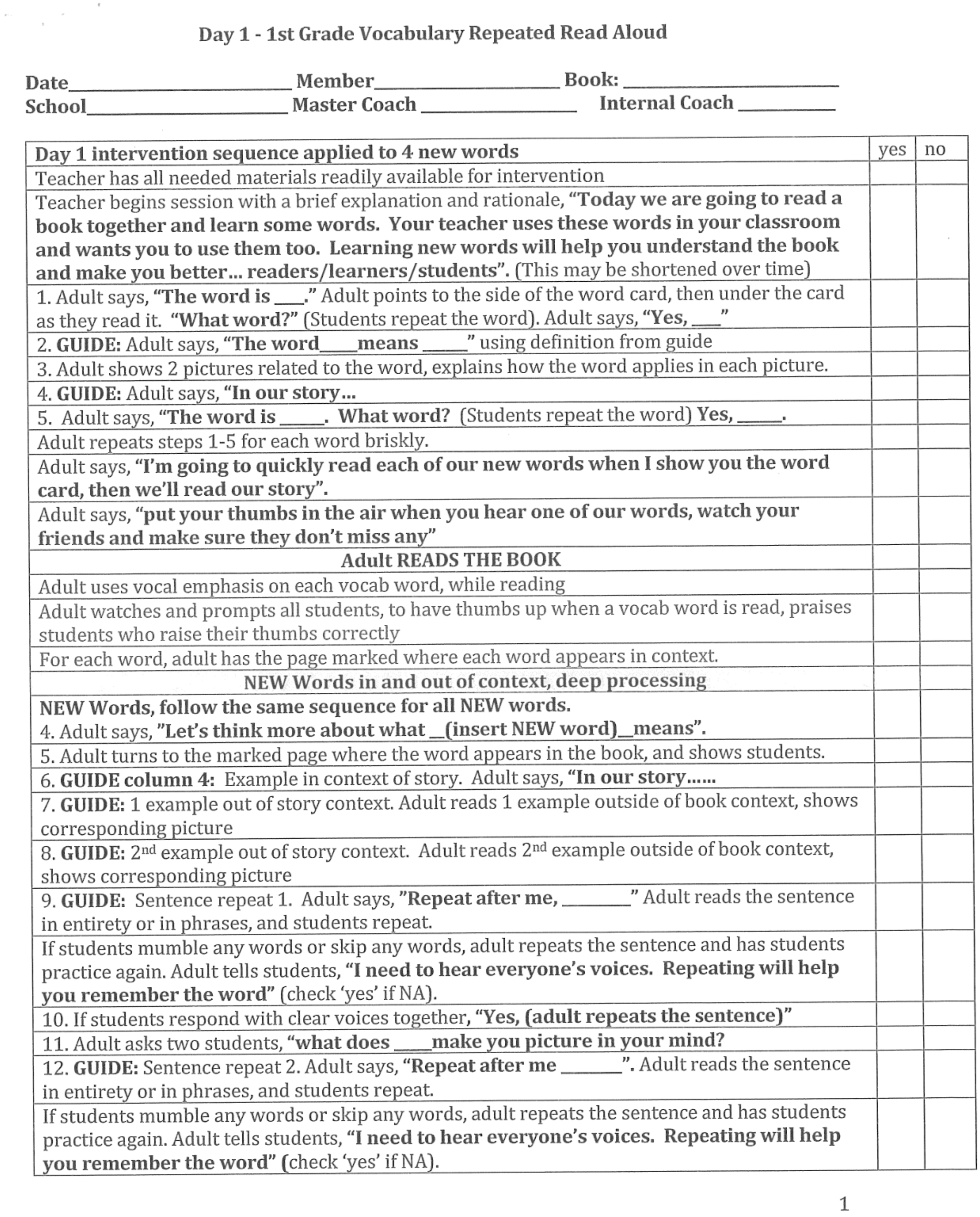
APPENDIX A—Logic Model

Inputs	Activities	Outputs	Outcomes	Impact
Training	Training & Professional Development in Reading Corps core competencies, as well as MAVRIC core functions related to vocabulary intervention and assessment	Reading Corps Members acquire requisite skills in reading intervention and assessment	Members implement core intervention and assessment skills with students who need additional vocabulary support	Student vocabulary skills improve
Coaching Model	School-based (Internal) and external (Master) coaches conduct fidelity observations of MAVRIC interventions and assessments Coaches also support MAVRIC Alignment and coordination with schools	Fidelity checklists are collected to produce data regarding implementation accuracy Coaches provide immediate feedback regarding implementation; help problem-solve implementation challenges	Member implementation of core intervention and assessment skills is maintained at a high level of accuracy and with increasing levels of technical expertise	Students receive intervention in accord with how it was empirically-tested Data are collected with high accuracy
School Partnerships	Deliver interventions and plan for successful integration of the innovations within partnership sites	Schools and Reading Corps Members collaboratively plan MAVRIC activities	MAVRIC activities are implemented at appropriate times during school day	Students are ready-to-learn when exposed to MAVRIC activities; students miss minimal other school-based learning activities.
Intervention Resources	MAVRIC task force develops intervention scripts and materials	Reading Corps Members deliver effective, efficient	Student vocabulary skills improve	Improved reading skills and school performance.

