USAID
FROM THE AMERICAN PEOPLE

LAC Reads
capacity program

EARLY GRADE READING IN LATIN AMERICA AND THE CARIBBEAN: A SYSTEMATIC REVIEW

LAC READS CAPACITY PROGRAM

This report is made possible by the generous support of the American people through the United States Agency for International Development (USAID). The contents are the responsibility of the American Institutes for Research and do not necessarily reflect the views of USAID or the United States Government.

Early Grade Reading in Latin America and the Caribbean: A Systematic Review

December 2016

LAC Reads Capacity Program

Contents

Tables

Figures

Boxes

Abbreviations

AIR	American Institutes for Research
BLDS	The British Library of Development Studies
CASP	Critical Appraisal Skills Programme
CI	Confidence Interval
DEC	Development Experience Clearinghouse
DfID	The U.K. Department for International Development
DOAB	Directory of Open Access Books
DOAJ	Directory of Open Access Journals
EFA	Education for All
EGR	Early Grade Reading
ELIDS	Institute of Development Studies
ICC	Intracluster Correlation Coefficient
ICT	Information and Communication Technology
IPA	Innovations for Poverty Action
IRB	Institutional Review Board
JOLIS	The Joint Libraries of the World Bank and International Monetary Fund
J-PAL	The Abdul Latif Jameel Poverty Action Lab
LAC	Latin America and the Caribbean
LLECE	Latin American Laboratory for Assessment of the Quality of Education
LRCP	LAC Reads Capacity Program
M&E	Monitoring & Evaluation
MOE	Ministry of Education
NGOs	Nongovernmental Organizations

OAS	Organization of American States
PA	Phonological Awareness
PICO	Population, Intervention, Comparison and Outcome
RAs	Research Associates
RCT	Randomized Controlled Trial
REDALYC	Red de Revistas Científicas de América Latina y el Caribe, España y Portugal
SciELO	Scientific Electronic Library Online o Biblioteca Científica Electrónica en Línea
SERCE	Second Regional Comparative and Explanatory Study
SMDs	Standardized Mean Differences
TDE	Teste do Desempenho Escolar
TERCE	Third Regional Comparative and Explanatory Study
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNHCR	The United Nations High Commissioner for Refugees
UNICEF	The United Nations Children's Fund
USAID	United States Agency for International Development
WHO	World Health Organization
3ie	The International Initiative for the Impact Evaluation

Acknowledgements

We are grateful to our colleagues on the project team, at AIR, and in the sector who contributed to this report formally or informally at various stages of the process. This included participating in planning discussions, reviewing or commenting on the systemic review technical guidelines document and/or the draft report, and providing feedback during panel presentations at the 2015 USAID Education Summit and the 2016 CIES conference. We also appreciate the numerous researchers and authors who were contacted and who provided additional details on their work.

In particular we would like to recognize the contributions of the LRCP AOR, Michael Lisman, and the entire LAC Bureau education team, as well as education team members in the USAID LAC Missions, and colleagues at Mathematica Policy Research and at the Institute of Education Science.

Executive Summary

Educational policy around early grade reading (EGR)[1] in the Latin American and Caribbean region (LAC) has long suffered from a disjuncture between school practice and research. Studies exist in the global literature on how pedagogical programs should be designed to promote gains in EGR outcomes but it is unclear whether findings in other regions can be extrapolated to the LAC region. Also, within the LAC region itself, research on EGR is fragmented and often of poor quality. There is no comprehensive or systematic overview of the EGR research literature specific to the LAC region. As a result, policy makers, pedagogy and curriculum specialists, and other stakeholders in the region are unable to determine what is relevant and are thus unable to shape policy, practice and programs in an evidence-driven manner.

This report aims to assist policy makers, international funders, nongovernmental organizations (NGOs), practitioners, researchers, and other relevant stakeholders in the LAC region by synthesizing the evidence on what works to improve reading outcomes in the LAC region. We address several research questions through our systematic review. First, we examine the effectiveness of various programs implemented in the LAC region that aim to improve early grade reading outcomes, including teacher training, school feeding, computer-aided instruction, programs with an emphasis on nutrition, and Information and Communication Technology (ICT) programs. Second, we assess the fidelity of implementation of programs that aim to improve reading outcomes. Third, we examine the factors that predict early grade reading outcomes. Fourth, we examine the experiences and perspectives of various stakeholders about early grade literacy in the LAC region. For this purpose, we use a mixed-methods systematic review, in which we synthesize the evidence from both quantitative and qualitative research.

It is important to differentiate research results on the basis of the quality of the methodology so that policy makers can make decisions that are based on valid findings. To that end, we reviewed

[1] For the purposes of this study, early grade reading (EGR) is defined as embracing the period from birth to end of Grade 3.

and appraised the quality of all of the different methodological approaches used by the evaluations. We used separate quality appraisal tools and methods of synthesis for each type of research study because we recognized the importance of choosing the right quality criteria for different types of research. The quality appraisal for each research type enables us to highlight the results of high-quality studies.

Policy makers and practitioners need guidance in order to make use of evidence that is voluminous, diverse, and fragmented across disciplines. For research to be relevant to policy, it must be captured and consolidated in a reliable and accessible manner. As Mark Lipsey noted in 1999, "Practice and policy, therefore, are best guided by a cumulation of research evidence sufficient to balance the idiosyncrasies of individual studies and support more robust conclusions than any single study can provide." Ten years later, Lipsey (2009) elaborated further, "The most useful guidance for practitioners, and the most informative perspective for program developers and researchers, will not come from lists of the names of programs shown by research to have positive effects. Rather, they will come from identification of the factors that characterize the most effective programs and the general principles that characterize what works" (p. 126).

Although systematic reviews and meta-analyses were originally conducted in the U.S. social sciences, they are equally relevant for donors of development programs and policy makers in low- and middle-income countries (Glass, 1976; Waddington et al., 2012). Recent systematic reviews have addressed several important questions about the effectiveness of development programs in low- and middle-income countries (Brody et al., 2015; Duvendack et al., 2011; Evans & Popova, 2015; Spier et al., 2016). These systematic reviews allow for a synthesis of "all the existing high-quality evidence using transparent methods to give the best possible generalized statements about what is known" (Waddington et al., 2012, p. 360).

In this systematic review, we use quality criteria and synthesis methods that are aligned with the research questions about early grade reading in Latin America and the Caribbean. First, we use an adapted version of a risk of bias (RoB) assessment tool developed by Hombrados and Waddington (2012) and a combination of meta-analyses and narrative reviews to examine the effects of different types of education- and non-education-focused programs on early grade reading outcomes. This risk of bias assessment tool uses quality criteria that enable us to determine the quality of experimental and quasi-experimental studies. Second, we use a narrative review of quantitative studies that focus on the mechanisms underlying changes in early grade reading outcomes. For these studies we rely on an adapted version of the tool created by Hombrados and Waddington (2012) that focuses specifically on quantitative studies that aim to determine the predictors of reading outcomes. Third, we use a narrative review to synthesize the qualitative evidence with a focus on reading in Latin America and the Caribbean. To determine the quality of qualitative studies, we use an adapted version of the Critical Appraisal Skills Programme (CASP) Qualitative Research Checklist.

It is important to obtain a comprehensive review of the evidence on early grade reading in the LAC region, but achieving comprehensiveness is not always straightforward. For example, Evans and Popova (2015) demonstrate that differences in search protocols across systematic reviews may result in different conclusions about the impact of education programs in international development. Although systematic literature reviews typically seek to cover as many relevant data sources as possible, databases that are generally used for systematic reviews never represent the

universe of knowledge. To maximize the comprehensiveness of the systematic review, the search strategy needs to be as broad as possible to retrieve as many potentially relevant items as are available (Schuelke-Leech, Barry, Muratori, & Yurkovich, 2015). In this systematic review, we use both well-established approaches, such as searching academic databases and the grey literature, and novel computational approaches, such as WikiLabeling, to maximize the comprehensiveness of the systematic review.

As recommended by Waddington et al. (2012) in their toolkit for systematic reviews of effects in international development, we first analyzed which populations, interventions, comparators, and outcomes are relevant to early grade reading outcomes in Latin America. Our inclusion criteria for the review were based on these factors. We only included literature that is relevant for the literacy of children in early grades in the LAC region. This literature included both studies with an emphasis on education and studies with a focus on enabling factors that are linked to education programs or reading outcomes. For example, we included studies that mostly focused on nutrition that may indirectly influence reading outcomes if those studies also included an outcome measure related to early grade reading. We developed a conceptual framework to identify these enabling factors.

We included 108 studies with a focus on early grade reading outcomes in the LAC region. We retrieved these studies after a comprehensive online search of databases, journals, and other websites with a focus on international development. This search initially resulted in 9,696 articles. We then used novel computational techniques, specifically Wikilabeling, and a manual review of the abstracts against our inclusion criteria. Following this review we were left with a total of 162 studies that went for full text review. During this phase, an additional 54 articles were removed as not relevant, resulting in 108 studies included in the final review.[2]

The 108 included articles were comprised of quantitative intervention research, quantitative nonintervention research, qualitative intervention research, and qualitative research. We included 23 experimental and quasi-experimental studies with a focus on the effects of specific development programs on early grade reading outcomes, 62 quantitative studies that had an emphasis on early grade reading outcomes but did not emphasize a specific intervention, 16 qualitative studies without a focus on a specific intervention, and 8 qualitative studies that focused on a specific intervention.

The vast majority of studies included in our review of evidence were published journal articles and came from either North or South America; significantly fewer articles were from Central America and the Caribbean. Most articles were published in English or Spanish. We found no articles in any regional languages, which may be due in large part to publication bias and availability of journals in languages that are not national languages.

More than 90% of the articles were focused on high- to upper-middle-income countries. The disproportionate emphasis on high-income and upper-middle-income countries can be explained by the limited available resources and capacity for conducting high-quality research in low-income and lower-middle-income countries.

[2] One mixed-methods study was counted twice as it was reviewed by both qualitative and quantitative reviewers.

An analysis of the quantitative intervention studies indicates that impact evaluations with an emphasis on early grade reading outcomes only focus on a small portion of the intervention types that can influence early grade reading outcomes. We only found three topic areas with more than two impact evaluations that focus on early grade reading outcomes: (1) teacher training, (2) nutrition interventions, and (3) ICT programs.

Although the majority of the included impact evaluations used a randomized controlled trial (RCT), only eight of the studies were rated as having a low risk of selection bias. Of the eight studies with a low risk of selection bias, two focus on child nutrition, three focus on ICT, one focuses on parental and community participation, one focuses on teacher practices for reading, and two focus on teacher training. These data show that there is little strong evidence regarding the impact of development programs on early grade reading outcomes.

Nonetheless, the diverse set of studies enabled the team to address a wide range of research questions:

1. What are the existing intervention- and nonintervention-based studies and what is the existing literature from or on the LAC region involving reading programs, practices, policies, and products focused on improving reading skills for children from birth through Grade 3?
2. What is the quality of the existing EGR evidence (quantitative intervention and nonintervention and qualitative intervention and nonintervention) in the LAC region and what is its practical use for varied LAC region stakeholders?
3. What are the gaps in the evidence base on EGR in the LAC region as compared to what we know globally about best practices in EGR?
4. What is the impact of reading programs, practices, policies, and products aimed at improving the reading skills of children from birth through Grade 3 on reading outcomes in the LAC region?
5. What strategies have been successful and what is the evidence for this success? Which strategies were unsuccessful and why?
6. What are examples of effectively using evidence/knowledge to shape and/or improve EGR policy and practice in the LAC region?

To address these research questions, we relied on a broad conceptual framework that examines how various factors can influence early grade reading outcomes in the LAC region. This conceptual framework explains how programs or initiatives can contribute to improving early grade reading outcomes in a sustainable manner. We also consider mechanisms that may influence how stakeholders interact with programs or practices as well as external or contextual factors that influence implementation and the linkages in the conceptual framework.

Although we only found a very limited number of high-quality quantitative intervention studies, they did indicate examples of development programs that are likely to have positive effects on early grade reading outcomes in specific circumstances and contexts. Specifically, we found evidence that teacher training programs can positively affect early grade reading outcomes in high-income economies when they are well implemented and complemented by the sustained coaching

of teachers. In addition, we found some evidence that nutrition programs can have positive effects on early grade reading outcomes in contexts where stunting and wasting are high, such as Guatemala. However, we also found evidence indicating that the distribution of laptops to children can have adverse effects on early grade reading outcomes, particularly when the distribution of laptops is not complemented by additional programs.

The findings of the quantitative nonintervention studies indicate that phonemic awareness, phonics, fluency, and comprehension are associated with reading ability. The research also indicates that poverty and child labor are negatively correlated with early grade reading outcomes. This finding on the importance of poverty and socioeconomic factors for early grade reading outcomes supports the quantitative intervention result that nutrition programs may be effective in improving early grade reading outcomes. Finally, the quantitative nonintervention studies show that the quality of preschool is positively associated with early grade reading outcomes. Triangulating this result with the quantitative findings on the impact of teacher training suggests that teacher training combined with sustained coaching could possibly positively affect early grade reading outcomes through its influence on the quality of preschool.

Both qualitative and quantitative studies indicated that consideration of context is key to improving reading outcomes. This lends credence to the conceptual framework, which suggests that enabling factors and assumptions in part determine the potential for success of various programs or strategies. The most frequently discussed topic in qualitative nonintervention articles is the need to promote social learning to improve early grade reading.

However, we found strong evidence for publication bias in the studies that focus on the effects of teacher practices and parental involvement on early grade reading outcomes in the LAC region; that is, there are likely to be a large number of additional studies that have not been published on similar topics because they did not find statistically significant effects. Findings from statistically unsuccessful interventions are also important, and publishing only the results of programs that show positive and statistically significant effects on early grade reading outcomes impedes policy makers' ability to make evidence-based decisions.

Recommendations

The primary end goal of all activities within the LAC Reads Capacity Program is to enhance the capacity of key stakeholders (e.g., the Ministry of Education and the government, international funders and intergovernmental entities, international NGOs, academics, and researchers and practitioners) to use evidence to choose, develop, implement, and evaluate early grade reading (EGR) strategies, programs, practices, and interventions. Through this systematic review of the EGR evidence from the LAC region we identified multiple gaps in the evidence-base. These gaps indicate that key stakeholders face significant challenges when attempting to make evidence-based decisions.

Next we present our recommendations based on the evidence gaps to highlight specific areas in need of further funding or research. This research could help to fill in the evidence gaps and provide more robust and comprehensive evidence on what works in early grade reading in the LAC region.

Recommendations Based on the Evidence Gaps:

- Ensure that language assessments include multiple reading constructs and differentiate between those constructs so it is easier to identify the effects of interventions on individual constructs.

- Fund long-term mixed-methods experimental or quasi-experimental research on the effects of preschool and early childhood education on early grade reading outcomes.

- Include several early grade reading constructs in administrative data to enable researchers to conduct high-quality research on the mechanisms underlying early grade reading using large sample sizes.

- Document ongoing research to minimize publication bias so that unpublished research is available to policy makers as well and to ensure that hypotheses are pre-specified.

- Register ongoing research on early grade reading in a central, publicly available location so that everyone can see what is being done and seek to complement and add to the research base.

- Develop more interdisciplinary mixed-methods research on early grade reading that includes more than one reading construct and large sample sizes.

- Fund rigorous research that allows for an examination of the causal effects of specific development programs on early grade reading outcomes. These studies include both experimental and quasi-experimental studies with a sufficient sample size. These studies also need to be supplemented with qualitative research.

- Pursue more research on EGR strategies for students with disabilities.

- Pursue more research on reading in indigenous languages.

- Conduct more research on the linkages between the development of prewriting and writing skills and early grade reading outcomes.

In addition to reviewing the EGR evidence from the LAC region, the LAC Reads Capacity Program also collects and catalogues EGR pedagogical resources (e.g., supplementary reading materials, assessments, instructional materials, videos), and other EGR documents that are neither research-based evidence nor resources (e.g., policy documents, project reports, best practices documents) from the LAC region. These resources can serve as additional support for stakeholders to improve their practice. In order to support stakeholders to improve their practice, the program is developing a resource database making these resources available to a wide audience through the program's website at http://www.lacreads.org. We are also conducting a stakeholder analysis in the region with our local partner organizations to identify key EGR stakeholders, determine their interests and needs, and how the evidence from this review and the resources collected can best be used to support EGR capacity and achievements in the region.

Introduction

In this section we provide the rationale for conducting this systematic review, introduce the concept of systematic reviews and specifically systematic reviews in international development, define our operational terms, provide some background on literacy in the LAC region, and outline our conceptual framework and research questions.

Rationale for This Review

Globally, there is considerable evidence on what predicts successful literacy acquisition, effective reading interventions, and best practices for fostering fluent reading in different contexts. For instance, the literature suggests that it is important to focus early grade instruction and assessments on key reading subskills, such as phonemic awareness, vocabulary knowledge, and reading comprehension (Adams, 1990; Snow, Burns, & Griffin, 1998). Time spent on reading tasks and access to print materials is significantly related to early reading development (Cipielewski & Stanovich, 1992; Cunningham & Stanovich, 1991). The quality of the home and community culture of reading is also positively associated with literacy outcomes (Foy & Mann, 2003; Sénéchal & LeFevre, 2002). In addition, research suggests that children's literacy develops best if they first learn to read a language they already understand, and that first language skills transfer and support second language reading in significant and predictable ways (August & Shanahan, 2006; Cummins, 1979; Koda & Reddy, 2008; Nakamura & De Hoop, 2014). Children's literacy learning is also influenced by enabling factors inside and outside the education sector, such as teacher attendance, state capacity, and the nutrition of children.

Critical gaps in the literature and challenges remain, however, both outside and inside the LAC region. For example, the majority of the findings discussed earlier are based on findings outside the LAC region, and it is unclear whether these findings can be extrapolated to the LAC region. Second, the majority of the evidence on reading, both inside and outside the LAC region, is based on correlations and does not allow for causal claims about the impact of development programs on reading outcomes. Third, the existing evidence on early reading acquisition in LAC-relevant contexts is neither synthesized nor readily accessible. Fourth, it is not easy for policy makers, practitioners, and other decision makers to ascertain the quality of the existing studies on literacy in the LAC region. These four factors limit the possibility of evidence-informed policy making. The USAID-funded LAC Reads Capacity Program (LRCP) commissioned this systematic review to help resolve these challenges.

This systematic review intends to inform the work of the LCRP, which is led by the American Institutes for Research (AIR) in partnership with Juárez and Associates. The LCRP aims to increase the impact, scale, and sustainability of early grade reading interventions in the LAC region through the development of state-of-the-art knowledge resources and the provision of technical assistance (TA) to governments in the LAC region and other selected key stakeholders. Such resources and TA should enable and enhance the efforts of governments in the LAC region to boost early grade reading outcomes, particularly for disadvantaged children. A key aim of the LAC Reads Capacity Program is to facilitate stakeholder understanding and application of evidence-based, context-appropriate approaches to improving early grade reading.

The LCRP initiated this systematic review of the evidence around early grade reading from or on the LAC region in order to

- Provide a summary of the evidence on improving early grade reading skills in the LAC region over the last 25 years (since Education for All);
- Organize, categorize, and conduct a quality review of the existing evidence to help users of this review make more informed decisions based on the quality of the evidence provided;
- Identify gaps in the existing LAC early grade reading evidence, both to inform our audiences about them and to encourage researchers in the LAC region to continue adding to the knowledge base for EGR practices in the region; and
- Synthesize the existing EGR literature from the LAC region and package it in accessible ways for our different stakeholder groups (ministries of education, USAID missions, NGOs, academic institutions, teacher training institutions, teachers unions, private foundations, etc.).

Operational Definitions

This section provides the operational definitions of key terms used in this review.

Early grade reading is defined by USAID as pertaining to Grades 1–3 of primary schooling. Our systematic review focused on students in Grades 1–3, but we also broadened the definition to include children from birth, as there is a large evidence base on the importance of developing early receptive and expressive language skills, exposure to print, and prereading and writing activities for improving later reading success. We included any children from birth through Grade 3 regardless of age as we are aware that in many countries, late entry, low internal efficiency, and grade retention policies cause significant overage problems.[3]

Evidence refers to a research or empirically derived body of facts that can be used to make informed decisions about education interventions (e.g., policies, practices, or programs).

LAC region includes all countries in the Latin America and Caribbean region, as follows: Antigua and Barbuda, Argentina, Aruba, Bahamas, Barbados, Belize, Bermuda, Bolivia, Brazil, British Virgin Islands, Cayman Islands, Chile, Colombia, Costa Rica, Cuba, Curacao, Dominica, Dominican Republic, Ecuador, El Salvador, French Guiana, Grenada, Guadeloupe, Guatemala, Guyana, Haiti, Honduras, Jamaica, Martinique, Mexico, Mont Serrat, Netherlands Antilles, Nicaragua, Panama, Paraguay, Peru, Puerto Rico, Saint Barthelemy, Saint Kitts and Nevis, Saint Lucia, Saint-Martin, Sint Maarten, Saint Vincent and the Grenadines, Suriname, Trinidad and Tobago, Turks and Caicos Islands, Uruguay, U.S. Virgin Islands, and Venezuela.