	Research-based word lists are used to identify vocabulary words for instruction	vocabulary interventions Reading Corps Members teach research-based vocabulary words		
Assessment Resources	Assessments with adequate technical characteristics are used by Reading Corps Members Members have assessments for use in identifying at-risk students (e.g., IGDIs Picture Naming, Screening; 4,000 Word Listening Test Members have assessments for use in monitoring student progress/intervention effectiveness (e.g., IGDIs Picture Naming, Progress; mastery measures of taught words)	Accurate assessment data are collected Students needing support are identified and provided support; those not needing support receive classroom instruction only Members can determine intervention effectiveness with individual students	Identification and evaluation data are accurate Vocabulary resources are used accurately and efficiently Modifications to instruction can be made in response to student learning	Trustworthy data are used for identification of students, evaluating outcomes, and for monitoring student progress Students who need support are provided support; students for whom support is unnecessary are not provided valuable resources Student vocabulary skills improve

Implementation and Impact Evaluation	Implementation data are collected for assessments and interventions Outcome data are collected before, during, and after intervention occurs	Implementation accuracy is calculated and reported for assessments and interventions Vocabulary improvement is calculated and reported for the intervention period The percentage of students who are no longer at-risk is calculated and reported	The degree of implementation accuracy for MAVRIC components is known; feedback to improve implementation is provided. Comparison between MAVRIC participants and non-participants provides evidence of effects of participation	Evidence supporting continued research and/or replicability and effectiveness of MAVRIC is identified