[3] We also included studies that present early grade reading outcomes and reading outcomes for other students. These studies differentiated the results for early grade readers and other students in some but not all cases. To mitigate this concern, we contacted the authors of primary studies to obtain information about separate effect sizes for early grade readers and other students. We incorporated this information in the synthesis of the evidence when we were able to. However, in some cases authors of primary studies either did not respond or were not able to provide all information we requested. In those cases we had to assume that the effects of programs on reading outcomes were the same for early grade readers and other students.

Meta-analysis is "the statistical pooling of information on study effect sizes" (Waddington et al., 2012) to determine the impact of programs on specific outcomes. Meta-analysis enables researchers to estimate the average effect size of specific programs on early grade reading outcomes and to assess the variation in effect sizes across different contexts. The advantage of meta-analysis over narrative synthesis is that it enables researchers to increase their sample size by pooling studies together, which increases statistical power or the ability to find small but meaningful effects with sufficient precision (Waddington et al., 2012). However, in some instances it might not be appropriate to conduct meta-analyses because studies are insufficiently comparable (Snilstveit et al., 2012). In those cases we applied a narrative synthesis.

Narrative synthesis is an "approach to the synthesis of evidence relevant to a wide range of questions including but not restricted to effectiveness [that] relies primarily on the use of words and text to summarize and explain—to 'tell the story'—of the findings of multiple studies (Popay et al., 2006)."

PICO criteria stands for population, intervention, comparison, and outcome. The Campbell Collaboration and Cochrane, the foremost research networks promoting best practices in systematic reviews worldwide, recommend using the PICO categories for formulating questions and search strategies for systematic reviews.

Systematic review is a review of the evidence around a particular topic that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to extract and analyze data from the studies that are included in the review.

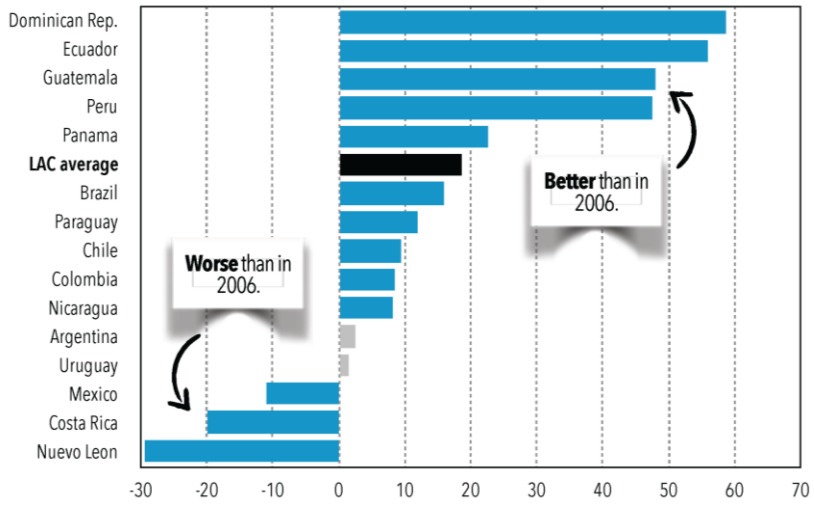
Background on Literacy and Evidence-Informed Policy in the LAC Region

The LAC region has experienced some positive trends in primary education and early grade reading and literacy outcomes in the last decade. Early grade reading outcomes for third grade students in the LAC region increased in the large majority of the countries as shown in Figure 1. This improvement was accompanied by a reduction in the rural-urban gap in reading proficiency. The Second Regional Comparative and Explanatory Study (SERCE) and the Third Regional Comparative and Explanatory Study (TERCE) are regional assessments that are comparable over time and conducted by the Latin American Laboratory for the Assessment of the Quality of Education (LLECE). The tests evaluate third and sixth graders in reading, math, and science but for the purposes of this report, we only refer to the reading scores from the third grade tests. The third grade reading results from the 2006 SERCE and the 2013 TERCE in urban and rural regions in each participating country show a significant reduction in the gaps in reading scores in the majority of countries with the exclusion of Colombia, Nicaragua, and the Dominican Republic (UNESCO, 2014, p.8). In addition, the pupil-teacher ratios (PTRs) in almost all 29 countries in the LAC region with data decreased from 26:1 in 1999 to 23:1 in 2012 (UNESCO, 2014, p. 8). Furthermore, the percentage of trained teachers for most countries in the region increased significantly (UNESCO, 2014, p. 8).

Synthesizing the evidence across different contexts is critical because it is difficult to speak in general terms about the state of early grade reading in the LAC region. The LAC region is composed of more than 40 countries and territories on two continents with five different official

languages (English, Spanish, French, Dutch, and Portuguese) and many more regional languages. Figure 1 shows that even though most countries saw improvements in early grade reading outcomes, students in Mexico, Costa Rica, and Nuevo Leon actually performed worse. The improvements in early grade reading outcomes in the LAC region are thus not universal.

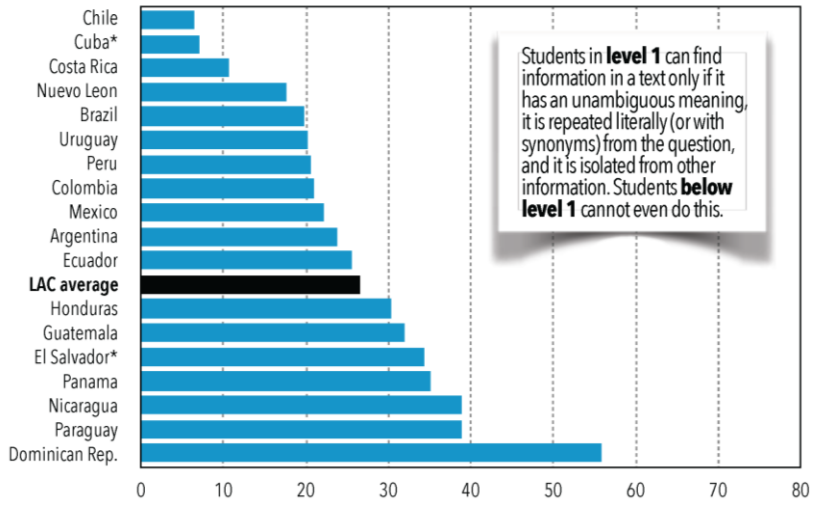
Figure 1. Change in Mean Scores in Third Grade Reading, 2006–2013



Source: From *Are Latin American Children's Reading Skills Improving? Highlights of the Second and Third Regional Comparative and Explanatory Studies (SERCE & TERCE).* Washington, DC: American Institutes for Research; p. 15. Reprinted with permission. **Notes:** (1) Only changes shown in blue or black are statistically significant. (2) Honduras did not participate in 2006 and Cuba did not participate in 2013, so they were excluded from this graph. (3) The mean score for the region includes all countries in this graph with equal weights.

There are various other challenges with early grade reading in the LAC region. First, there are still great disparities among the poor, rural, indigenous, and other disadvantaged groups in the region. Second, despite the improvements in reading outcomes for third graders, one in four third graders performed so poorly that they were categorized in the lowest level of the reading test, and less than 5% of the third graders performed so well that they were categorized as achieving the highest levels of reading. Figure 2 depicts these challenges by demonstrating that there are still a significant number of third graders scoring at the lowest levels of reading. In fact, more than 60% of third grade students have only achieved basic reading skills (Levels 1 and 2).

Figure 2. Percentage of Third Graders Scoring at Level 1 or Below on Reading, 2013

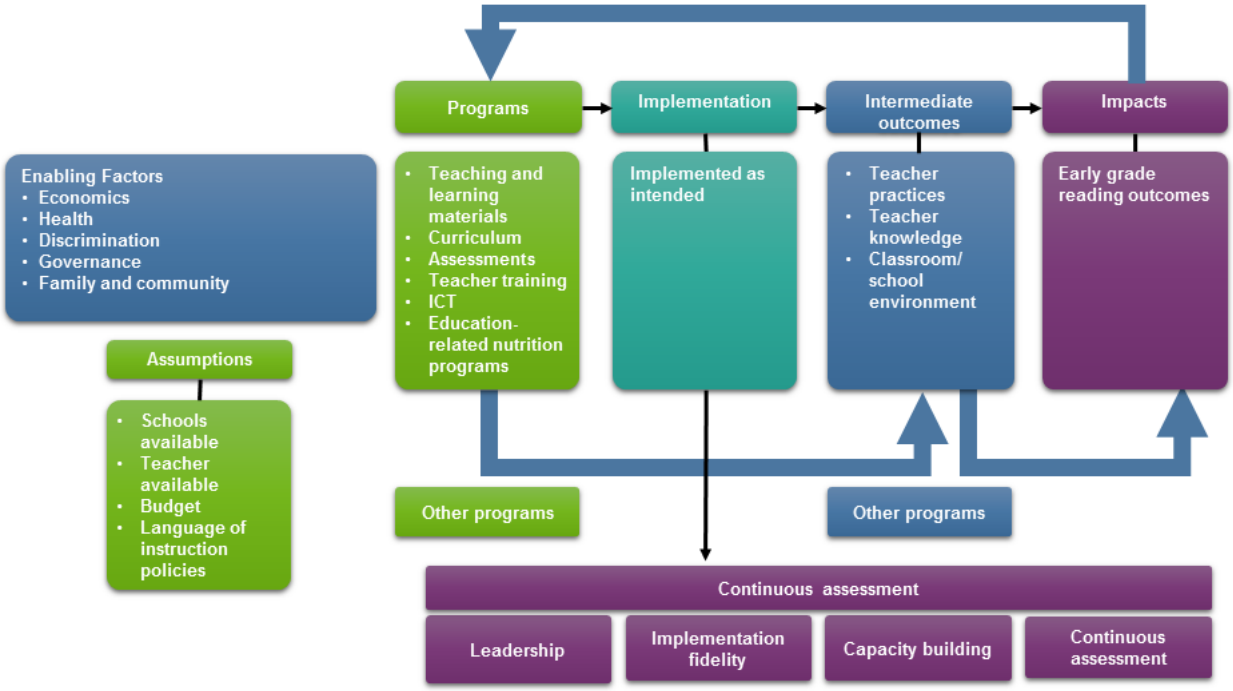


Source: From *Are Latin American Children's Reading Skills Improving? Highlights of the Second and Third Regional Comparative and Explanatory Studies (SERCE & TERCE)*. Washington, DC: American Institutes for Research; p. 19. Reprinted with permission. **Notes:** (1) Lowest levels include level 1 and below. (2) The mean score for the region includes all countries except for Cuba, El Salvador, and Honduras with equal weights. (3) Cuba's and El Salvador's scores are from 2006.

Evidence-informed early grade reading policy can contribute to mitigating some of the concerns associated with early grade reading outcomes in the LAC region, and some trends in the region suggest that there is potential for a move towards more evidence-informed policy. Positive changes in the enabling environment can for example contribute to evidence-informed policy. These changes include a positive trend in economic growth and the use of findings of impact evaluations in the LAC region. Economic growth allows policy makers to distribute resources to more effective programs. Furthermore, the previous use of findings from rigorous impact evaluations of cash transfer programs in the LAC region demonstrates the potential for using evidence to inform policy decisions. For example, impact evaluations of cash transfer programs in Mexico, Brazil, Colombia, and Argentina contributed to the scaling up of these programs (e.g. Agosto et al., 2012; Langou & Forteza, 2012). However, up until now, education policies to improve reading outcomes have only been informed by evidence to a limited extent. The limited use of evidence to inform education policies may be associated with the major differences across countries in the LAC region. These differences limit the possibilities to learn about experiences from other contexts. This systematic review will help the LAC Reads Capacity Program to contribute to evidence-informed policy by synthesizing evidence across different contexts. The systematic review will enable us to examine contextual factors that constrain or enable early grade reading outcomes in addition to an analysis of what works to improve early grade reading outcomes.

To synthesize the evidence, it is important to use a theory-based approach. Therefore, we developed a conceptual framework (Figure 3) that examines how various factors can influence early grade reading outcomes in the LAC region. The team describes this conceptual framework below.

Figure 3. Conceptual Framework



The team believes that EGR practice- and policy-relevant research should be built on a conceptual framework that maps out the linkages across enabling factors, education- and noneducation-related programs or initiatives that are associated with literacy, intermediate outcomes, and reading outcomes, as well as the assumptions that underlie this framework. This conceptual framework explains how programs or initiatives can contribute to improving early grade reading outcomes in a sustainable manner.

We also consider mechanisms that may influence how stakeholders interact with programs or practices, as well as external or contextual factors that influence implementation and the linkages in the conceptual framework. Importantly, the linkages in the conceptual framework can be moderated by the enabling environment. This enabling environment consists of the institutions and other contextual characteristics that need to be in place to enable the implementation of successful programs that are effective in improving early grade reading outcomes. For example, teacher training programs are likely to be more effective in an environment with a sufficient number of qualified teachers with the incentive to attend school. Similarly, teaching students how to read is likely to be more effective in an environment in which students are not stunted or wasted. Finally, a strong governance structure sets the stage for high-quality education by ensuring that schools and teachers are available and have a budget within which they can implement programs or practices.

Our conceptual framework begins with the enabling factors and assumptions that are necessary for any intervention or program to be able to impact EGR outcomes in LAC. As discussed above, these factors refer to assumptions that need to be in place to enable successful programs that are effective in improving reading outcomes. Then, education programs are implemented along with other noneducation programs that may have complementary or indirect effects or moderate effects on the education programs. Successful implementation enables the achievement of intermediate

outcomes, such as changes in teacher knowledge and practices, which can in turn improve EGR. Finally, we also include key elements for sustainability—namely, leadership, implementation fidelity, capacity building, and continuous assessment—that enable implementation to continue producing outcomes and impacts. Sustainability also depends on overcoming potential barriers, including financing, motivation at the community level, turnover in the government, and prioritization of these goals among competing initiatives.

Measurement, learning, and evaluation can help sustain programs through producing data that inform policy makers, address program weaknesses, and improve implementation. Ultimately, appropriate use of data could enhance the quality of programs or practices and the fidelity of their implementation, thus improving EGR. We utilize this conceptual framework as we assess how the available data on EGR in LAC contributes to the process.

Systematic Reviews in Development

This section describes the characteristics and significance of international development systematic reviews, different methodological approaches, and our rationale for adopting a broad review approach. Although systematic reviews were originally conducted in the U.S. social sciences, they are equally relevant for donors of development programs, policy makers, and other stakeholders in low and middle-income countries (Glass, 1976; Waddington et al., 2012). Recent systematic reviews have addressed important questions about the effectiveness of development programs through systematic reviews (Brody et al., 2015; Duvendack et al., 2011; Evans & Popova, 2015; Spier et al., 2016). These systematic reviews allow for a synthesis of "all the existing high-quality evidence using transparent methods to give the best possible generalized statements about what is known" (Waddington et al., 2012, p. 360).

Systematic reviews in the field of medical research lend themselves to specific and focused research questions such as "Does vitamin C reduce the incidence, duration or severity of the common cold when used either as a continuous regular supplementation every day or as a therapy at the onset of cold symptoms?" Reviews in international development, however, often tend to be broader, partly because there is less high-quality intervention research upon which to rely but also because contextual factors play a major role in determining which programs are most effective in early grade reading outcomes and under which conditions (Waddington et al., 2012). In our case, the research question of interest is also much broader as we are looking to identify all existing EGR evidence from the LAC region and are not focused on a particular type of intervention.

There are two distinct camps when it comes to broadly or narrowly defining the research question for a systematic review as described by Waddington et al. (2012) in their paper titled "How to Do a Good Systematic Review of Effects in International Development: A Tool Kit." These two camps are described as "splitters" and "lumpers." "Splitters" contend that only studies that are similar on the PICO criteria (population, intervention, comparison, and outcome) should be compared. "Lumpers" argue that broader reviews promote policy relevance because they compare a range of interventions focused on a common goal (e.g., early grade reading improvement), which enables policy makers to select the most effective intervention or evidence relevant to their context. Broadening the scope of a review also enables researchers to assess generalizability across a wider range of contexts, study populations, and behaviors (Grimshaw et al., 2003; Shadish, 2002,) and to address a wider range of research questions that are relevant for policy and practice. For

example, the review could better understand the mechanisms underlying changes in reading outcomes by including qualitative studies or quantitative studies that do not focus on a specific program (Snilstveit, Oliver, & Vojtkova, 2012). In sum, broadening the scope of the systematic review allows for a synthesis of "all the existing high-quality evidence using transparent methods to give the best possible generalized statements about what is known" (Waddington et al., 2012, p. 360). However, the majority of existing systematic reviews with an emphasis on low- and middle-income countries only focus on experimental and quasi-experimental studies.

The focus on experimental and quasi-experimental methods in systematic reviews may be too stringent to make them relevant for stakeholders. In order to capture the variety of evidence that exists on EGR in the LAC region, we broadened our review to include (1) studies that are not focused on interventions, (2) qualitative research, and (3) less rigorous experimental and quasi-experimental studies. Reviews that are too tightly defined can result in few studies that meet the inclusion criteria which makes it difficult to draw any conclusions. "The benefits of including broader evidence are to provide more detailed information on where existing studies fall down, and where new primary studies are required" (Waddington et al., 2012). For example, Snilstveit et al. (2012) argue that systematic reviews need to include qualitative studies and use narrative review methods to enable researchers to address a wider range of research questions that are relevant for both policy and practice. Similarly, stakeholders may be interested in the mechanisms underlying changes in reading outcomes. For this purpose, we need to complement systematic reviews in international development with narrative syntheses of the evidence underlying changes in reading outcomes and what predicts these changes. This narrative synthesis requires the inclusion of qualitative studies and quantitative studies that focus on the predictors of reading outcomes in addition to the inclusion of experimental and quasi-experimental studies.

Approach

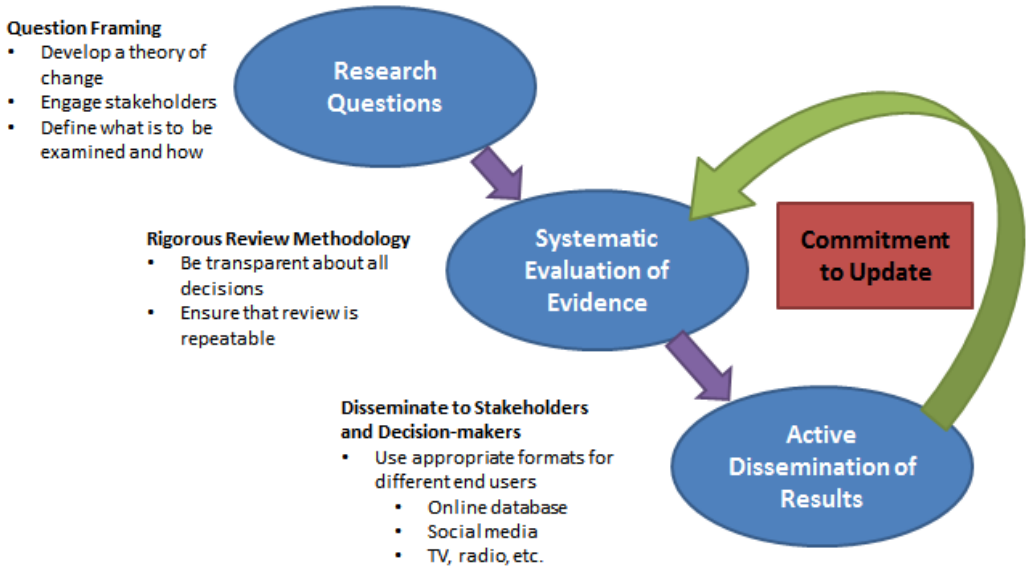
To ensure a comprehensive review of existing articles, the systematic review team used both well-established approaches, such as searching academic databases and grey literature, as well as innovative computational approaches to maximize the comprehensiveness of the search. Although systematic literature reviews typically seek to cover as many relevant data sources as possible, databases never represent the universe of knowledge. Evans and Popova (2015) demonstrate that differences in search protocols across systematic reviews may result in different conclusions about the impact of education programs in international development. To maximize the comprehensiveness of the review, we employed a search strategy that aimed to retrieve as many potentially relevant items as possible (Schuelke-Leech et al., 2015).

As recommended by Waddington et al. (2012) in their toolkit for systematic reviews of effects in international development, we based our inclusion criteria on which populations, interventions, comparators, and outcomes are relevant for early grade reading outcomes in Latin America. We only included literature that is relevant for the literacy of children in early grades in the LAC region. This literature includes both studies with an emphasis on education and studies with a focus on enabling factors that are linked to education programs or early grade reading outcomes. These studies did not necessarily have to focus on the education sector. We also included studies that focused on enabling factors or programs that can indirectly influence early grade reading outcomes, such as poverty, programs with an emphasis on nutrition, and teacher attitudes. We

developed a conceptual framework to identify the education programs and enabling factors that are potentially relevant for early grade reading outcomes in the LAC region.

The systematic review team used separate quality criteria and synthesis methods when evaluating each type of early grade reading research on LAC. First, we used an adapted version of a risk of bias assessment tool developed by Hombrados and Waddington (2012) and a combination of meta-analyses and narrative review methods to examine the effects of different types of education- and noneducation-focused programs on early grade reading outcomes. This risk of bias assessment tool uses criteria that enable us to determine the quality of experimental and quasi-experimental studies. Second, we conducted a narrative review of quantitative studies that focus on the predictors of early grade reading outcomes. We relied on an adapted version of the Hombrados and Waddington (2012) tool, which focuses specifically on quantitative studies that aim to determine the predictors of reading outcomes. Third, we used a narrative review to synthesize the qualitative evidence with a focus on reading in Latin America and the Caribbean. To determine the quality of qualitative studies, we use an adapted version of the Critical Appraisal Skills Programme (CASP) Qualitative Research Checklist. **Figure 4** represents the elements of the systematic review from developing a research question to disseminating the results.

Figure 4. Conceptual Elements of the Systematic Review



Following the systematic review, we will disseminate the research findings to key stakeholders that are identified on the basis of a comprehensive stakeholder mapping and analysis. The stakeholder analysis aims to identify and systematize key EGR organizational and individual stakeholders within the framework of the LRCP. The team will then be able to better tailor the results of the systematic review to the specific needs of various practitioners so they can both utilize existing knowledge as well as work to fill gaps in knowledge as appropriate for each context. The overarching practitioner groups include: (1) Ministries of Education and governments; (2) international funders and governmental organizations; (3) international NGOs; (4) academics and researchers; and (5) practitioners, such as teachers and teacher trainers. We present preliminary recommendations by overall stakeholder group at the end of this report.

Research Questions

To increase the policy relevance of the systematic review, this review employed the "lumping" approach by looking at a broad range of study types and designs that are all focused on early grade reading in the LAC region. The review is defined by the following research questions:

1. What are the existing intervention- and nonintervention-based studies and what is the existing literature from or on the LAC region involving reading programs, practices, policies, and products focused on improving reading skills for children from birth through Grade 3?
2. What is the quality of the existing EGR evidence (quantitative intervention and nonintervention and qualitative intervention and nonintervention) in the LAC region and what is its practical use for varied LAC region stakeholders?
3. What are the gaps in the evidence base on EGR in the LAC region as compared to what we know globally about best practices in EGR?
4. What is the impact of reading programs, practices, policies, and products aimed at improving the reading skills of children from birth through Grade 3 on reading outcomes in the LAC region?
5. What strategies have been successful and what is the evidence for this success? Which strategies were unsuccessful and why?
6. What are examples of effectively using evidence/knowledge to shape and/or improve EGR policy and practice in the LAC region?

Methods

This section describes the steps we followed to design and carry out this systematic review of early grade reading research in the LAC region. It details each phase of the process and includes descriptions of search strings, inclusion criteria, quality review protocols, synthesis and triangulation methodology, and procedures.

Systematic Review Phases

The systematic review included the following phases:

1. Establishing a conceptual framework
2. Developing the research questions
3. Determining the relevant population, intervention, comparisons, and outcomes (PICO)
4. Determining the relevant study types
5. Developing the search strategy
6. Searching for evidence
7. Extracting data from identified sources
8. WikiLabeling

9. Applying inclusion criteria and recording key indicators
10. Reviewing full text using quality review protocols
11. Analyzing data
12. Mapping the gaps in the evidence
13. Triangulating findings

In the following sections, we explain each step within these phases.

1. *Establishing a Conceptual Framework*

The research team initiated the systematic review by jointly developing a conceptual framework for the research, which shows the perceived relationship between the enabling environment, interventions, and EGR outcomes as well as the underlying assumptions. We developed the model on the basis of the existing EGR literature from the United States, Europe, and developing country contexts as well as the researchers' own knowledge about and experience with EGR and the LAC region.

2. *Developing the Research Questions*

The research questions include both descriptive questions about what EGR programs are implemented in the LAC region and more analytically oriented questions related to what works to improve reading outcomes, how these programs work, and how enabling and implementation factors influence these relationships. The combination of descriptive and analytical questions enables us to examine both the gaps in the existing literature and the knowledge we can obtain from the existing literature. We developed the research questions in consultation with USAID.

3. *Determining the PICO*

We determined the relevant population, interventions, comparisons, and outcomes on the basis of our research questions, knowledge about the LAC region, and our knowledge about experimental and quasi-experimental methods. We defined the relevant population as children in early grades in Latin America and the Caribbean. Furthermore, we decided to not restrict our sample on the basis of interventions because there are many interventions that can directly or indirectly influence early grade reading outcomes. We only determined appropriate comparisons for our synthesis of experimental and quasi-experimental studies. Other studies do not require a control or comparison group to enable a rigorous study. For the experimental and quasi-experimental studies, we included all randomized controlled trials (RCTs) and studies with multivariate analyses that included a comparison group. Finally, we included all quantitative studies that included a measure of early grade reading as an outcome variable. We did not determine appropriate outcomes for qualitative studies because high-quality qualitative studies do not require quantitative outcome measures.

4. *Determining the Relevant Study Types*

To answer our research questions, we included four study types. The first types are experimental and multivariate nonexperimental studies that include a control or comparison group. We defined these studies as "quantitative intervention studies." We included these studies to determine the

impact of specific programs on early grade reading outcomes. The second study type consists of qualitatively oriented studies with a focus on interventions. These studies usually emphasize the process of program implementation or experiences of beneficiaries about the performance of the program. We defined these studies as "qualitative intervention studies." The third type of study, quantitative studies, emphasize predictors of reading outcomes and do not focus on the effects of a specific program. We defined these studies as "quantitative nonintervention studies." We included these studies to increase our understanding of intermediate outcomes and their ability to predict reading outcomes. Fourth, we included qualitative studies that discuss literacy in the LAC region but do not include an emphasis on a specific program. We defined these studies as "qualitative non-intervention studies." We included these studies to assess the experiences and perspectives of key stakeholders, including students, teachers, and policy makers, concerning literacy and reading.

For the quantitative intervention studies, we only discuss experimental designs using random assignment to the intervention and nonexperimental designs that use multivariate regression or other multivariate analysis in this review. Our online database includes quantitative intervention studies without a control or comparison group and nonexperimental quantitative intervention studies that use univariate analyses methods. However, these studies cannot be considered sufficiently rigorous to be included in this systematic review. We include multivariate nonexperimental designs such as regression discontinuity designs, "natural experiments," and studies in which students or schools self-select into the program. To be included, the studies needed to collect cross-sectional or longitudinal data for both beneficiaries and control or comparison groups and use propensity score or other types of matching, difference-in-difference estimation, instrumental variables regression, multivariate cross-sectional or longitudinal regression analysis, or other forms of multivariate analysis, such as the Heckman selection model. The studies do not necessarily have to demonstrate baseline equivalence to be included in the review.

For the other study types, we included all studies that can be considered "empirical research." We define empirical research as studies based on data or an empirically derived body of facts that can be used to make informed decisions about education interventions (e.g., policies, practices, or programs).

5. *Developing the Search Strategy*

To develop and refine the search strategy, we relied on our PICO criteria and consultations with other researchers, librarians, computer scientists, and content experts. Through this process, we selected the most relevant databases for our review.[4] The primary requirement for selected databases—ability to search the full database—is critical to ensure that the selection process was impartial. For example, Google Scholar is a source of unpublished or "grey" literature. However, it does not provide an interface that allows a systematic search and retrieval of all potentially relevant documents. Rather, the query yields only the top results as defined by the Google search algorithm. After selecting appropriate databases, the team drafted, tested, and refined the initial search queries overall and by database specifications to identify the search string that best captured the most potentially relevant evidence for the population, topic, and time frame of interest. We searched English, Spanish, Portuguese, and Dutch language databases.[5]

[4] The full list of databases is included in the discussion under "Searching for Evidence."
[5] None of the French databases we located were specific to education in the LAC region.

The systematic review team constructed a database query by identifying search terms using the **PICO criteria** (population, intervention, comparison, and outcomes), which are the standard criteria used in the Cochrane and Campbell systematic reviews (http://handbook.cochrane.org/). The terms below represent the keywords and phrases that were identified for our English search. Their equivalents in the other target languages are not listed here but are available upon request.

- **Population:**
  - Birth to grade 3, 0-10, early childhood, pre-school, pre-primary, primary, kindergarten, grade 1, grade 2, grade 3, day care, early-grade, elementary
  - Latin America, Caribbean, Central America, South America, Antigua and Barbuda, Argentina, Aruba, Bahamas, Barbados, Belize, Bermuda, Bolivia, Brazil, British Virgin Islands, Cayman Islands, Chile, Colombia, Costa Rica, Cuba, Curacao, Dominica, Dominican Republic, Ecuador, El Salvador, French Guiana, Grenada, Guadeloupe, Guatemala, Guyana, Haiti, Honduras, Jamaica, Martinique, Mexico, Mont Serrat, Netherlands Antilles, Nicaragua, Panama, Paraguay, Peru, Puerto Rico, Saint Barthelemy, Saint Kitts and Nevis, Saint Lucia, Saint-Martin, Saint Vincent and the Grenadines, Sint Maarten, Suriname, Trinidad and Tobago, Turks and Caicos, Islands, Uruguay, US Virgin Islands, Venezuela
- **Intervention**:[6] We did not search for terms such as "program" or "intervention." We accepted all evidence-based research about literacy.
- **Comparison**: We did not search for a control or comparison group. We included all evidence-based research about literacy regardless of the use of a control or comparison group.
- **Outcomes**: We included no search terms associated with outcomes. We included any study that had an early grade literacy-related outcome. This outcome could focus on students, teachers, or parents/community.

We also included time frame (1990–2015) in the search parameters. We selected this time frame because it provided us with access to a large amount of relevant evidence; we also wanted to be more inclusive and make sure we did not leave out any important evidence. In addition, this time frame focuses on the period after the Education for All (EFA) movement and the World Conference on Education for All held in 1990 in Jomtien, Thailand, which expanded the focus of the education agenda from access to quality and brought a new interest in the quality of education students were receiving (World Conference on Education for All, 1990). Two of the six goals adopted at the Jomtien conference led to greater interest and support for early grade reading development. They were: Goal 1, the expansion of early childhood care and development activities; and Goal 3, improvement in learning achievement. Based on the aforementioned PICO criteria and time frame, we constructed a search string in five languages—English, Spanish,

[6] Terms pertaining to intervention and control variables such as "intervention," "evaluation," "effectiveness," "outcomes if specified," and so on are used to describe studies and may not appear in the title or abstract of the papers (and therefore will not be retrievable in many databases). Therefore, in most cases, we did not include methods terms in searches (see Brunton et al., 2012).

French, Portuguese, and Dutch—to cover the variety of literature most likely to address early grade reading in the LAC region.

English
(Read* OR Litera* OR writ* OR communic*) AND (primary sch* OR primary grad* OR “grades 1 through 3” OR “grades 1 to 3” OR “grades 1-3” OR "first through third" OR "Grade 1" OR first grade* OR "grade 2" OR second grade* OR “grade 3” OR third grade* OR early grade* OR elementary OR kinder* OR pre-school* OR preschool* OR prekindergarten* OR preK OR pre-K OR “early childhood”) AND (Latin America* OR Caribbean OR South America* OR Antigua* and Barbuda OR Argentin* OR Aruba OR Bahama* OR Barbados OR Beliz* OR Bermud* OR Bolivia* OR Brazil* OR “British Virgin Islands” OR “Cayman Islands” OR Chile* OR Colombia* OR Costa Ric* OR Cuba* OR Curaca* OR Dominica* OR “Dominican Republic” OR Ecuador* OR El Salvador* OR French Guiana* OR Grenada* OR Guadeloup* OR Guatemala* OR Guyana* OR Haiti* OR Hondura* OR Jamaica* OR Martinique OR Mexic* OR Mont Serrat OR “Netherlands Antilles” OR Nicaragua* OR Panama* OR Paraguay* OR Peru* OR “Puerto Ric*” OR “Saint Barthelemy” OR “Saint Kitts and Nevis” OR Saint Lucia* OR “Saint-Martin” OR “Saint Vincent and the Grenadines” OR “Sint Maarten” OR Surinam* OR "Trinidad and Tobago" OR “Turks and Caicos” OR Uruguay OR “Virgin Islands” OR Venezuela)
Spanish
(Leer OR Lecto-escritura OR Alfabetiz* OR “Ambiente letrado”) AND (“la escuela primaria” OR “grados de primaria" OR "grados 1ero a 3ero" OR "grados 1 a 3" OR "grados 1-3" OR "de primer grado a tercer grado" OR "Grado 1" OR "primer grado" OR "primeros grados" OR "primer grado" OR "grado 2" OR "segundo grado" OR "grado 3" OR "tercer grado "OR “grados iniciales” OR “grados tempranos” OR “educación preescolar” OR “Educación maternal” OR “jardín de infancia” OR “Jardines de infancia” OR Kinder* OR preescolar OR pre-kindergarten OR “primera infancia " OR “Educación Inicial”) AND (“Latino América” OR Caribe OR “Sud América” OR “América del Sur” OR “Antigua y Barbuda” OR Argentin* OR Arub* OR Baham* OR Barbados OR Belice* OR Bermud* OR Bolivi* OR Brasil OR “Islas Virgenes Birtánicas” OR “Gran Cayman” OR Chil* OR Colombi* OR “Costa Rica” OR Cub* OR Curaca* OR Dominica* OR “República Dominicana” OR Ecuador* OR “El Salvador” OR “Guayana Francesa” OR Grenada* OR Guadalupe OR Guatemal* OR Guyana* OR Guayana OR Haiti* OR Hondur* OR Jamaic* OR Martinic* OR Méxic* OR “Mont Serrat” OR “Antillas Holandesas” OR Nicaragu* OR Panamá* OR Paraguay* OR Perú* OR “Puerto Ric*” OR “San Bartolomé” OR “Saint Kitts y Nevis” OR “Saint Lucia” OR “Saint-Martin” OR “Saint Vincente y Granadines” OR “San Martín” OR Surinam OR “Trinidad y Tobago” OR “Turks y Caicos” OR Uruguay OR “Islas Vírgenes” OR Venezuel*)
French
(lire OR “la lecture” OR l'écriture OR écrire OR “l’Alphabétisation” OR “environnement lettré” OR “lire-écrire”) AND ("l'école primaire" OR “Enseignement primaire” OR “l'école élémentaire” OR “première année” OR “deuxième année de cycle 2” OR “cours préparatoire” OR “CP” OR “troisième année de cycle 2” OR “cours élémentaire1re année” OR “CE1” OR “première année du cycle 3” OR “cours élémentaire 2e année” OR “CE2” OR “maternelle” OR “Préscolaire” OR “petite enfance”) AND (“Amérique latine” OR Caraïbes OR "Amérique du Sud" OR "Antigua-et-Barbuda" OR Argentine OR Aruba OR Antilles OR Bahamas OR Barbade OR Belize OR Bermudes OR Bolivie OR Brésil OR "Îles Vierges britanniques" OR "Grand Cayman” OR Chili OR Colombie OR “Costa Rica” OR Cuba OR Curaçao OR Dominique OR “République dominicaine" OR Equateur OR “El Salvador” OR Guyane OR Grenade OR Guadeloupe OR Guatemala OR Haïti OR Honduras OR Jamaïque OR Martinique OR Mexique OR “Mont Serrat" OR Nicaragua OR Panama OR Paraguay OR Pérou OR "Puerto Rico" OR “San Bartolomé" OU "Saint Kitts-et-Nevis" OR “Saint Lucia" OR “Saint-Martin” OR “Saint-Vincent-et-Grenadines” OR Suriname OR “Trinité-et-Tobago” OR “îles Turks et Caicos” OR Uruguay OR Venezuela)
Portuguese
(Leitura OR Escrever OR Alfabetização OR “Alfabetização Inicial” OR “Alfabetização Infantil” OR “Alfabetização Emergente” OR “Alfabetização de Crianças” OR “Meio de Alfabetização” OR “Ambiente

Escritura" OR "Compreensão de leitura" OR "Literatura Infantil" OR "tradições orais indígenas" OR "alfabetização inicial endógena na língua maternal") AND ("Escola Primária *" OR "graus elementares" OR "graus primeiro-terceiro" OR "graus 1-3" OR "graus 1-3 "OR " primeiro grau para a terceira série" OR "Grau 1" OR "primeiro grau" OR "séries iniciais" OR "pré-escolar" OR "jardim de infância" OR Creche OR Maternal OR Kinder OR pré-escola OR pré-jardim de infância* OR "primeira infância" OR "Educação da Primeira Infância") AND ("America Latina" OR Caribe OR "América do Sul* OR "Antígua e Barbuda " OR Argentina OR Aruba, OR Bahamas OR Barbados OR Belize OR Bermuda OR Bolívia OR Ilhas Virgens OR Brasil OR Gran Cayman Británicas OR Chile* OR Colômbia* OR Costa Rica* OR Cuba, OR Curacao OR Dominicana* OR Equador OR "El Salvador" OR Grenada OR Guiana OR Guadalupe OR Guatemala* OR Haiti OR Honduras OR Jamaica OR Martinica OR México OR "Mont Serrat" OR "Antilhas Holandesas" OR Nicarágua OR Panamá* OR Paraguai* OR Peru* OR "Porto Rico" OR "São Bartolomeu" OR "São Cristóvão e Nevis" OR "Santa Lúcia" OR "São Martin" OR "São Vicente e Granadinas" OR Suriname OR "Trinidad e Tobago" OR "Turcas e Caicos" OR Uruguai OR Venezuela) AND (meninas OR meninos OR crianças* OR bebês OR infantil)
Dutch
(Lezen* OR Alfabetisering) AND ("basisschool*" OR "basisonderwijs*" OR "groep 3 tot en met 5" "groep 3 tot 5" OR "groep 3-5" OR "groep 3"" OR "groep 4"OR "groep 5"OR kleuterschool* OR peuterspeelzaal* OR kinderopvang* OR brede school* OR "vroegste kinderjaren") AND ("Latijns Amerika*" OR Latijns-Amerika OR "Zuid Amerika* OR "Zuid-Amerika* OR Centraal-Amerika" OR Centraal Amerika" OR Antigua* en Barbuda OR Argentinie* OR Argentinië* OR Aruba OR Bahama's OR Barbados OR Belize OR Bermuda OR Bolivia* OR Brazilië* OR Brazilie* OR "Britse Maagdeneilanden" OR "Kaaimaneilanden" OR Chilli* OR Colombia* OR Columbia* OR "Costa Rica*" OR Cuba* OR Curacao OR Curaçao OR Dominica* OR "Dominicaanse Republiek" OR Ecuador* OR "El Salvador*" OR "Frans Guyana*" OR Grenada* OR Guadeloupe OR Guatemala* OR Guyana* OR Haiti* OR Haïti* OR Honduras OR Jamaica* OR Martinique OR Mexico OR Montserrat OR "Nederlandse Antillen" OR Nicaragua* OR Panama* OR Paraguay* OR Peru* OR "Puerto Rico" OR "Saint Barthelemy" OR "Saint Barthélemy" OR "Saint Kitts en Nevis" OR "Saint Lucia*" OR "Saint-Martin" OR "Saint Vincent en de Grenadines" OR "Sint Maarten" OR Suriname OR Trinidad en Tobago OR "Turks- en Caicoseilanden" OR "Turks en Caicoseilanden" OR Uruguay OR "Maagdeneilanden" OR Venezuela)

We aimed to make the search strings as broad as possible to retrieve the maximum amount of potentially relevant items from all databases (Schuelke-Leech et al., 2015). In theory, the use of one standardized search string ensures an unbiased search strategy across all databases. In practice, using one standardized search string is challenging because the search rules are not standardized across repositories. For example, SAGE Publications has an interface that looks for two-word and longer phrases encapsulated in double quotation marks (e.g., *"early grade."*). In contrast, the Thomson Reuters Web of Science research platform instructs users to include search terms/phrases in parentheses: *(early grad*)*. The rules of using Boolean logic, including wildcards (e.g., "*" and "?"), are also different across various data sources. Furthermore, some databases impose limits on the number of queries and the length of search strings. As a result, the team modified the search string according to each database and documented the iterative process of modifying the search strings (see Appendix B).