APPENDIX B—Sample (1st Grade) Implementation Fidelity Checklist

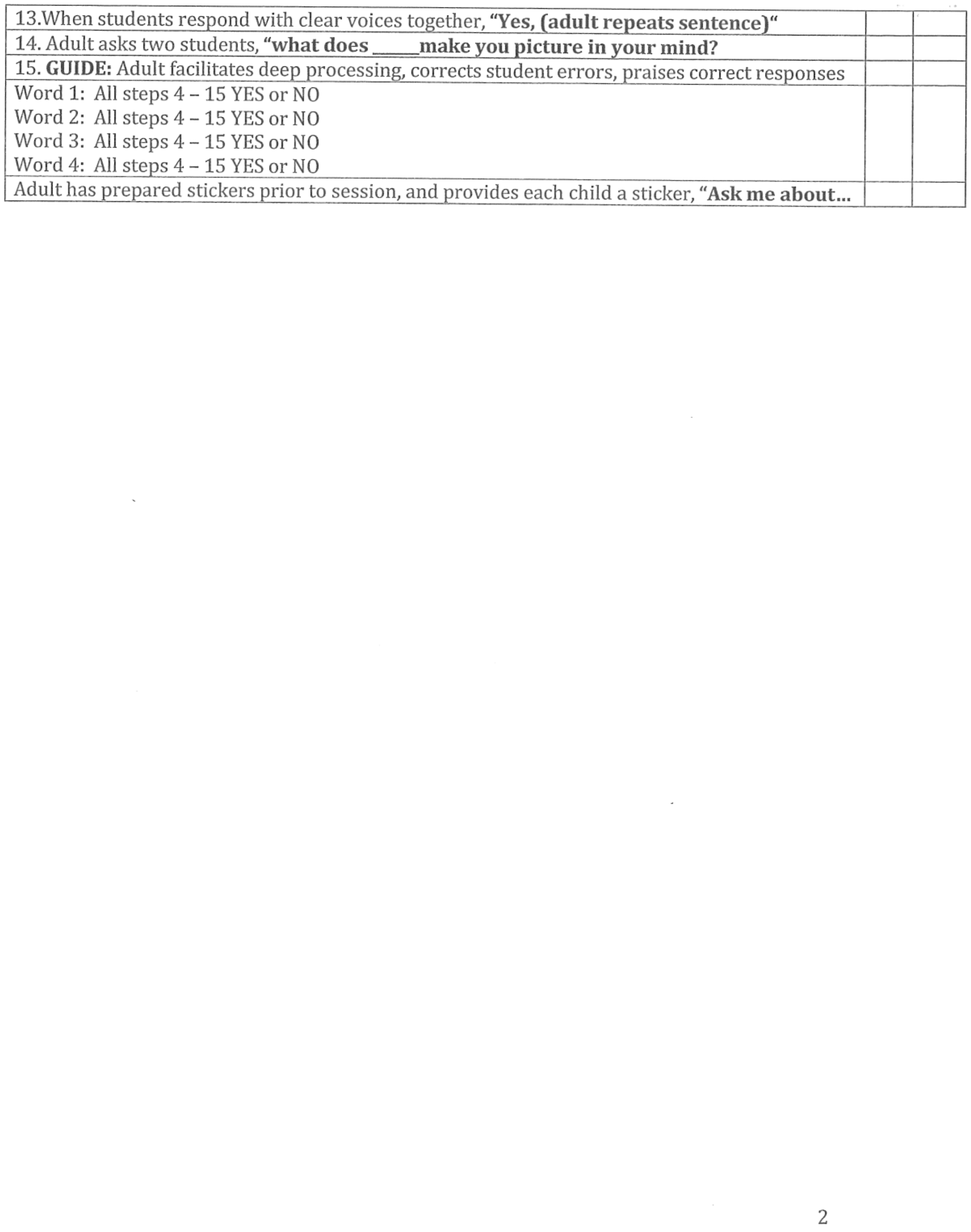


Day 1 - 1st Grade Vocabulary Repeated Read Aloud

Date______________________ Member__________________ Book: ____________________
School_____________________ Master Coach ______________ Internal Coach __________

Day 1 intervention sequence applied to 4 new words	yes	no
Teacher has all needed materials readily available for intervention		
Teacher begins session with a brief explanation and rationale, **"Today we are going to read a book together and learn some words. Your teacher uses these words in your classroom and wants you to use them too. Learning new words will help you understand the book and make you better… readers/learners/students".** (This may be shortened over time)		
1. Adult says, **"The word is ___."** Adult points to the side of the word card, then under the card as they read it. **"What word?"** (Students repeat the word). Adult says, **"Yes, ___"**		
2. **GUIDE:** Adult says, **"The word____means _____"** using definition from guide		
3. Adult shows 2 pictures related to the word, explains how the word applies in each picture.		
4. **GUIDE:** Adult says, **"In our story…**		
5. Adult says, **"The word is _____. What word?** (Students repeat the word) **Yes, _____.**		
Adult repeats steps 1-5 for each word briskly.		
Adult says, **"I'm going to quickly read each of our new words when I show you the word card, then we'll read our story".**		
Adult says, **"put your thumbs in the air when you hear one of our words, watch your friends and make sure they don't miss any"**		
Adult READS THE BOOK		
Adult uses vocal emphasis on each vocab word, while reading		
Adult watches and prompts all students, to have thumbs up when a vocab word is read, praises students who raise their thumbs correctly		
For each word, adult has the page marked where each word appears in context.		
NEW Words in and out of context, deep processing		
NEW Words, follow the same sequence for all NEW words. 4. Adult says, **"Let's think more about what _(insert NEW word)_means".**		
5. Adult turns to the marked page where the word appears in the book, and shows students.		
6. **GUIDE column 4:** Example in context of story. Adult says, **"In our story……**		
7. **GUIDE:** 1 example out of story context. Adult reads 1 example outside of book context, shows corresponding picture		
8. **GUIDE:** 2nd example out of story context. Adult reads 2nd example outside of book context, shows corresponding picture		
9. **GUIDE:** Sentence repeat 1. Adult says, **"Repeat after me, _______"** Adult reads the sentence in entirety or in phrases, and students repeat.		
If students mumble any words or skip any words, adult repeats the sentence and has students practice again. Adult tells students, **"I need to hear everyone's voices. Repeating will help you remember the word"** (check 'yes' if NA).		
10. If students respond with clear voices together**, "Yes, (adult repeats the sentence)"**		
11. Adult asks two students, **"what does ____make you picture in your mind?**		
12. **GUIDE:** Sentence repeat 2. Adult says, **"Repeat after me ______".** Adult reads the sentence in entirety or in phrases, and students repeat.		
If students mumble any words or skip any words, adult repeats the sentence and has students practice again. Adult tells students, **"I need to hear everyone's voices. Repeating will help you remember the word" (**check 'yes' if NA).		

1



13.When students respond with clear voices together, **“Yes, (adult repeats sentence)“**		
14. Adult asks two students, **“what does _____make you picture in your mind?**		
15. **GUIDE:** Adult facilitates deep processing, corrects student errors, praises correct responses		
Word 1: All steps 4 – 15 YES or NO Word 2: All steps 4 – 15 YES or NO Word 3: All steps 4 – 15 YES or NO Word 4: All steps 4 – 15 YES or NO		
Adult has prepared stickers prior to session, and provides each child a sticker, **“Ask me about…**		