6. *Searching for Evidence*

Following the development of the broad search strings, research associates (RAs) at AIR used the search terms and strings (in each of the target languages) to conduct an initial search of online databases and development-focused websites, reviewed bibliographies of accepted articles to find other potentially relevant studies, and sent out emails to EGR experts in the LAC region and

beyond in order to cast a broad net and capture as much of the evidence base as possible. We used three primary methods to search for EGR evidence:[7]

A. Internet searches of predefined online databases, journals, and international development organizations

The review team worked with other researchers, librarians, computer scientists, and content experts to identify appropriate online databases, journals, and international development organizations for our search.

i. Online Databases:

- 3ie
- British Library for Development Studies
- Campbell Collaboration
- Cochrane Library
- Dissertation Abstracts
- Directory of Open Access Journals (DOAJ)
- Directory of Open Access Books (DOAB)
- Development Experience Clearinghouse (DEC)
- Education International
- JSTOR Arts & Sciences I–X Collections and JSTOR Business III Collection
- SAGE Publications
- ScienceDirect
- Taylor & Francis
- Wiley
- WorldCat
- Within EBSCO:
  - Academic Search Premier
  - EconLit
  - Education Source
  - ERIC (Education Resource Information Center)
  - Psychology & Behavioral Sciences Collection
  - PsycINFO
  - SocINDEX with Full Text

ii. Development-Focused Databases/Websites:

- The U.K. Department for International Development (DfID)
- The United States Agency for International Development (USAID)
- The Joint Libraries of the World Bank and International Monetary Fund (JOLIS)
- The British Library for Development Studies (BLDS)
- Institute of Development Studies (eldis)
- The International Initiative for Impact Evaluation (3ie)

[7] Searches were conducted in English, Spanish, French, Portuguese and Dutch as appropriate. No documents were excluded because of language.

- The Abdul Latif Jameel Poverty Action Lab (J-PAL)
- Innovations for Poverty Action (IPA)
- World Health Organization (WHO)
- United Nations Educational, Scientific and Cultural Organization (UNESCO)
- The United Nations Children's Fund (UNICEF)
- The United Nations High Commissioner for Refugees (UNHCR)
- Population Council
- World Vision
- Save the Children
- Plan International
- Organization of American States (OAS)

iii. LAC Region Databases and Websites:

- Latindex
- Red de Revistas Científicas de América Latina y el Caribe, España y Portugal (Redalyc)
- Scientific Electronic Library Online o Biblioteca Científica Electrónica en Línea (SciELO)
- Consejo Latinoamericano de Ciencias Sociales (CLACSO)
- Dialnet
- eRevistas

B. Review of Bibliographies of Articles and Reports for Other References

To ensure we captured all of the relevant and applicable literature in the region, we reviewed the bibliographies of accepted articles and reports to identify relevant and high-quality studies that might fit our criteria. We then searched for these studies and applied our inclusion criteria to them.

C. Surveys With Experts in the Field

EGR experts—particularly those from the wider LAC region—were asked to provide additional sources of evidence that may not have been captured through the online evidence search. We used the snowball approach of contacting researchers and scholars in the field of EGR, who then shared the contacts of others, to share their relevant research, recommend additional research, and forward our request for evidence to their colleagues. Although we did not ultimately identify any new evidence meeting our inclusion criteria during the time frame of the search, we did identify EGR resources for categorization and cataloguing for a separate component of the LRCP work.

7. *Extracting Data From Identified Sources*

We imported all citations found through the above search methods into the Mendeley reference management software (http://www.mendeley.com/). Mendeley automatically extracted bibliographic data from each book, article, or reference and removed all duplicates. At this stage, we were able to identify and export 9,696 unique documents.

The primary focus of the preceding stage ("Initial search for evidence") was to retrieve as many potentially relevant documents from all data sources as possible. However, different data sources have different search functionalities and interfaces. For example, the SAGE Publications website only allowed us to search by a limited number of keywords (e.g., "early grade" AND literacy OR "early grade" AND reading). As a result, we had to limit our results by a number of journal categories

(e.g., Special Education, Regional Studies, Language & Linguistics). In contrast, we were able to use the full search string at the ScienceDirect website (see Annex A on the ScienceDirect website). To overcome these differences in search capabilities, we exported all 9,696 documents into a comma-separated value (CSV) file and applied a "standardized" search string across all documents using the same algorithm in Python.[8] The following sections examine the additional steps that we took to identify the most potentially relevant articles, review them manually, and apply the strict inclusion criteria.

8. *WikiLabeling*

An AIR data scientist applied Wikipedia-based labelling and classification techniques to the extracted data to categorize text into meaningful categories and to increase the relevance of retrieved results using the well-known online encyclopedia, Wikipedia (Egozi, Markovitch, & Gabrilovich, 2011; Gabrilovich & Markovitch, 2006). Due to the broad and inclusive nature of our search strings, much of the initial evidence we captured was not actually relevant to our review. Therefore, we applied Wikipedia-based labelling to help us identify the most relevant pages. The process of identifying these pages is two-fold: first, experts need to share a list of potentially relevant categories. Next, we had to mine Wikipedia to find pages associated with exactly these or similar categories. We then validated the resulting list with the experts again. For example, "learning outcomes," originally proposed by our experts, maps directly to "outcome-based education" within Wikipedia. Wikipedia's innate hierarchical structure allowed us to make our categories less ambiguous and better organize them into a meaningful list (Box 1).

Box 1. List of Relevant Categories That Have Individual Wikipedia Pages

- Dual language
- Emergent literacies
- First language
- Fluency
- Free writing
- Grammar
- Language education
- Language proficiency
- Listening
- Literacy
- Orthography
- Outcome-based education
- Phonemic awareness
- Phonics
- Phonological awareness
- Reading (process)
- Reading comprehension
- Second-language
- Second language acquisition
- Spoken language
- Transitional bilingual education
- Understanding
- Vocabulary
- Writing

We combined the WikiLabeling results with the "standardized" search term strategy described in the previous section. Although WikiLabeling is generally effective at assessing the overall context of a document and its relevance to a given subject, the search term strategy helps narrow down the

[8] Python is a programming language with multiple relevant modules and libraries that was used in this project (https://www.python.org/).

search by a number of very specific keywords and phrases, such as individual countries and the region name. We used this approach to categorize documents in all target languages (English, French, Spanish, Dutch, and Portuguese).

The "standardized" search term strategy and WikiLabeling are complementary in several important ways:

- Search terms and regular expressions help discover individual words and phrases within a document, no matter where they appear. For example, the geographic region may be mentioned only in the discussion part of a paper when writing about broader potential impacts. Meanwhile, the main body of the paper might have nothing to do with Latin America or the Caribbean (for example, we have seen some studies evaluating an intervention in sub-Saharan Africa, which mention other developing countries that could learn from this experience). In contrast, Wikipedia-based labelling assesses the entire context of the document by comparing all words and phrases used in academic papers and comparing them to the ones used to describe individual concepts, such as "language education" or "phonological awareness."
- Search strings can cover a wide range of inclusion criteria and be structured to include three or four different variables. WikiLabeling looks into every concept individually and therefore provides a more in-depth assessment of the relevance of a document for the subject of focus.
- Search term strategies are more flexible and do not depend on the user community curating an online encyclopedia every day. However, the continuous curation in Wikipedia helps improve the quality of knowledge and introduce new meaningful concepts into the scientific language through discovery and analysis applied in WikiLabeling.

For this systematic review, we used the search term strategy followed by Wikipedia-based labelling and classification to define which documents were most likely to be relevant for the subject in focus. This approach has four major advantages:

1. **Cost effectiveness:** Our initial search returned 9,696 unique literature references from all sources. A typical approach is to use a number of research assistants to review these documents and establish their relevance based on inclusion criteria. Such an approach is costly and time consuming. Machine learning through WikiLabeling allowed us to significantly reduce the cost by removing the most likely irrelevant results and ranking the remaining potentially relevant documents (just like Google ranks search results on its page after the user inserts their query). This approach significantly reduced the time research assistants required to review documents.
2. **Robustness:** Computational approaches are largely systematic and unbiased in how they decide the relevance of documents on a given subject. Both the search term strategy and Wikipedia-based labelling apply standardized approaches and offer several methods of robust evaluation and validation. Systematically tracking the actions and decisions of a diverse group of research assistants is virtually impossible.

3. **Better coverage:** Advances in library science and information retrieval allow more and more databases to be more structured and searchable by the research community. It is also important to note that good knowledge of machine learning techniques allows for retrieving a wider set of potentially relevant documents from a bigger number of data sources.
4. **Less subject to human error:** Although humans continue to play a critical role in curating and validating the final document selections, a computational model applies only one set of relevance parameters through all documents. As mentioned, no matter how firm and well formulated the inclusion criteria are, individual researchers can bring their expertise into the selection process.

Importantly, our approach supplements but does not replace the human review of potentially relevant articles. We built in several quality control procedures to ensure that our algorithm did not lead to the exclusion of relevant papers. We created four samples, with 100 abstracts each. Within each sample we included a set of 80 randomly selected abstracts that were retrieved by the search strategy, WikiLabeling, or both. The remaining 20 documents were randomly selected from the subset not retrieved by any of our approaches (i.e., 8,145 documents that were considered as irrelevant by the search strategy, WikiLabeling, or both). We then distributed these samples to four senior reviewers and reading experts and asked them to identify the irrelevant articles. This process enabled us to check for both false negatives (articles not retrieved through our search approach–the 20–but which were deemed relevant) as well as false positives (articles retrieved through our search approaches–the 80–but which were deemed irrelevant).

Table 1 shows that the number of false negatives was not very high (an average of 3 percent), which indicates that our methods were able to capture the majority of all potentially relevant studies among the 9,696 originally downloaded from all data sources. At the same time, the false positive error rate was rather high (0.67 average). There are two reasons for such a high rate of false positives: (1) it is generally impossible to integrate all inclusion criteria—such as the quality of the study and methodology—into the automated computational techniques; and (2) our goal was to improve the overall coverage of the systematic review while also reducing the cost of reviewing thousands of documents, which led us to use a more flexible search and retrieval strategy while removing a large number of clearly irrelevant studies. Thus, we erred on the side of sensitivity rather than specificity to not miss any relevant research.

Table 1. False Negative and False Positive Errors

Error Type	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	AVERAGE
False Negative	0	0.04	0.02	0.06	**0.03**
False Positive	0.66	0.68	0.66	0.68	**0.67**

9. *Applying Inclusion Criteria and Recording Key Indicators*

After narrowing down our list of articles through WikiLabeling, we imported all relevant citations back into the Mendeley reference manager software. We divided citations among RA reviewers, who applied the predetermined inclusion criteria (see Table 2) to each title and abstract. We chose to err on the side of sensitivity rather than specificity during our initial title and abstract review. Our inclusion criteria were purposefully broad because we did not want to miss any relevant

citations due to narrow inclusion criteria. Any article that did not meet one of the following five threshold criteria laid out in Table 2 was automatically excluded from further review.

Table 2. Initial Inclusion Criteria for EGR Evidence

#	Category	Criteria	Notes
1	**Year of Publication**	Include *literature* from the last 25 years, a time frame spanning 1990–2015. We will update the search in each subsequent year of the 5-year project.	• If unpublished, the research must have been conducted in that time frame.
2	**Relevance to the Region**	The evidence must be from or on the LAC region including any or all of the following: Antigua and Barbuda, Argentina, Aruba, Bahamas, Barbados, Belize, Bermuda, Bolivia, Brazil, British Virgin Islands, Cayman Islands, Chile, Colombia, Costa Rica, Cuba, Curacao, Dominica, Dominican Republic, Ecuador, El Salvador, French Guiana, Grenada, Guadeloupe, Guatemala, Guyana, Haiti, Honduras, Jamaica, Martinique, Mexico, Mont Serrat, Netherlands Antilles, Nicaragua, Panama, Paraguay, Peru, Puerto Rico, Saint Barthelemy, Saint Kitts and Nevis, Saint Lucia, Saint-Martin, Saint Vincent and the Grenadines, Sint Maarten, Suriname, Trinidad and Tobago, Turks and Caicos Islands, Uruguay, US Virgin Islands, Venezuela	• We will not include research on migrants from the LAC region residing outside the region.
3	**Relevance to the Population**	Boys or girls ages birth through Grade 3 in the LAC region. If the children are enrolled in Grade 3 or below, they fall within our population regardless of the age.	• We included all research that focuses at least partly on this age group even if other populations of interest were included.
4	**Relevance to the Topic**	The literature must have a focus on reading or literacy (which includes reading and writing).	• We included all research focusing at least partly on reading or literacy even if it addressed multiple areas. We did not include research that could have an effect on reading but does not actually discuss that link (e.g., IQ studies). • Research on writing was included automatically if it also discussed the link to reading or literacy.
5	**Is It Research?**	There must be a research question or research objective and a methodology that matches that objective.	• If the document was a literature review or systematic review, then we did not include it in our review. We instead focused on the primary studies cited in that literature review.

During the title and abstract review, reviewers selected "yes," "no," "unclear," or "not rated" on the Excel spreadsheet for each of the inclusion criteria (i.e., published since 1990, from or on the

LAC region, ages birth to Grade 3, reading or literacy focused, and includes a research question or objective). Here is an explanation of each option:

- Marking "yes" for any of the five criteria indicated that the reviewer should continue onto the next criterion on the coding sheet. If the reviewer marked "yes" to all of the inclusion criteria, then they were required to fill in the remaining indicators outlined in Table 3.
- Marking "no" indicated that the reviewer should stop because the study did not meet the criteria for further review. In this case, the remaining inclusion criteria were automatically marked as "unrated," signifying that the study failed to meet one of the inclusion criteria and thus, whether it met the other criteria was no longer relevant.
- Marking "unclear" indicated that the study was tagged for review by a senior reviewer. At this stage, we followed the motto "When in doubt–include," and maintained a record of all excluded articles indicating for what criteria they were excluded.

Reviewers then used the same Excel spreadsheet to record key indicators (Table 3) for literature that met all five inclusion criteria.[9]

Table 3. Key Indicators

Categories	Selection choices
Abstract Number	
Citation Information	
Abstract	
Document Reviewer Name	
Country(ies) of Focus	Antigua and Barbuda, Argentina, Aruba, Bahamas, Barbados, Belize, Bermuda, Bolivia, Brazil, British Virgin Islands, Cayman Islands, Chile, Colombia, Costa Rica, Cuba, Curacao, Dominica, Dominican Republic, Ecuador, El Salvador, French Guiana, Grenada, Guadeloupe, Guatemala, Guyana, Haiti, Honduras, Jamaica, Martinique, Mexico, Mont Serrat, Netherlands Antilles, Nicaragua, Panama, Paraguay, Peru, Puerto Rico, Saint Barthelemy, Saint Kitts and Nevis, Saint Lucia, Saint-Martin, Sint Maarten, Saint Vincent and the Grenadines, Suriname, Trinidad and Tobago, Turks and Caicos Islands, Uruguay, US Virgin Islands, Venezuela, or multiple countries
Region	South America, Central America, Caribbean, North America
World Bank Income Level	Low income, lower-middle income, upper-middle income, high income non-OECD, high income OECD
Type of Document	Journal article, technical report, dissertation/thesis, book chapter, other
Full Text Available to AIR	Yes, No, Other
Full Text Available to Public	Yes, No, Other
How Was Document Located?	Source bibliography, hand search of journal, online source, in-person contact, recommended by a content expert

[9] After an initial review of a subset of citations, we refined our key indicators as needed to make them more explicit and relevant to the types of evidence we found during the search.

Categories	Selection choices
Language of Publication?	English, Spanish, French, Dutch, Portuguese, Bilingual, Other
Target Group	Early childhood, pre-primary (pre-k or kindergarten), primary, out-of-school children (school-age children who are not enrolled), other
Type of Evidence	**Quantitative:** Intervention-based: Experimental, Quasi-Experimental, Multivariate Regression, Univariate Regression, Graphics, Other **Quantitative**: Nonintervention-based: Psychology, linguistics, reading science studies (methods include structural equation models, multivariate and univariate regressions, lab-type pilot studies, writing system analyses, other) **Qualitative:** Intervention, nonintervention: Case study, focus groups, interviews, multiple methods, other **Mixed Methods:** Includes both quantitative and qualitative methodologies

10. *Reviewing Full Text Using Quality Review Protocols*

We compiled all of the full-text articles and books that met all inclusion criteria, as well as those that were still unclear after the title and abstract review, and assigned them to senior researchers based on language and type of study. The senior researchers reviewed the articles using separate quality review protocols based on the type of study as follows:

- **Quantitative intervention studies**: An adapted version of a risk of bias (RoB) assessment tool[10] developed by Hombrados and Waddington (2012)
- **Quantitative nonintervention studies**:[11] An adapted version of the RoB tool for quantitative intervention studies, which removed any questions regarding interventions.
- **Qualitative intervention and nonintervention studies**:[12] An adapted version of the Critical Appraisal Skills Programme (CASP) Qualitative Research Checklist
- **Mixed-methods studies**: Both the RoB tool and the qualitative protocols were applied

Two or more reviewers read and rated all quantitative intervention studies to ensure consensus. The reviewers resolved disagreements in assessments through discussion or by third-party adjudication. We reread studies several times if something was unclear and maximized the use of all the available information from the studies. We based our assessment on the reporting in the primary studies, erring on the side of caution. For example, in those cases in which it was not clear whether standard errors were clustered, we assumed the standard errors were not clustered and took that into consideration in our risk of bias assessment.[13]

[10] See Appendix C for the full RoB tool and the justification of the risk of bias for the included studies.
[11] See Appendix D for the full RoB tool for quantitative nonintervention studies.
[12] See Appendix E for the full quality review tool for qualitative studies.
[13] We contacted authors after finishing the first draft of this report in those cases where information was unclear or incomplete. This approach enabled us to retrieve some extra information, which we incorporated in the synthesis of the results. However, not all authors responded. In addition, not all authors were able to address our questions. In those cases where authors did not respond or were not able to provide the required information we had to make assumptions on how to interpret the findings of the studies and we erred on the side of caution.

For the other types of studies, pairs of reviewers rated the same studies at the outset to ensure a common understanding of the quality categories, but the remaining articles were reviewed by single reviewers because of time constraints.

Risk of Bias Assessment for Quantitative Intervention Studies.

We used an adapted set of criteria to determine the rigor of the quantitative intervention studies and to assess risk of bias in experimental and quasi-experimental studies (Hombrados & Waddington, 2012). Specifically, we assessed the risk of the following biases:

1. **Selection bias and confounding**, based on the quality of the identification strategy used to determine causal effects and assessment of equivalence across the beneficiaries and comparison or control group
2. **Performance bias**, based on the extent of spillovers to the students in the control or comparison groups and contamination of the control or comparison group
3. **Outcome and analysis reporting biases**, including:
    - The use of potentially endogenous control variables
    - Failure to report nonsignificant results
    - Other unusual methods of analysis
4. **Other biases**, including:
    - Courtesy and social desirability bias
    - Differential attrition bias
    - Strong researcher involvement in the implementation of the intervention and the Hawthorne effect

Risk of Bias Assessment for Quantitative Nonintervention Research. The quantitative nonintervention quality review tool assesses the relevance, data and methodology, and analytical approach of the research by eliciting reviewers' responses to 18 quality criteria questions. Upon reading the full-text article, reviewers must respond to each question by selecting "Yes," "No," "Unclear," or "N/A" and provide a justification for their rating, citing the text whenever possible. Finally, reviewers must provide a summary of the article's main findings and their relevance to target stakeholder groups.

Quality Review for Qualitative Intervention and Nonintervention Research. We designed the qualitative intervention quality review tool to assess the research design, data analysis, ethical considerations, and the relevance to practice. The tool examines reviewers' responses to 11 main questions, each of which has multiple subquestions. Upon reading the full-text article, reviewers must select either "High," "Medium," "Low," "N/A," or "Not Mentioned" for each of the 11 questions and subquestions and provide a justification for their rating. The justification should also be supported with text and page numbers from the article. Reviewers are encouraged to comment on both strengths and weaknesses when applicable. The 11 qualitative review questions were divided into three categories: research design, ethics and reflexivity, and relevance to the field as shown below:

Research Design:

- Clear statement of research?
- Appropriateness of qualitative methodology?
- Addresses the aims of the research?
- Was the data collected in a way that addressed the research issue?
- Was the data analysis sufficiently rigorous?
- Is there a clear statement of findings?

Ethics and Reflexivity:

- Has the relationship between researcher and participants been adequately considered?
- Have ethical issues been taken into consideration?
- Appropriate recruitment strategy?

Relevance to the Field:

- How valuable is the research?
- Information for stakeholders to assess replicability?

In addition to these 11 quality criteria, reviewers were asked to summarize the main findings of each article in regards to how they might affect various stakeholder groups in the LAC region such as policy makers, international NGOs, teacher training institutes, and researchers. Finally, reviewers were asked to review the bibliography for each article and to list any other potentially relevant references for further review.

Quality Review for Mixed-Methods Research. Reviewers completed both a quantitative and a qualitative quality review protocol for mixed-methods articles.

11. *Analyzing Data*

We used different types of analyses for each type of research. First, we implemented a combination of meta-analysis and narrative synthesis to analyze the effectiveness of programs that could potentially influence reading outcomes. We calculated the standardized mean difference and the standard error for each of the studies included in the meta-analysis. Where possible, we used a stratified meta-analysis to differentiate the results of studies with a low, medium, or high risk of bias and to differentiate between the findings of RCTs and nonexperimental studies. Second, we used a narrative review to examine the main lessons from the included qualitative intervention and nonintervention studies. To identify these lessons, we relied mostly on the findings of high-quality studies. Third, we analyzed the main lessons about the predictors of early grade reading outcomes in the LAC region from quantitative nonintervention studies. For this purpose, we again relied on the studies that were identified as higher quality in the risk of bias assessment.

Quantitative Intervention Studies. For the analysis of the quantitative intervention studies, we first calculated effect sizes for each of the included quantitative studies that were eligible for inclusion in the meta-analysis. We then conducted a meta-analysis, which is a way of statistically pooling the effect sizes from different studies in order to identify patterns among study results, sources of disagreement among those results, or other interesting relationships that may come to light in the context of multiple studies. We conducted separate meta-analyses to determine the impact of teacher training, ICT, and nutrition programs on early grade reading outcomes (for more information, see the section on the meta-analysis) because we found a sufficient number of high-quality and relatively homogeneous studies about the effects of these programs on early grade reading outcomes. We also conducted a narrative synthesis for studies that could not be included in the meta-analysis but showed evidence concerning the impact of specific programs on early grade reading outcomes. The narrative synthesis included studies which emphasized the effects of preschool, school governance, specific teacher practices, and parental involvement on early grade reading outcomes. The following section will discuss (1) calculation of effect sizes, (2) meta-analysis, and (3) narrative review.

Measures of Treatment Effects. We extracted information from each quantitative study that was eligible for inclusion in the meta-analysis to estimate standardized effect sizes. In addition, we calculated standard errors and 95 percent confidence intervals if possible. We calculated the Hedges' g sample-size-corrected standardized mean differences (SMDs) for continuous outcome variables, which measures the effect size in units of standard deviation of the outcome variable.

We first calculated SMDs (Cohen's d) by dividing the mean difference with the pooled standard deviation by applying the formula in equation 1:

(1) $\text{SMD} = \frac{Yt-Yc}{Sp}$

SMD refers to the standardized mean differences, Yt refers to the outcome for the treatment group, Yc refers to the outcome for the comparison group, and Sp refers to the pooled standard deviation.

The pooled standard deviation Sp can be calculated by relying on the formulas in equations 2 and 3:

(2) $\text{Sp} = \frac{\sqrt{((SDy^2)*(nt+nc-2))-(\frac{\beta^2*(nt*nc)}{nt+nc})}}{nt+nc}$

(3) $\text{Sp} = \frac{\sqrt{(nt-1)*st^2+(nc-1)*sc^2}}{nt+nc-2}$

We used equation 2 for regression studies with a continuous dependent variable. In this equation, SDy refers to the standard deviation for the point estimate from the regression, nt refers to the sample size for the treatment group, nc refers to the sample size for the control group, and β refers to the point estimate. We used equation 3 when information was available about the standard deviation for the treatment group and the control group.

We corrected the standardized mean difference for small sample size bias by relying on equation 4, which transforms Cohen's d to Hedges' g.

(4) SMDcorrected = SMDuncorrected * $(1 - \frac{3}{4*(nt+nc-2)-1})$

We also relied on equation 5 to estimate the standard error of the standardized mean difference:

(5) SE= $\sqrt{\frac{nt+nc}{nc*nt} + \frac{SMD^2}{2*(nc+nt)}}$

We adjusted standard errors for those studies that use outcome variables that are clustered at a higher level of aggregation than the student level but do not take this into consideration in the estimation of the standard errors and confidence intervals. For these studies, we applied corrections to the standard errors and confidence intervals using the variance inflation factor (Higgins & Green, 2011):

$$SEcorrected = SEuncorrected \times \sqrt{(1 + (m-1) * ICC)}$$

Here, m is the number of observations per cluster, and ICC is the intracluster correlation coefficient. To identify the ICC, we relied on a study by Yoshikawa et al. (2015), who estimated the ICC for reading outcomes of students clustered in schools in Chile. They found an ICC of 0.10. Although this estimate is most likely not externally valid for the rest of the LAC region, it is our best estimate of the ICC that is available to us. Thus, we rely on this estimate for our effect size calculations.

In those cases in which we were not able to retrieve the missing data, we extracted or imputed effect sizes and associated standard errors based on commonly reported statistics such as the t or F statistic or p or z-values using David Wilson's practical meta-analysis effect-size calculator. Where studies did not report sample sizes for the treatment and the control or comparison group, we assumed equal sample sizes across the groups. We report our format for the calculation of effect sizes in Appendix F.

Methods for Handling Dependent Effect Sizes. We included only one effect size per study in a single meta-analysis. Where studies reported more than one effect size on the basis of different statistical methods, we selected the effect size with the lowest risk of bias. Where studies presented several impact estimates for different variables that measure the same reading construct, we used a sample-size weighted average to measure a "synthetic effect size." Examples of reading constructs include decoding, vocabulary acquisition, and reading comprehension. Importantly, there were insufficient studies that reported impacts on more than one reading construct. The majority of the studies that we were able to include in the meta-analysis only determined the impact of the evaluated program on a standardized language assessment for the grade level. Furthermore, the majority of the studies did not provide enough information about the assessment of reading to determine which reading constructs were measured. For example, none of the included studies provided details about the contents of the assessment test. Thus, we did not conduct separate meta-analyses for more than one reading construct because there was insufficient information about effect sizes for different reading constructs. Therefore, we assumed that the effect sizes were similar for different reading constructs or calculated synthetic effect sizes. This approach does not allow us to examine separate impacts on different reading constructs. Furthermore, it requires the assumption that effect sizes are not dependent upon the specific reading construct that is used as an outcome variable. These assumptions are not necessarily realistic, but we needed to make them

in order to enable a meta-analysis across studies. To mitigate these concerns, we complemented the meta-analysis with a narrative review approach. In addition to the meta-analysis for early grade reading outcomes, we were able to conduct a meta-analysis to determine the effects of nutrition programs on early grade spelling outcomes.

We also calculated synthetic effect sizes for different grades and different age groups and assumed homogenous effects across age groups when heterogeneous effects were not reported. We did not find sufficient studies that reported separate effects for different grades or age groups to report separate meta-analyses by grade or age group. We also found several studies that only reported average effects for students that meet our inclusion criteria (Grade 3 and below) and students that did not meet our inclusion criteria. We include heterogeneous effect sizes for Grade 3 and below when this information is available as in Osorio and Linden (2009). However, other studies only report average effects for students in different age groups. For example, Beuermann, Cristia, Cueto, Malamud, & Cruz-Aguayo (2015) reported average impacts of the provision of one laptop per child for children in Grades 2, 4, and 6. In this case we decided to include a homogenous effect size that assumes the effects are equivalent for each of these age groups. Again, this assumption may not be realistic, but we needed to make this assumption to enable a meta-analysis. To mitigate this concern, we complemented the meta-analysis with a narrative review.

Meta-Analysis. We conducted separate meta-analyses to determine the effects of nutrition programs, teacher training programs, and ICT programs because these were the three topics for which we had sufficient numbers of studies for a meta-analysis. When the number of studies allowed for it, we examined the heterogeneity of the effect sizes for each outcome across studies. We examined heterogeneity by using I-squared and Q as well as tau-squared and the visualization of the forest plots (Borenstein, Hedges, Higgins, & Rothstein, 2009). We used Stata to conduct the meta-analysis.

When the number of studies allowed for it, we performed a sensitivity analysis for two methodological effect size moderators:

- Risk of bias status for each risk of bias category; and
- Study design (RCTs versus quasi-experimental studies).

We also considered the methodology and the risk of bias of the included studies in the interpretation of the meta-analysis when our sample size did not allow for such stratified meta-analyses.

We started with separate meta-analyses of RCTs and quasi-experimental evaluations for determining the effects of each of the programs. Then we used an iterative approach to determine the potential bias from pooling RCTs and quasi-experimental evaluations and studies with low, medium, and high risk of bias for each of the types of bias we assessed in our risk of bias assessment. We used random-effects meta-analysis because the average effect of programs that influence reading outcomes is likely to differ across contexts due to differences in program design or contextual characteristics. This approach is in line with the approach used in a recent systematic review on the effects of women's self-help groups on women's empowerment (Brody et al., 2015).

Publication Bias. We used two methods to determine the potential for publication bias. First, we assessed the potential for publication bias using funnel plots. Second, we conducted Egger's test to determine the potential for publication bias in studies that focus on reading outcomes.

Qualitative Studies and Quantitative Nonintervention Studies.

After using the quality protocols to review full-text qualitative and quantitative nonintervention articles, we coded the protocols using NVivo qualitative data analysis software (QSR International Pty Ltd., Version 10, 2012). NVivo is traditionally used to manage and code empirical data (Bhattacharyya, 2004; Caldeira & Ward, 2003; Patashnick & Rich, 2004). It is also used for secondary data in document analysis, such as reports, websites, and other sources. A team of analysts trained in using the qualitative software program conducted the data analysis process by coding and analyzing the quality ratings and justifications for each study.

In order to code and analyze the quality ratings and justifications for each article, we created three separate NVivo files for the qualitative intervention research, qualitative nonintervention research, and quantitative nonintervention research. Once we coded all of the quality criteria and justifications in NVivo, reviewers compared the quality of each criterion across all articles of a particular research type. For example, a reviewer could compare the quality of the statement of research across all qualitative intervention studies. We then wrote up a synthesis of the findings for each quality criterion for each research type using the NVivo coding structure.

In order to synthesize the study findings for each research type, we also used NVivo as a tool for the qualitative research. Analysts created separate NVivo files for intervention and nonintervention research and imported the reviewers' statements of findings for each included study. They then coded these statements of findings into topic nodes (these were predetermined by literacy experts as covering the main areas of early grade reading). Once the coding was complete, the analysts were able to see all of the findings for each topic area and could then write up the analysis and implications by topic area.

We only included findings for high-quality and medium-quality articles in our synthesis. To determine which qualitative studies were of sufficient quality to report on the findings, we created an Excel file with all 26 qualitative intervention and nonintervention studies as well as their ratings on each of the quality criterion. This sheet enabled us to see all of the ratings in one view and determine if a study was strong enough to be included. We could then refer back to the original protocol and the reviewers' justifications to make sure that the study met certain criteria such as having a research question, matching methodology, transparent methods of analysis, substantiated findings, and so forth.

In order to synthesize the findings of the quantitative nonintervention research, we first determined which studies should be included in the analysis. To do this we referred to the quality protocols filled out by the reviewers for each article and only included studies that were considered high quality in all the pertinent ratings in the protocol. For instance, if there was missing information about data administration or no information provided about how the language of testing was determined, we did not dismiss the study; however, if the reviewers judged that there were serious problems with the method or sample selection, we did not include the study in our analysis.

12. *Mapping the Gaps*

"Evidence-gap maps present a visual overview of existing systematic reviews or impact evaluations in a sector or subsector, schematically representing the types of interventions evaluated and outcomes reported" (Snilstveit et al., 2012, p. 1). In this systematic review, we present a visual overview of the evidence about reading outcomes in the LAC region. In addition, we present additional gap maps that show the studies by topic area and country to give the reader a visual representation of the gaps in research topics. We go beyond a focus on effectiveness by creating gap maps for quantitative nonintervention and qualitative studies in addition to quantitative intervention studies.

To create the evidence-gap map for the quantitative intervention studies, we coded the intervention types and outcome measures and linked these to various characteristics of the evaluated programs. This process allowed us to create evidence-gap maps that demonstrate what evidence is available on the impact of teacher training, nutrition, ICT, preschool, school governance, teacher practices, and parental involvement on reading outcomes. In this process we differentiated evidence-gap maps by methodology (experimental versus nonexperimental), risk of bias, socioeconomic condition (high-income versus upper-middle income, lower-middle-income, and low-income country), and country.

We created a separate evidence-gap map for quantitative nonintervention and qualitative intervention and nonintervention studies that shows the study topics, type of research, and country of the research for the medium and high quality studies. The following list of topics was developed by a team of literacy experts as representative of the various topic areas within the field of early grade reading. During the analysis of findings, reviewers categorized each study into one of the following topic areas:

- Assessment
- Child nutrition
- Curriculum
- Disabilities
- General pedagogical approaches
- Information and Communication Technologies (ICT)
- Parental and community participation
- Poverty
- Preliteracy/emergent literacy
- Preschool
- Reading in bilingual/multilingual contexts
- Reading materials
- Reading skills
- School governance

- Teaching practices for reading
- Teacher training
- Writing

These categories formed the content coding scheme, which was geared toward separating the raw data from the protocols into "large buckets," with lower level subcodes used to identify data that addressed specific subtopics. The team defined each broad category and subcode to ensure consistency across coders and over time. To correctly analyze studies for the content review, the team used utilization-focused sampling among the included articles to "select a set of cases concerning a problem or issue where sufficient depth and detail in specific cases will support rigorously identifying key factors that can credibly inform future decision making" (Patton, 2015a, p. 270)

13. *Triangulating Findings*

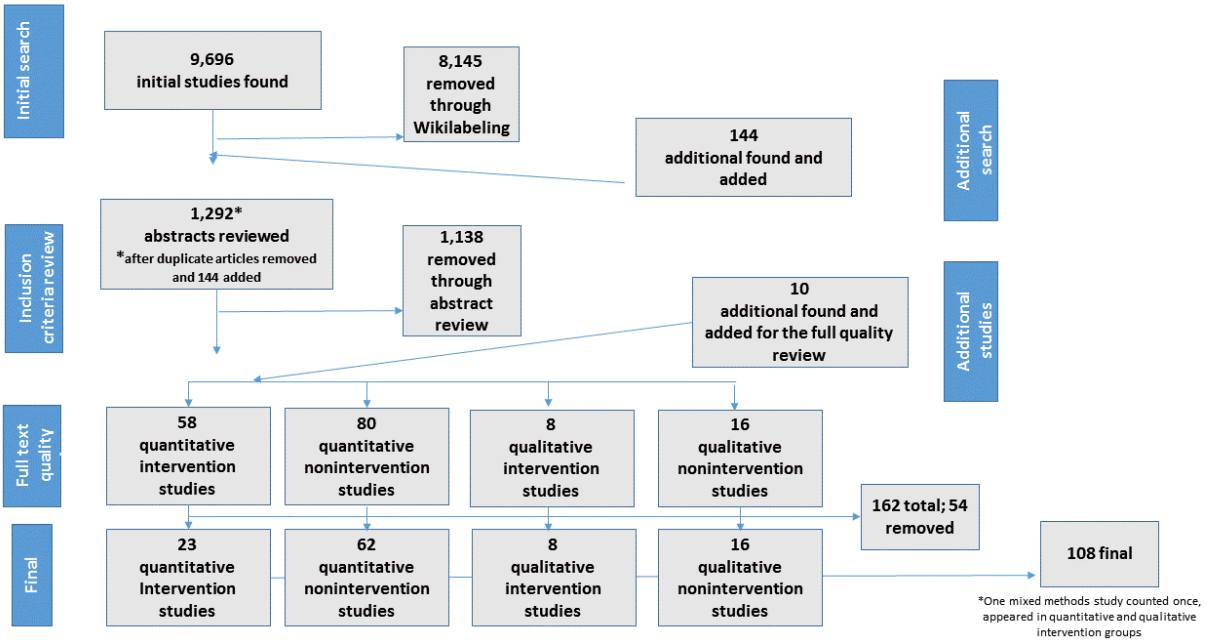
After conducting the quality review and synthesis of articles and mapping the gaps, reviewers triangulated the different syntheses by linking the evidence back to the conceptual framework. We examined the impact of the different programs on reading outcomes and triangulated these findings with the qualitative research articles to examine whether the fidelity of implementation or experiences and perspectives of different stakeholders may have influenced the impact of these programs. In addition, we assessed the predictors of reading outcomes to increase our understanding about the linkages between intermediate outcomes, such as teacher knowledge and behavior, and reading outcomes. Finally, we used the information from the qualitative research to examine whether and where any links in the conceptual framework broke down. Findings from the qualitative synthesis and the quantitative nonintervention synthesis helped describe, explore, and interpret how specific programs improve reading outcomes.

The triangulation of findings from different research methods allowed us to define and test hypotheses using different methodologies that informed and supplemented each other. This approach allowed us to capture the state of the evidence on whether and how specific programs improve reading outcomes in Latin America as well as the gaps in the evidence.

Results of the Analyses

Our literature search aimed to answer the first research question: What are the existing intervention- and nonintervention-based studies and what is the existing literature from or on the LAC region involving reading programs, practices, policies, and products focused on improving reading skills for children from birth through Grade 3? We searched the available studies and found 9,696 unique results. We then conducted a literature review and several types of risk of bias and other quality assessments to assess the quality of the existing EGR evidence (quantitative intervention and nonintervention and qualitative intervention and nonintervention) in the LAC region and its practical use for varied LAC region stakeholders. Finally, for qualifying quantitative articles, we sought to determine the impact of reading programs, practices, policies, and products aimed at improving the reading skills for children from birth through Grade 3 on reading outcomes in the LAC region through meta-analysis and narrative synthesis. Figure 5 depicts the systematic review phases from initial search through quality review. It indicates the number of studies that passed into each subsequent phase of review as well as the numbers of studies that were removed at each phase.

Figure 5. Systematic Review Phases: Initial Search to Quality Review



We conducted the search from July to August 2015 and applied the WikiLabeling approach in September 2015.

Initial Search: We found 9,696 studies using our search strings and modified strings for all online sources. We applied WikiLabeling in order to identify the most relevant of the 9,696 documents and removed 8,145 documents that were identified as irrelevant.

Inclusion Criteria Review: We retrieved 144 additional articles through other search engines that we identified as having potentially relevant research. We reviewed these articles against the inclusion criteria along with the articles identified through WikiLabeling for a total of 1,292 articles reviewed. During this stage, reviewers applied the five inclusion criteria to titles and abstracts and an additional 1,138 articles were rejected. (See Appendix G for details on the number of articles rejected for each inclusion criterion).

Full-Text Quality Review: One hundred sixty-two articles moved on to the next phase, full-text review. The articles reviewed during this phase included 152 articles that either met all five inclusion criteria or met all criteria with one or more criteria listed as unclear (i.e., it could not be determined from reviewing the abstract whether it met the criteria), plus 10 additional studies that were identified through web searches or snowballing of references and met all inclusion criteria. These articles were reviewed in their entirety.

Final: In the final full-text article review stage, we rejected an additional 54 articles for the following reasons:

- We were not able to access the full text of the article.
- During the inclusion criteria review, reviewers marked many articles as "unclear." Upon reviewing the full text, reviewers were able to determine that the articles did not meet the inclusion criteria.
- Unpublished dissertations and theses were excluded from the qualitative and quantitative nonintervention research due to the sheer volume.[14]

Characteristics of Included Studies

Table 4 summarizes the characteristics of all articles included in the final review. The articles are categorized by publication type, year of publication, region and country of focus, language of publication, research type, and the country of focus income level (as determined by the World Bank).

Table 4. Characteristics of Final Included Reviews

	N	%
Publication type		
Dissertation/thesis	3	3%
Journal article	97	90%
Technical report	5	5%
Working paper	3	3%
Year of publication		
1990–1995	5	5%
1996–2000	12	11%
2001–2005	15	14%

[14] We will add the evidence from these sources when we update our systematic review in subsequent years.

	N	%
2006–2010	28	27%
2010–2016	48	44%
Region and country of focus		
Caribbean		12%
Cuba	2	
Jamaica	6	
Puerto Rico	4	
Central		5%
Costa Rica	1	
Guatemala	4	
North		17%
Mexico	18	
South		63%
Argentina	10	
Brazil	27	
Chile	15	
Colombia	6	
Guyana	1	
Peru	7	
Uruguay	1	
Venezuela	2	
Multiple countries	3	3%
Language of publication		
English	61	56%
Portuguese	15	14%
Spanish	32	30%
Type of research		
Qualitative intervention	8	7%
Qualitative nonintervention	16	15%
Quantitative intervention	22	20%
Quantitative nonintervention	62	57%
Country of focus income level (World Bank)		
Lower middle income	5	5%
Upper middle income	79	73%
High income	20	18%
Not applicable/multiple countries	4	4%

Table 4 shows that the vast majority of studies included in our review of evidence were published journal articles and came from either North or South America with significantly fewer from

Central America and the Caribbean. Obtaining unpublished research is not an easy task especially when the systematic review team is located outside the LAC region. The only Central American countries represented were Costa Rica and Guatemala, and for the Caribbean, Puerto Rico, Jamaica, and Cuba were represented. Most articles were published in English or Spanish. We found no articles in any regional languages, which may be due in large part to publication bias and lack of availability of journals in languages that are not national languages. Alternatively, the limited number of articles in indigenous languages may be an indication of the limited resources available to indigenous populations to conduct research on early grade reading.

Another interesting finding shown in the table is that that more than 90% of the articles were focused on high- to upper-middle income countries. The disproportionate emphasis on high-income and upper-middle-income countries may be explained by the limited available resources and capacity for conducting high-quality research in low-income and lower-middle-income countries.

The following sections summarize the results of our systematic review of the literature. We describe the quality parameters and present results of our quality review for the four types of research, including: qualitative intervention, qualitative nonintervention, quantitative intervention, and quantitative nonintervention. We also triangulate the findings across research types.

Quantitative Intervention Research

Analyses

We included 23 experimental and quasi-experimental papers that focused on determining the effects of various development programs on early grade reading outcomes in Latin America. Together these 23 papers evaluate the effects of 24 unique programs or program components on reading outcomes. Of the 23 included papers, three papers estimate the impact of more than one program (Cardoso Martins et al., 2011; Larrain et al., 2012; Vivas, 1996). Three papers focused on the effect of the same program in Chile (Gomez Franco, 2015; Mendive et al., 2016; Yoshikawa et al., 2015). In our meta-analyses, we treat each of the 24 unique programs as one observation. Of the three papers that focused on the same program in Chile, we only include the paper with the lowest risk of selection bias in the meta-analysis (Yoshikawa et al., 2015), which is consistent with the approach followed by Brody et al. (2016). In our risk of bias assessment we include 26 evaluations, however. These evaluations include each of the three papers that estimate the impact of the same program and the 23 other unique programs that were evaluated.

The studies were diverse in terms of program characteristics, outcome measures, sample size, evaluation design, context, and analysis type. The estimated average effect size in a meta-analysis should be interpreted cautiously because of these differences. For this reason, we will place an equal weight on the narrative synthesis of the experimental and quasi-experimental studies as on the meta-analysis. This equal emphasis is in line with recommendations by Waddington et al. (2012) in their toolkit for systematic reviews in international development. They argue that the limitations of meta-analyses should be acknowledged when reviews are broad in scope.

Below we present the main characteristics of the included studies including program characteristics, outcome measures, sample size, study design, and analysis. Table 5 summarizes the characteristics of the evaluations.

Program Characteristics

This section presents the program characteristics of the 24 included experimental and quasi-experimental quantitative evaluations of the evaluated programs. The included studies focused on programs with an emphasis on teacher training, nutrition, ICT, preschools, specific teacher practices inside and outside the classroom, and specific parental practices outside the classroom. Of the included evaluations, five focused on the impact of specific teacher practices, such as reading aloud or the explicit instruction of new words (Larrain, Strasser, & Rosa Lissi, 2012; Neugebauer & Currie-Rubin, 2009; Vivas, 1996); three focused on the effects of parental involvement, for example by paired reading (Murad & Topping, 2000; Tapia & Benitez, 2013; Vivas, 1996); five examined the impact of programs with an emphasis on nutrition, such as food supplements and school meals (Adrogue & Orlicki, 2013; Ismail et al., 2014; Maluccio et al., 2009; Powell et al., 1998; Simeon, Grantham-McGregor, & Wong, 1995); four estimated the impact of ICT programs, such as the distribution of laptops to children and computer-aided instruction (Beuermann et al., 2015; Cristia et al., 2012; Ferrando et al., 2011; Osorio & Linden, 2009); two focused on the impact of preschool (Campos et al., 2011; Felicio, Terra, & Zoghbi, 2011); four estimated the impact of teacher training (Gomez Franco, 2015; Mendive et al., 2016; Pallante et al., 2015; Yoshikawa et al., 2015); and two emphasized the effects of school governance reforms (Bando, 2010; Lockheed, Harris, & Jayasundera, 2010).

Each of these programs may result in improvements in reading outcomes through different pathways. For example, teacher training programs may lead to improvements in teacher knowledge, which may in turn influence teacher practices, which could then result in improvements in reading outcomes. At the same time, nutrition programs may lead to improvements in children's dietary diversity and food security, which could in turn increase the concentration levels and reading outcomes of children. The different programs may also be complementary to each other. For example, distributing laptops to children is only likely to result in improvements in reading outcomes if teacher practices are appropriate. These interactions are taken into consideration in the interpretation of our results. However, we will start with a separate analysis for each of the different programs before we triangulate the findings.

Outcome Measures

This section presents the outcome measures that were used to determine the impact of the programs in the experimental and quasi-experimental studies. The included studies estimate the impact of programs on outcome measures such as reading comprehension, reading fluency, letter naming, word recognition, phonemic segmentation fluency, decoding, spelling, language test scores, and national literacy exam test scores. Two other studies focused on more intermediate outcomes such as reading practices (Beuermann et al., 2015; Tapia & Benitez, 2013).

Each of the outcome measures can be considered part of a different construct. Reading is a broad concept that can be subdivided into many different constructs. Authors of primary studies use a large number of different operational definitions to measure reading outcomes and practices. Some

studies construct indices based on different elements of reading outcomes, while others are more specific in their definition of reading outcomes or practices. Both approaches have their advantages. Relying on an index addresses the so-called "indicator soup" problem, which refers to the difficulty of organizing and interpreting results with many outcome variables (King, Samii, & Snilstveit, 2010). However, the construction of indices can also be accompanied by a loss of detail, for example when interventions have positive effects on decoding, but not on language comprehension.

To mitigate these concerns, we planned to use an iterative approach. We proposed to synthesize the evidence on what works to improve early grade reading outcomes by conducting two types of analyses. The first analysis would first pool all studies that include an outcome measure related to reading outcomes regardless of the specifics of the construct (except for reading practices). The second analysis would then examine the impact of the included programs on different components of reading outcomes, such as decoding, letter recognition, and reading comprehension.

Importantly, however, we were limited in our ability to conduct the second analysis because in several cases it was not entirely clear from the study report whether outcome measures should be considered a decoding, vocabulary acquisition, or a reading comprehension construct. Thus, in practice we only conducted a narrative review to determine the impact of the programs on specific components of reading outcomes. In some cases this narrative review was limited to analyzing the results of only one study because we did not encounter more than one study that focused on that specific reading construct.

Although the majority of the included studies only emphasized one outcome measure related to early grade reading, several studies included more than one outcome measure. Of the 24 program evaluations, 15 included only one outcome measure. Furthermore, of the 24 evaluations, eight evaluations relied on a language test score to measure the impact of the program, five evaluations assessed the impact of the program on reading comprehension, four determined the impact on vocabulary acquisition, two studies focused on early literacy or letter naming, and two evaluations emphasized the impact of the program on reading practices. Other outcome measures that were included in at least one study were word reading, phonemic segmentation, decoding, spelling, English language test scores, and an undetermined measure of literacy outcomes.

Some studies relied on existing or administrative data to determine the impact of the program, while others collected their own reading outcome data. Specifically, of the included studies, 12 studies relied exclusively on existing or administrative data to determine the impact of the program, while the remaining studies collected their own data. Unfortunately, none of the studies presented details about how the assessment test was aligned with the evaluated program so we were not able to assess over-alignment of the assessment test with the program design. It is important to note that the studies that relied on existing or administrative data had a much larger average sample size than the studies that collected their own data. We discuss the sample size of the included studies in more detail below following a discussion about the context in which the studies took place.

Context

The included experimental and quasi-experimental studies focused on a wide range of countries in Latin America and the Caribbean, but high-income economies are overrepresented considering the low number of high-income economies in Latin America and the Caribbean. Of the 24 included evaluations, eight focused on high-income economies, 14 focused on upper-middle-income countries, and only 2 focused on lower-middle-income economies. Among the high-income economies, Chile is particularly overrepresented, as four of the included experimental and quasi-experimental evaluations focused on Chile. In addition to these four studies, we found four evaluations with an emphasis on Brazil, three evaluations that focused on Peru and Jamaica, two evaluations that focused on Argentina, Venezuela, and Mexico, and individual evaluations that focused on Guatemala, Guyana, and Colombia. These findings indicate that experimental and quasi-experimental studies with an emphasis on reading outcomes focus disproportionally on higher-income economies and provide too little attention to low-income economies. Thus, we may not be able to extrapolate the findings of our synthesis of the quantitative intervention studies to low-income economies.

We hypothesized that the disproportionate emphasis on high-income economies may be associated with the limited resources and capacity of researchers in middle-income economies to conduct experimental and quasi-experimental evaluations. We examined the number of studies that were authored or co-authored by in-country researchers and we found some evidence that supports this hypothesis. Specifically, we found that the majority of the studies in high-income and upper-middle-income countries were implemented by researchers that were based in the countries of interest, while the majority of the studies in the lower-middle-income countries were conducted by researchers based in the United States.

Evaluation Design

To be included in this report, the quantitative studies needed to focus on program effectiveness by relying on either an experimental or a quasi-experimental design to determine the impact of the program of interest. The study designs of the included studies were diverse. Of the 24 included program evaluations, 16 relied on a randomized controlled trial to determine the impact of the programs. Of these 16 evaluations, seven used a cluster randomized controlled trial where the program was implemented at the school-level as opposed to the student-level. Of the eight remaining studies, five used propensity score matching designs and three used multivariate regression analyses to determine the impact of the evaluated programs on reading outcomes. Cluster randomized controlled trials are the strongest design for making causal claims about the impact of education programs, but under certain conditions, student-level randomized controlled trials or quasi-experimental designs can also determine causal effects. We will discuss these conditions in more detail in our critical appraisal of the quality of the studies.

Sample Size

The included evaluations were diverse in terms of their sample size. Of the 24 included evaluations, six studies relied on a sample size <100 students to determine the impact of the program, seven studies determined the impact of the programs with a sample size <1000 students, and 11 studies focused on a sample size of more than 1,000 students.

The majority of the small sample studies focused on a program that was implemented for academic purposes, while the studies with sample sizes over 1000 students often focused on the effectiveness of government-supported programs. Of the 24 included evaluations, 9 focused on the effectiveness of programs that were implemented for academic purposes. These studies usually serve to examine the mechanisms that influence how development programs can influence reading outcomes. Of the 15 remaining studies, 13 estimated the impact of government-supported programs on reading outcomes or practices. These studies serve to assess whether development programs implemented at scale are effective in improving reading outcomes. Both types of studies are important for different purposes and each study type suffers from different biases. Studies that are implemented for academic purposes often rely on a sample size that is too small to determine small or medium but meaningful effects of an intervention with sufficient precision. Large-scale government-supported programs often require less rigorous quasi-experimental designs (as opposed to cluster randomized controlled trials) and usually have to rely on data from national exams to determine the impact of reading programs. We will examine these limitations in more detail in the section on the critical appraisal of the studies.

Table 5. Summary of Quantitative Intervention Studies

Study title	Authors (year)	Location researcher	Implementer	Context	Outcomes	Sample	Study design	Intervention/ program	Analysis
Lectura compartida de cuentos y aprendizaje de vocabulario en edad preescolar: un estudio de eficacia	Larrain, Strasser, & Lissi (2012) Experiment 1	Chile	Academic	Santiago, Chile	Vocabulary acquisition	112 children from 3 public kindergartens	RCT	Shared book reading without word elaboration	T-test
Lectura compartida de cuentos y aprendizaje de vocabulario en edad preescolar: un estudio de eficacia	Larrain, Strasser, & Lissi (2012) Experiment 2	Chile	Academic	Santiago, Chile	Vocabulary acquisition	62 children from 3 public kindergartens	RCT	Shared book reading with word elaboration	T-test
Desarrollo de Habilidades Conductuales Maternas Para Promover la Alfabetización Inicial en Niños Prescolares	Tapia & Benitez, (2013)	Mexico	Academic	Mexico	Literacy practices in mothers	20 women with limited literacy practices whose preschool children showed low levels in pre-academic and linguistic skills	RCT	Training for mothers to conduct activities and strategies to promote language and pre-academic skills related to early literacy in preschool children using joint reading of stories and puppet play	ANOVA
The Effects of Early Childhood Education on Literacy Scores Using Data from a New Brazilian Assessment Tool	Felicio, Terra, & Zoghbi, (2011)	Brazil	Ministry of Education	Brazil	Literacy scores	1986 second grade students	Quasi-experimental	Enrolment in preschool	Propensity score matching

Study title	Authors (year)	Location researcher	Implementer	Context	Outcomes	Sample	Study design	Intervention/ program	Analysis
Exploring Teachers' Read-Aloud Practices as Predictors of Children's Language Skills: The Case of Low Income Chilean Preschool Classrooms	Gomez Franco, (2014)	Chile	Ministry of Education	Chile	Teachers' speech characteristics Teachers' read-aloud strategies	913 students across 47 preschools	RCT	Teacher training program for teachers in preschool	ANOVA
The Impact of Improving Nutrition During Early Childhood on Education Among Guatemalan Adults	Maluccio et al. (2009)	United States	Institute of Nutrition of Central America and Panama	Guatemala	Grades completed Reading comprehension Nonverbal cognitive skills	1,471 adults from four villages	RCT	Provision of nutrient dense drink to children	OLS regression analysis
Parents as Reading Tutors for First Graders in Brazil	Murad & Topping (2000)	United States	Academic	Brazil	Reading comprehension Reading fluency	48 students from a single school	RCT	Training for parents to participate in paired reading with their children	Bivariate comparison
The Effect of a Multicomponent Literacy Instruction Model on Literacy Growth for Kindergartners and First-Grade Students in Chile	Pallante & Kim (2013)	Chile	Collaborative Language and Literacy Instruction Project	Chile	Letter naming Word reading Vocabulary Phonemic segmentation fluency	312 kindergartners; 305 first graders	Cluster RCT	Teacher training during 5 workshops for teachers complemented with teaching resources	Difference-in-difference regression analysis

Study title	Authors (year)	Location researcher	Implementer	Context	Outcomes	Sample	Study design	Intervention/ program	Analysis
Letter Names and Phonological Awareness Help Children to Learn Letter–Sound Relations	Cardoso-Martin et al. (2011) Experiment 1	Brazil	Academic	Brazil	Letter sound learning Decoding	32 children/ 20 children	RCT	Phonological training for children about the shapes and names of letters	ANOVA
Letter Names and Phonological Awareness Help Children to Learn Letter–Sound Relations	Cardoso-Martin et al. (2011) Experiment 2	Brazil	Academic	Brazil	Letter sound learning Decoding	10 children/ 10 children	RCT	Phonological training for children about the shapes and names of letters with extra emphasis on beginning and middle sound letters	ANOVA
Treatment of *Trichuris trichiura* Infections Improves Growth, Spelling Scores and School Attendance in Some Children	Simeon et al. (1995)	United States	Academic	Jamaica	Spelling Reading	407 children	RCT	Deworming program	OLS regression analysis
Do In-School Feeding Programs Have Impact on Academic Performance and Dropouts? The Case of Public Argentine Schools	Adrogue & Orlicki (2013)	United States	Ministry of Education	Argentina	Language test score	4,466 schools	Difference-in-difference	School feeding program	OLS regression analysis

Study title	Authors (year)	Location researcher	Implementer	Context	Outcomes	Sample	Study design	Intervention/ program	Analysis
Technology and Child Development: Evidence from the One Laptop Per Child Program	Cristia et al. (2012)	Peru	Ministry of Education	Peru	Language test scores	4,111 students	RCT	Provision of 1 laptop per child	OLS regression analysis
Nutrition and Education: A Randomized Trial of the Effects of Breakfast in Rural Primary School Children	Powell et al. (1998)	United States	Academic	Jamaica	Spelling Reading	814 students	RCT	Provision of nutritious breakfast by schools	OLS
Learning with the XOs: The Impact of the Ceibal Plan	Ferrando et al. (2011)	Uruguay	Ministry of Education	Uruguay	Reading	1,365 students	Difference-in-difference	Provision of 1 laptop per child	Propensity score matching
Guyana's Hinterland Community-Based School Feeding Programme	Ismail et al. (2012)	United States	Ministry of Education	Guyana	English language test scores Reading test scores	410 treatment and 783 control students	Quasi-experimental	Community-based school feeding program.	Propensity score matching
The Contribution of Quality Early Childhood Education and its Impacts on the Beginning of Fundamental Education	Campos et al. (2011)	Brazil	Ministry of Education	Brazil	National literacy exam test scores	605 treatment and 157 comparison students	Nonexperimental	Preschool	Hierarchical regression analysis

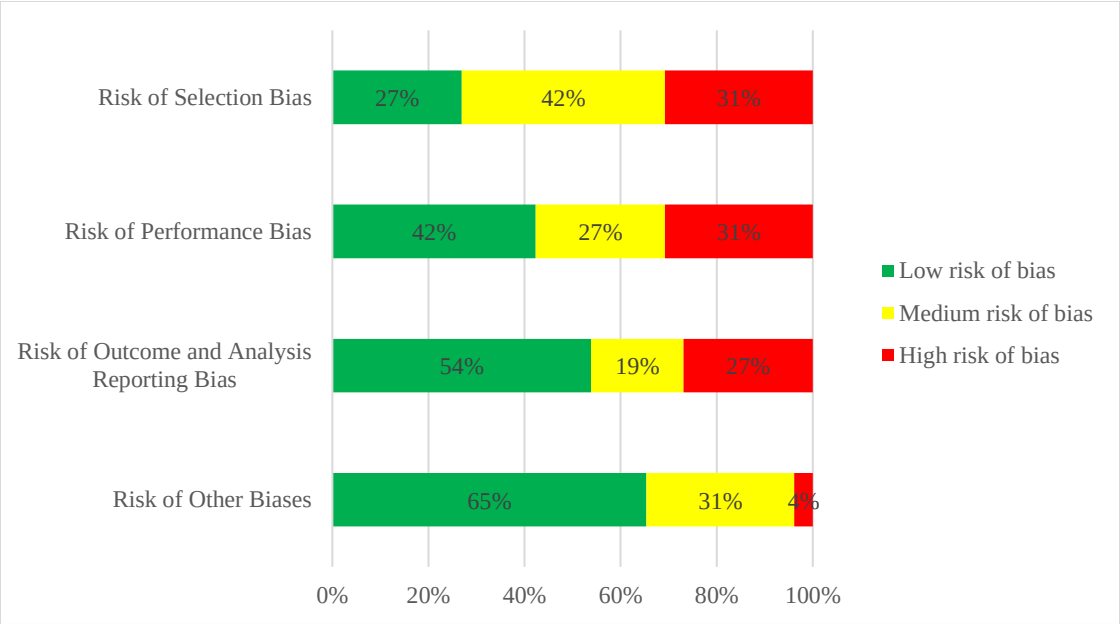
Study title	Authors (year)	Location researcher	Implementer	Context	Outcomes	Sample	Study design	Intervention/ program	Analysis
School Improvement Plans and Student Learning in Jamaica	Lockheed et al. (2006)	United States	Ministry of Education	Jamaica	National literacy and language school-level performance data	71 treatment and 67 comparison schools	Quasi-experimental	The program provided support to schools on the basis of needs identified through the preparation of a School Development Plan	Propensity score matching
Opening the Black Box: Intervention Fidelity in a Randomized Trial of a Preschool Teacher Professional Development Program	Mendive et al. (2016)	Chile	Ministry of Education	Chile	Child language and literacy	1,033 treatment students and 843 control students	RCT	Teacher training program	OLS regression analysis
Effects of Story Reading on Language	Vivas (1996) Experiment 1	Venezuela	Academia	Venezuela	Language comprehension and expressive language	72 treatment and 84 control students	RCT	Systematic, story-reading-aloud program at school	ANOVA
Effects of Story Reading on Language	Vivas (1996) Experiment 2	Venezuela	Academia	Venezuela	Language comprehension and expressive language	72 treatment students and 66 control students	RCT	Systematic, story-reading-aloud program at home	ANOVA
The Use and Misuse of Computers in Education: Evidence from a Randomized Experiment in Colombia	Osorio & Linden (2009)	United States	Ministry of Communication	Colombia	Language test scores	4,327 treatment students and 3,889 control students	RCT	Provision of computers for instruction	OLS regression analysis

Study title	Authors (year)	Location researcher	Implementer	Context	Outcomes	Sample	Study design	Intervention/ program	Analysis
Experimental Impacts of a Teacher Professional Development Program in Chile on Preschool Classroom Quality and Child Outcomes	Yoshikawa et al. (2015)	United States	Ministry of Education	Chile	Child language and literacy	1,033 treatment students and 843 control students	RCT	Teacher training program	OLS regression analysis
Home Computers and Child Outcomes: Short-Term Impacts from a Randomized Experiment in Peru	Beuermann et al. (2015)	United States	Ministry of Education	Peru	Time spent reading Language Test Scores	Total sample of 2,817	Difference-in-difference	Provision of educational laptop	OLS
Read-Alouds in Calca, Peru: A Bilingual Indigenous Context	Neugebauer & Currie-Rubin (2009)	United States	Academia	Peru	Vocabulary acquisition Reading comprehension	29 treatment students and 26 control students in 2 treatment classrooms and 2 control classrooms	RCT	Read-aloud program	OLS

Critical Appraisal

For our critical appraisal of the studies, we relied on a risk of bias assessment tool with 71 questions with which we could accurately determine four types of risk of bias. The tool is an adapted version of a risk of bias assessment tool developed by Hombrados and Waddington (2012). We examined the risk of selection bias and confounding, performance bias, outcome and analysis reporting bias, and other biases. The complete risk of bias assessment tool and a detailed assessment of the risk of bias of each individual study are included in Appendix C. Figure 6 shows the distribution of low-, medium-, and high-risk bias across the included studies for each of the risk of bias categories.

Figure 6. Risk of Bias Assessment of Quantitative Intervention Studies



In general, there was agreement among the reviewers concerning assessments of the risk of selection bias, but initially there were more disagreements about the risk of performance bias, outcome and analysis reporting bias, and other biases. We reached consensus after a detailed discussion about each of the individual studies.

Selection Bias and Confounding

Selection bias is associated with lack of equivalence in observable or unobservable characteristics across treatment and control/comparison groups. Selection bias may result from self-selection into the program, which could lead to differences between students who participate in the program and students who do not participate in the program or targeting of a program to schools or students with specific characteristics. Self-selection may result in differences in unobservable characteristics because participants in development programs are usually more motivated or entrepreneurial (Waddington et al., 2012). The targeting of a program to schools or students with specific characteristics by an implementing agency is more likely to result in differences in observable characteristics. Quasi-experimental methods such as propensity score matching are

usually a good alternative to randomized controlled trials when a program is targeted to specific students or schools because in those cases it remains feasible to control for observable characteristics in the estimation of the impact of the program (Diaz & Handa, 2006). However, quasi-experimental methods such as propensity score matching usually do not allow for resolving selection bias when selection bias is caused by self-selection because propensity score matching does not enable researchers to control for unobservable characteristics.

Of the 26 included studies, seven were rated as having a low risk of selection bias, five were rated as having a medium risk of selection bias, and eight were rated as having a high risk of selection bias. The nine studies with a low risk of selection bias were all cluster RCTs with a sufficient sample size to detect small but meaningful effects of the evaluated program on reading outcomes. For example, Cristia et al. (2012) used an RCT, in which 160 schools in Peru were randomly assigned to a program where each student received a laptop. The study relied on national test score data for more than 4,000 students. Similarly, Osorio and Linden (2009) used a cluster RCT with a sample of 5,201 students across 97 schools in Colombia to determine the impact of a program that distributed computers to support education.

We rated RCTs with a small sample size and quasi-experimental evaluations that used propensity score matching with a large sample as having a medium risk of selection bias. RCTs with a small sample size may suffer from lack of equivalence across the treatment and the control group because randomization requires a sufficient number of units of observation to guarantee equivalence across observable and unobservable characteristics. For example, Larrain, Strasser, and Lissi (2012) relied on a sample size of 62 children from three public kindergartens to determine the impact of more complex word elaboration on vocabulary acquisition. Such sample sizes are usually not sufficient to detect small but meaningful effects of a program on reading outcomes. Furthermore, the likelihood of publication bias is higher for studies with such low sample sizes because it is more likely that studies with such small sample sizes and statistically insignificant effects are not accepted for publication in peer-reviewed journals (Borenstein et al., 2009). As a result, the inclusion of studies with small sample sizes may result in an overestimate of the impact of development programs on reading outcomes. The majority of the included RCTs with a small sample size also only showed limited or no baseline data to demonstrate equivalence in observable characteristics. For example, Larrain et al. (2012) did not show baseline values for the beneficiary and control students. Furthermore, Murad and Topping (2000) only showed evidence for nonsignificant differences at baseline, but did not present the actual values of the baseline data.

We rated studies that relied on propensity score matching and a large sample size as having a medium risk of selection bias because propensity score matching does not enable researchers to entirely control for self-selection. The quasi-experimental studies we included did involve some self-selection in all cases. For example, De Felicio, Terra, and Zoghbi (2011) relied on propensity score matching to determine the impact of preschool on early grade reading outcomes in Brazil. However, participation in preschool is entirely dependent on self-selection, so the use of propensity score matching does usually not allow for demonstrating causal effects of participation in preschool in these specific cases.

Finally, we rated RCTs with a very small sample size and problems in the implementation of the randomization and nonexperimental studies that relied on ordinary least squares (OLS) regression analysis without a baseline as having a high risk of selection bias. Problems in the implementation

of the randomization included control students that switched to the treatment group (crossovers), deliberate exclusion of part of the sample that did not comply with the randomization, and too high or unknown attrition rates. For example, Gomez Franco (2015) excluded teachers who did not comply with the instructions provided during teacher training from his analysis on the impact of a teacher training program for preschool teachers. The exclusion of these teachers from the analysis is likely to result in significant overestimates of the impact of the program. Rugerio Tapia et al. (2013) also relied on a sample of 10 beneficiary mothers and 10 control mothers to determine the impact of a program that encourages mothers to jointly read with their children. This sample size is likely to result in lack of equivalence across beneficiary and control mothers. Mendive et al. (2015) determined the impact of a preschool professional development program for teachers by relying on a sample with attrition rates over 50%. Such attrition rates are very likely to result in selection bias as well due to lack of equivalence across beneficiary and control students. OLS regression analysis without a baseline also does not allow for addressing selection bias. Thus, these studies should be considered as having a high risk of selection bias. For example, Campos et al. (2011) used hierarchical regression analysis to determine the impact of participation in preschool on early grade reading outcomes in Brazil. The use of hierarchical regression analysis does not enable researchers to control for bias from unobservable characteristics and is thus likely to result in biased impact estimates.

Performance Bias

Performance bias refers to bias that results from spillovers or contamination. Spillovers are indirect benefits of the program that result from interaction with the treatment group. These indirect benefits may in turn result in underestimates of the impact of the program if they are not taken into consideration in the analysis. For example, Miguel and Kremer (2004) found evidence that the effects of deworming on school enrolment were considerably underestimated when control students interacted closely with treatment students because control students are less likely to be infected with intestinal worms if they interact with dewormed treatment students. Similarly, control students may be positively affected by a program if beneficiary students help them with their homework. Contamination refers to benefits for the control group because of the unintentional assignment of the program to the control group. For example, on the ground program implementers may not know about the random assignment of schools to a program and as a result start implementing the program in the control schools. Spillovers and contamination are less likely when the assignment of the program happens at the school level. In those cases, the likelihood of interaction between treatment students and control students is lower than when treatment and control students come from the same school. Furthermore, program implementers are also less likely to make mistakes in the allocation of benefits when program assignment is at the school level than when program assignment is at the classroom or student level.

Of the 26 included evaluations, 11 studies were rated as having a low risk of performance bias, seven studies were rated as having a medium risk of performance bias, and eight studies were rated as having a high risk of performance bias. We rated studies that relied on comparisons between students in schools and found no evidence or only marginal evidence for contamination of the control group as low risk of performance bias. For example, Adrogue and Orlicki (2013) used a difference-in-difference analysis to identify the impact of an in-school feeding program on reading outcomes in Argentina. Their comparison across schools is not likely to suffer from bias due to

spillovers or contamination because there is no evidence of interaction between the beneficiary and comparison students.

We rated studies that relied on comparisons across students in different classrooms but within the same school and studies that found some evidence for contamination of the control or comparison group as having a medium risk of bias. For example, Murad and Topping (2000) used a sample where the beneficiary and control students came from the same school. In this case, there is a risk of spillovers because of the possibility of interaction between the beneficiary and the comparison students. This interaction may in turn result in indirect benefits for the comparison students, which could lead to underestimates of the impact of the program.

Finally, we rated studies that relied on comparisons between students in the same classroom and studies that found major evidence for contamination of the control group as having a high risk of performance bias. For example, one study randomly assigned students in the same classroom to a school breakfast program without taking into consideration the likely option of sharing food between students (Powell et al., 1998). In this case, the risk of contamination was considered high because of a high likelihood of food sharing. This contamination could then result in underestimates of the impact of the program.

Outcome and Analysis Reporting Bias

Outcome and analysis reporting bias refers to bias that results from the failure to report certain (usually nonsignificant) results and the use of unusual or incorrect methods of analysis. The failure to report specific results may indicate evidence for publication bias. For example, researchers may have incentives to only report statistically significant results and fail to report results that are not statistically significant. This failure to report results may lead researchers to overestimate the impact of programs on reading outcomes because the meta-analysis may only include statistically significant results. Unusual estimation methods may also be an indication for outcome and analysis reporting bias. For example, researchers may choose arbitrary thresholds to ensure that results become statistically significant. Alternatively, researchers may also choose to include certain control variables and exclude other control variables to ensure that results are statistically significant. Finally, incorrect estimation methods may also result in a bias in the impact estimates. For example, researchers may choose to include potentially endogenous control variables, which may result in a bias in the impact estimates.

Of the 26 included studies, we rated 14 studies as having a low risk of outcome and analysis reporting bias, five studies as having a medium risk of outcome and analysis reporting bias, and seven studies as having a high risk of outcome analysis reporting bias. Specifically, studies that reported impact estimates on all relevant outcome variables associated with reading and used appropriate estimation methods were rated as having a low risk of outcome and analysis reporting bias. For example, Pallante and Kim (2013) report impact estimates on letter naming, word recognition, vocabulary acquisition, and phonemic segmentation. This wide range of outcome measures indicates that the authors did not selectively report the impact of the program on outcome measures where they found statistically significant effects.

Studies that were selective in their reporting of heterogeneous effect were rated as having a medium risk of outcome and analysis reporting bias. For example, Simon et al. (1995) only

reported positive and statistically significant heterogeneous effects of deworming on spelling outcomes.They did not report heterogeneous effects on reading outcomes, possibly because the results were not statistically significant. Nonetheless, the authors did present average impacts on all of the included outcome measures regardless of the statistical significance of the results. Similarly, Neugebauer and Currie-Rubin (2009) only presented impact estimates on an assessment test they developed themselves but not on a standardized assessment test.

Finally, studies that did not report nonsignificant impact estimates (even if they informally reported the lack of significance for these outcome variables in the text), studies that used arbitrary thresholds to determine the treatment status of certain students, and studies that switched control students to the control group when they did not comply with the program recommendations or program were rated as having a high risk of outcome and analysis reporting bias. For example, Mendive et al. (2015) used an arbitrary threshold to determine whether teachers were successfully implementing teacher practices following a teacher training program. They reported statistically significant effects of the compliance with appropriate teacher practices on reading outcomes. However, it remains unclear whether the results of the study were robust to the use of alternative thresholds. De Felicio et al. (2011) also reported only statistically significant effects of participation in preschool on reading outcomes, while they downplayed nonsignificant effects as irrelevant.

Other Biases

Other biases may include courtesy and social desirability bias, Hawthorne and John Hendry Effects, the inclusion of outcome variables that are not validated in the context of Latin America and the Caribbean, strong researcher involvement in the implementation of the program, and a failure to cluster standard errors when the program is assigned at a unit of intervention above the measurement level. Courtesy bias refers to a situation where the respondent gives the answer that he or she feels the interviewer wants to hear. Social desirability bias refers to a situation where the respondent gives the answer he or she believes is considered the socially correct answer. Self-reported data tend to suffer from courtesy and social desirability bias (White & Phillips, 2012). Hawthorne effects refer to a bias that results from extra motivation for the treatment group because the beneficiaries know that they are part of the treatment group while John Henry effect refers to the opposite effect, where control students are motivated to catch up with the treatment group. Bias may also result from the use of outcome variables that are not validated in the context of Latin America. For example, researchers may use tests that are contextually appropriate for the United States but not for the Latin American context.

Strong researcher involvement in the implementation of the program may result in a better or worse implementation of the program than should be expected when the program is implemented at scale. In addition, strong researcher involvement may increase the likelihood of the Hawthorne effect. Finally, a failure to cluster standard errors when that is considered appropriate, such as in cluster RCTs, may result in conclusions that are too optimistic about the statistical significance of the effects of development programs on reading outcomes.

Of the 26 included studies, we rated 17 studies as having a low risk of other biases, eight studies as having a medium risk of other biases, and one studies as having a high risk of other biases.

Studies that did not appear to suffer from any of the other biases mentioned above were rated as having a low risk of other bias.

Studies that experienced one (and only one) of the problems discussed above were rated as having a medium risk of bias. For example, Vivas (1996) did not account for clustering of the standard errors in the impact estimates of a story-reading-aloud program on reading outcomes in Venezuela. As a result, the study may have overestimated the statistical significance of the impact estimates. In another example, Mendive et al. (2015) used videos to measure the behavior of teachers but did not take into consideration the option that teachers may have changed their behavior due to the videos. This Hawthorne effect could have resulted in a bias in the impact estimates.

Finally, studies that suffered from more than one of the other biases discussed above were rated as having a high risk of other biases. These studies are likely to be biased because they suffer from more than one other methodological problem. For example, Gomez Franco (2015) did not account for clustering of the standard errors in the impact estimates of a teacher training program for teachers in preschool in Chile. Furthermore, the impact estimates presented in this study may also be biased due to the use of videos to measure teacher behavior.

Synthesis of Quantitative Studies

This section presents results from the meta-analysis and narrative review of the effects of different types of programs on reading outcomes. We present a separate analysis for each of the program types that were evaluated in the primary studies, including teacher training programs, ICT programs, nutrition programs, school governance programs, programs with an emphasis on teacher practices, and programs with an emphasis on parental involvement.

To synthesize the findings for each intervention type, we first conducted a meta-analysis for each of the RCTs, followed by a meta-analysis for each of the nonexperimental studies and an assessment of whether RCTs and nonexperimental studies can be credibly pooled in one meta-analysis. However, we were not able to adopt this approach in all cases because of the relatively small number of included studies.

Teacher Training Programs

Of the included studies, four presented an estimate of the impact of teacher training programs on reading outcomes. Of these studies, we were able to include two studies in our meta-analysis (Pallante et al., 2015; Yoshikawa et al., 2015). We did not include the other two studies because they evaluated the same program in Chile (Gomez Franco et al., 2015; Mendive et al., 2016) as Yoshikawa et al. (2015) and were rated as having a higher risk of selection bias. We summarize the evaluations that focused on the impact of teacher training in Table 6. This table also summarizes the outcome measures and the evaluation design that were used in the primary study. Despite the small number of studies, we still include a meta-analysis on the effects of teacher training programs on reading outcomes because both studies are RCTs with a low risk of selection bias in a very similar context.

Table 6. Primary Studies That Focus on the Impact of Teacher Training

Study	Definition of outcome variable(s)	Evaluation design	Included in meta-analysis?	Country
Gomez Franco, (2014)	Vocabulary acquisition Reading comprehension	Cluster RCT	No	Chile
Mendive et al. (2016)	Language test score Early literacy outcomes	Cluster RCT	No	Chile
Pallante & Kim (2013)	Letter naming Word reading Vocabulary acquisition Phonemic segmentation	Cluster RCT	Yes	Chile
Yoshikawa et al. (2015)	Language test score Early literacy outcomes	Cluster RCT	Yes	Chile

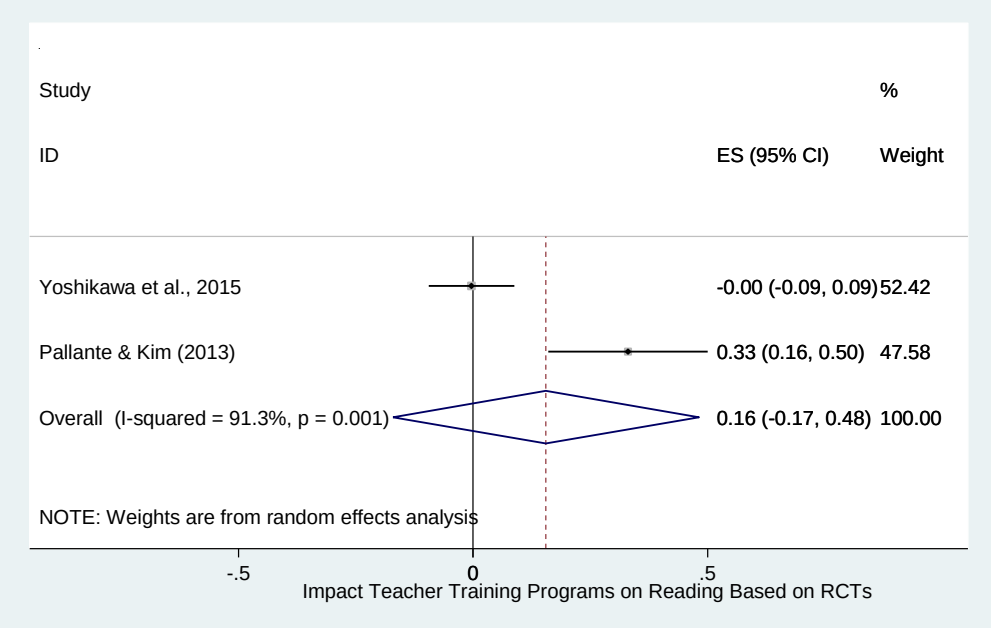
Meta-Analysis for Randomized Controlled Trials. The results of the meta-analysis for the RCTs are presented in Figure 7. We found no evidence that, on average, teacher training had a positive effect on reading outcomes (SMD = 0.16, 95% confidence interval (CI) = -0.17, 0.48; evidence from two studies). However, Pallante and Kim (2015) found a medium-sized, positive, and statistically significant effect on the reading outcomes of students in kindergarten and first grade in their evaluation of a teacher training program in Chile that targets phonological awareness, alphabetics and phonics, fluency, vocabulary, reading comprehension, and writing. This study focused on a comprehensive teacher training program that included a focus on coaching and sustained follow-up. In contrast, Yoshikawa et al. (2015) did not find positive effects of a teacher training program for teachers in prekindergarten classrooms in Chile. They did find positive impacts for emotional and instructional support of teachers, but the results suggested that these behavioral changes did not translate to positive effects on early grade reading outcomes. Mendive et al. (2016) demonstrated that the lack of positive effects on reading outcomes in Yoshikawa et al. (2015) may have resulted from problems in the implementation of the program. It is possible that teacher training programs need to be comprehensive and complemented by coaching and sustained follow-up in order to have positive impacts on reading outcomes. The coaching and sustained follow-up could result in improvements in the fidelity of implementation.

At the same time, however, we need to be careful in how we interpret the results because we only encountered two studies, which were both implemented in Chile. The effects of teacher training programs may well be different in a more representative sample of evaluations of teacher training programs. The results of our meta-analysis may not be externally valid and it is possible that the results cannot be extrapolated to the rest of the LAC region. We also do not interpret the heterogeneity in the effect sizes because of the small number of studies. Nonetheless, the results present indications that teacher training programs can be effective in improving early grade reading outcomes if they are implemented with fidelity and if teacher training programs are complemented by coaching.

We were not able to conduct a stratified meta-analysis by methodology or risk of bias because of the relatively small number of studies that focused on the impact of teacher training. However, both Pallante and Kim (2015) and Yoshikawa et al. (2015) present relatively strong study designs with a high internal validity. Both the risk of selection bias and the risk of performance bias were

low in these studies. Considering the quality of the studies, we are confident that our results apply to the context of Chile.

Figure 7. Impact of Teacher Training Programs on Reading Outcomes



ICT Programs

Of the 24 included studies, four estimated the impact of an ICT program on reading outcomes. We were able to include all of these studies in our meta-analysis. The evaluations that focused on the impact of ICT programs are summarized in Table 7.

Table 7. Primary Studies That Focus on the Impact of ICT

Study	Definition of variable	Evaluation design	Included in meta-analysis?	Country
Cristia et al. (2012)	Language test score	Cluster RCT	Yes	Peru
Ferrando et al. (2011)	Reading comprehension	Propensity score matching	Yes	Uruguay
Osorio & Linden (2009)	Language test score	Cluster RCT	Yes	Colombia
Beuermann et al. (2015)	Reading practices	Cluster RCT	Yes	Peru

Randomized Controlled Trials.

Figure 8 includes the results of the meta-analysis for the RCTs of ICT programs. We found no evidence to indicate that, on average, ICT programs had a positive effect on reading outcomes (SMD = 0.03, 95% confidence interval (CI) = -0.13, 0.19; evidence from three studies). The results of the one laptop per child program are particularly worrisome. The findings of Cristia et al. (2012)

suggest that the nationwide one laptop per child program had negative effects on early grade reading outcomes in Peru and may have resulted in adverse effects on the reading habits of children. Beuermann et al. (2015) showed evidence for negative but nonsignificant point estimates in their estimates of the impact of the program on the number of hours that children allocated to reading books in a smaller sample in Lima, Peru. A separate meta-analysis that focused on the impact of the one laptop per child program (see Figure 9) did not find evidence for statistically significant and negative effects of the program on reading outcomes if the sample was restricted to RCTs (SMD = -0.04, 95% confidence interval (CI) = -0.16, 0.08; evidence from two studies). In any case, it is important to be cautious when interpreting these results because we may not be able to extrapolate the results to outside Peru.

By contrast, Osorio and Linden (2009 found that a computer distribution program had a medium-sized, positive, and statistically significant effect on the reading outcomes of third grade students in Colombia. Interestingly, Osorio and Linden (2009) also found considerable evidence for challenges in implementing this program. In many cases, teachers did not use the computers in their instruction methods. Perhaps as a result, Osorio and Linden (2009) did not find any positive effects of the program in their full sample of students (third through 12th grade). However, our effect size calculations indicate that the distribution of computers to support computer-aided instruction may have positive effects on early grade reading outcomes for third grade students even in the presence of implementation problems. This finding indicates that the effects may be even larger when implementation problems can be prevented. Nonetheless, we only found one rigorous study that focused on the distribution of computers, so the results may not be externally valid outside Colombia.

Figure 8. Impact of ICT Program on Reading Outcomes on the Basis of RCTs

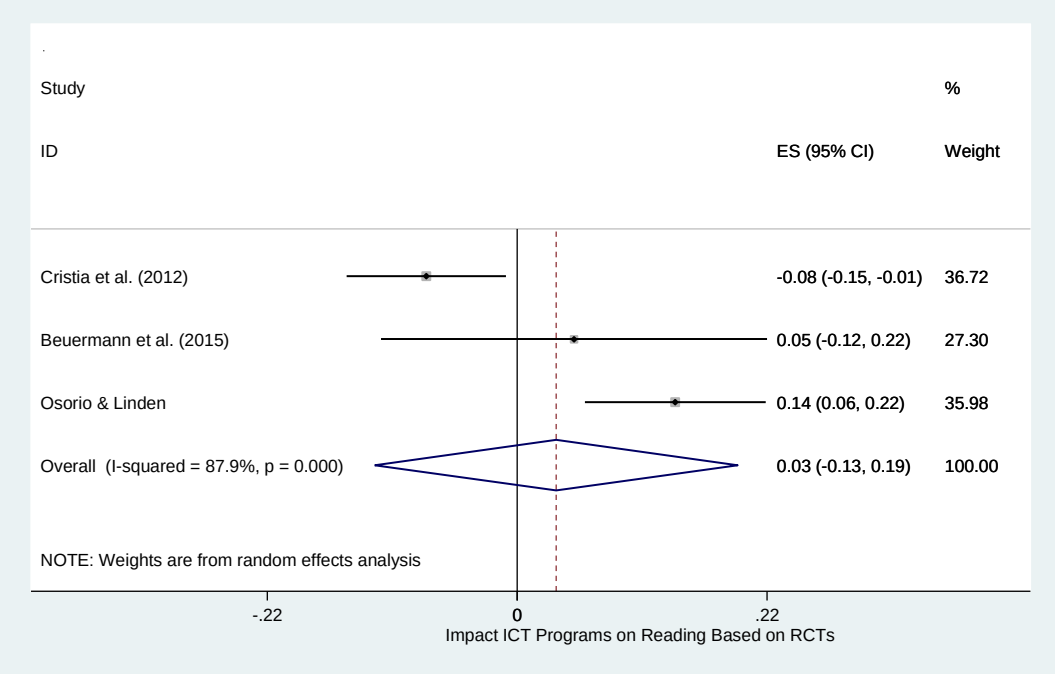
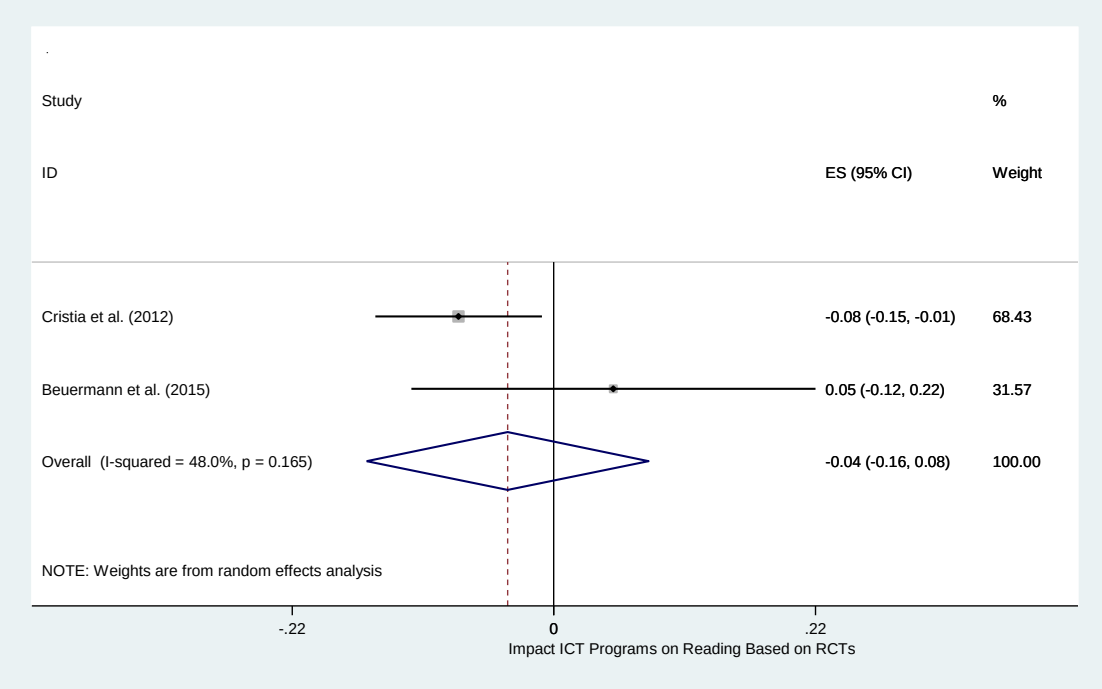


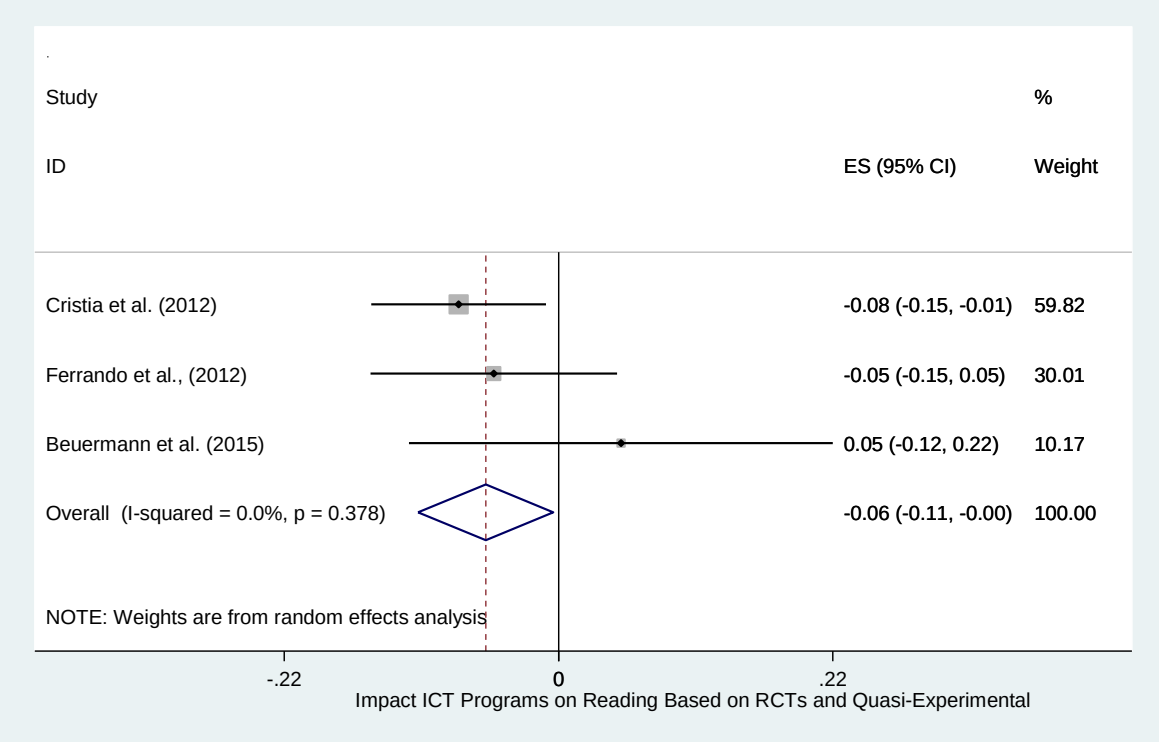
Figure 9. Impact of One Laptop per Child Program on Reading Outcomes on the Basis of RCTs



Quasi-Experimental Studies.

We found one quasi-experimental study that focused on the one laptop per child program in Uruguay. This study did not find evidence for statistically significant and positive or negative effects of this program on reading outcomes, but the point estimate is negative again. Furthermore, we found evidence for negative and statistically significant effects of the one laptop per child program on reading outcomes when we pooled the findings of this study in Uruguay with the findings of the RCTs in Peru in one meta-analysis (SMD = -0.06, 95% confidence interval (CI) = -0.11, 0.00; evidence from three studies). We report these results in Figure 10. It is important to be cautious when interpreting these results because of the medium risk of selection bias of the study in Uruguay. Nonetheless, the results are indicative of evidence that the one laptop per child program may have negative effects on reading outcomes in the LAC region.

Figure 10. Impact of One Laptop per Child Program on Reading Outcomes on the Basis of RCTs and Quasi-Experimental Studies



Together, the findings regarding the impact of ICT programs on reading outcomes in the LAC region suggest that ICT programs do not consistently have positive effects on early grade reading outcomes and may indeed have negative effects in some cases.

Nutrition Programs

Of the 24 included studies, five estimated the impact of a nutrition program on reading outcomes. We were able to include all of these studies in the meta-analysis. These studies are summarized in Table 8.

Table 8. Primary Studies That Focus on the Impact of Nutrition Programs

Study	Definition of variable	Evaluation design	Included in meta-analysis?	Country
Maluccio et al. (2009)	Reading comprehension	Cluster RCT	Yes	Guatemala
Adrogue & Orlicki (2013)	Language test score	Difference-in-Difference Analysis	Yes	Argentina
Ismail et al. (2012)	Reading test scores English test scores	Propensity Score Matching	Yes	Guyana

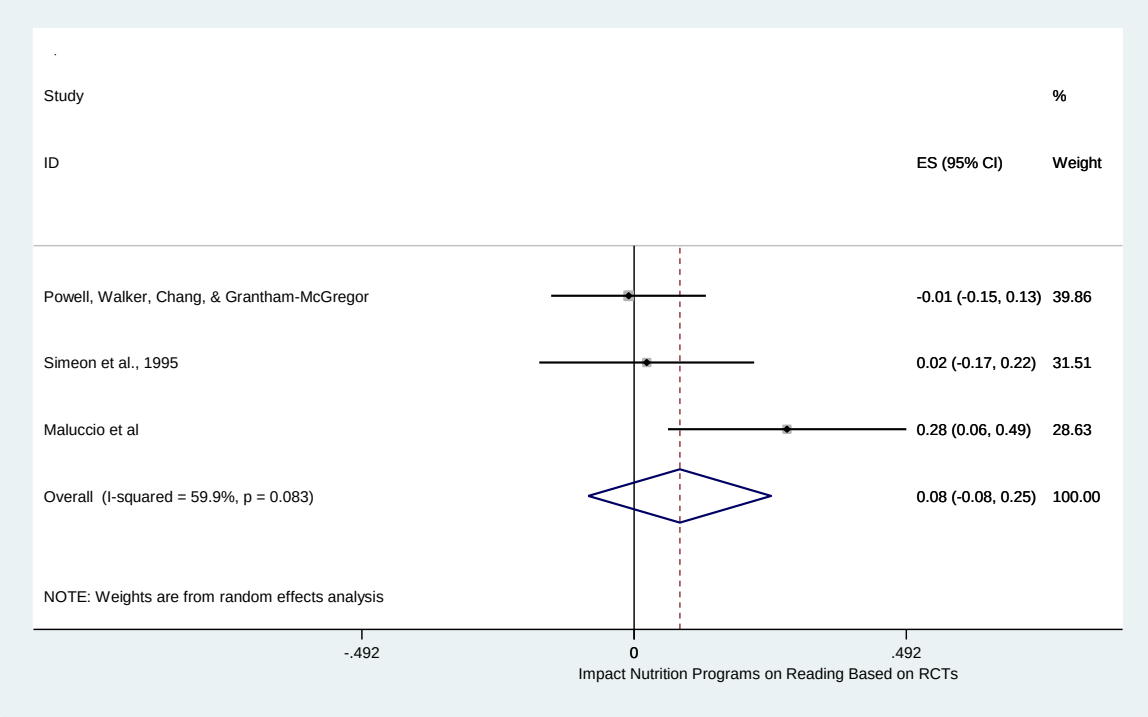
Study	Definition of variable	Evaluation design	Included in meta-analysis?	Country
Powell et al. (1998)	Reading comprehension spelling	RCT	Yes	Jamaica
Simeon et al. (1995)	Arithmetic Spelling Reading	RCT	Yes	Jamaica

Randomized Controlled Trials

We found no evidence that nutrition programs had positive and statistically significant average effects on reading outcomes in the LAC region on the basis of RCTs. Figure 11 shows the results from a meta-analysis in which we included impact evaluations of deworming and a school breakfast program in Jamaica and an impact evaluation of a program that includes the distribution of supplementary nutritious drinks in Guatemala 25 years after the start of the intervention (SMD=0.08, 95% confidence interval (CI)=-0.08, 0.25; evidence from 3 studies). The studies in Jamaica do not show evidence for positive effects of deworming and a school breakfast program on early grade reading outcomes. However, we need to be careful in the interpretation of these results because both studies have a high risk of performance bias. The studies use student-level randomized controlled trial designs. As a result the study are likely to underestimate the impact of the program because of the risk of spillovers and contamination.

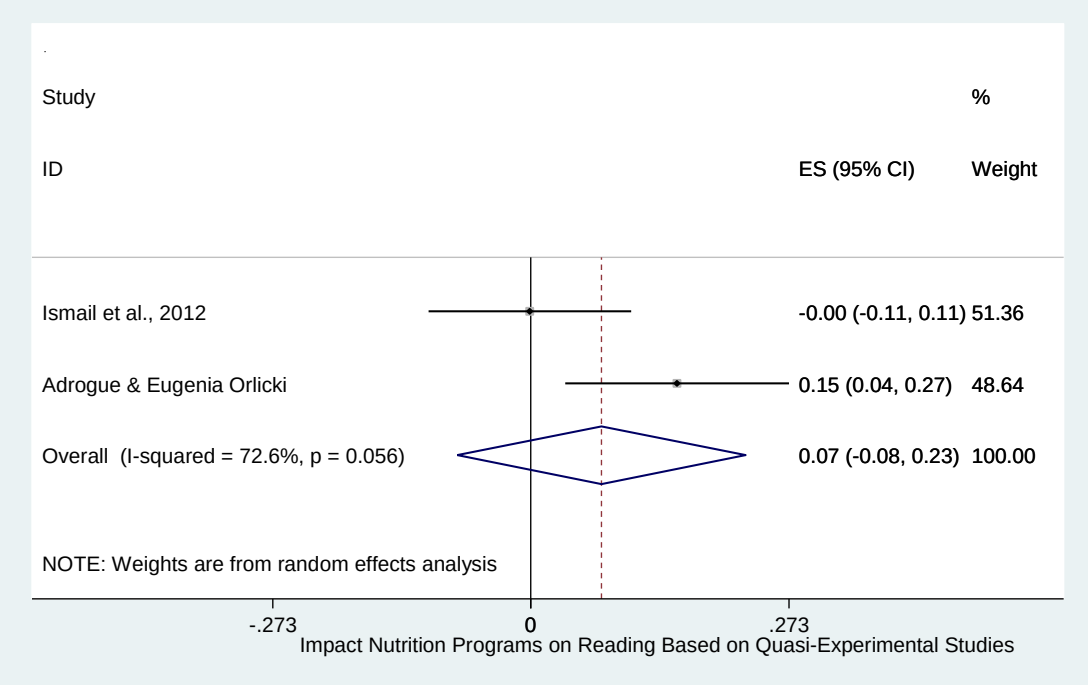
Maluccio et al. (2008) find evidence for positive effects of the distribution of nutritious supplements on reading outcomes in Guatemala. Although this study suffers from a medium risk of selection bias, the results look very promising particularly because of the long time frame of the study. However, the findings may be very context-specific. Guatemala has the highest rate of malnutrition in the LAC region (Maluccio et al., 2009). Thus, nutrition programs may be particularly effective in this context. This example shows the importance of taking into consideration enabling factors in the analysis of reading outcomes. Programs with a focus on nutrition may be very effective in improving reading outcomes in specific contexts where malnutrition rates are high.

Figure 11. Impact of Nutrition Programs on Reading Outcomes in the LAC Region Based on Randomized Controlled Trials



Quasi-Experimental Studies. We included two quasi-experimental studies of school feeding programs that estimated impacts on reading outcomes. These studies found no evidence that school feeding programs had positive and statistically significant effects on early grade reading outcomes in the LAC region (SMD = 0.07, 95% confidence interval (CI) = -0.08, 0.23; evidence from two studies). For example, Ismail et al. (2012) found no evidence of positive effects of a school feeding program on early grade reading outcomes in Guyana. Adrogue and Orlicki (2012) present some evidence that a school feeding program in Argentina had positive effects on early grade reading. However, these results are not very convincing because they are based on an evaluation design with a high risk of selection bias. Thus, we do not interpret this finding as rigorous evidence of the positive effects of school feeding programs on early grade reading outcomes in the LAC region. Nevertheless, we present the results of the meta-analysis in Figure 12.

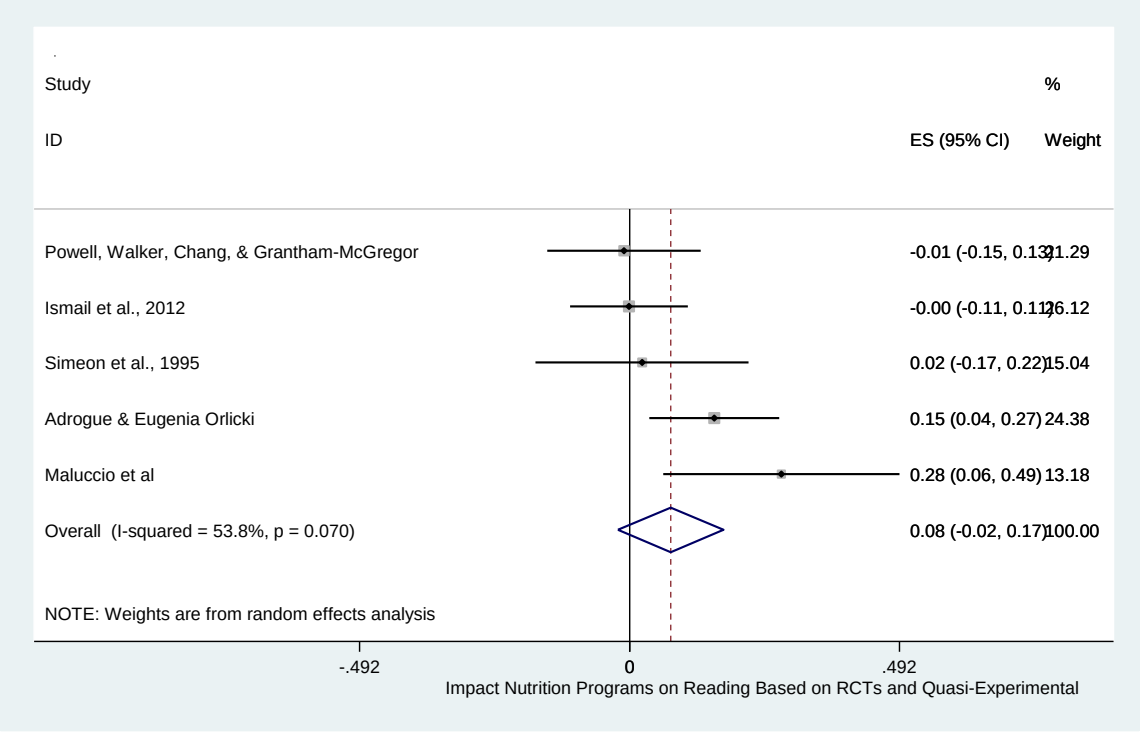
Figure 12. Impact of Nutrition Programs on Reading Outcomes in the LAC Region Based on Quasi-Experimental Studies



Despite the lack of evidence showing positive effects on early grade reading outcomes in the meta-analysis, we did find evidence for some positive effects of school feeding programs on English language outcomes. Ismail et al. (2012) reported positive and statistically significant effects of more than one standard deviation in their evaluation of a school feeding program in Guyana. This evidence indicates that school feeding programs may be effective in improving second language outcomes even if they are not able to improve early grade reading outcomes. We need to exercise caution in the interpretation of these findings, however, because the results are only based on one study with a medium risk of selection bias.

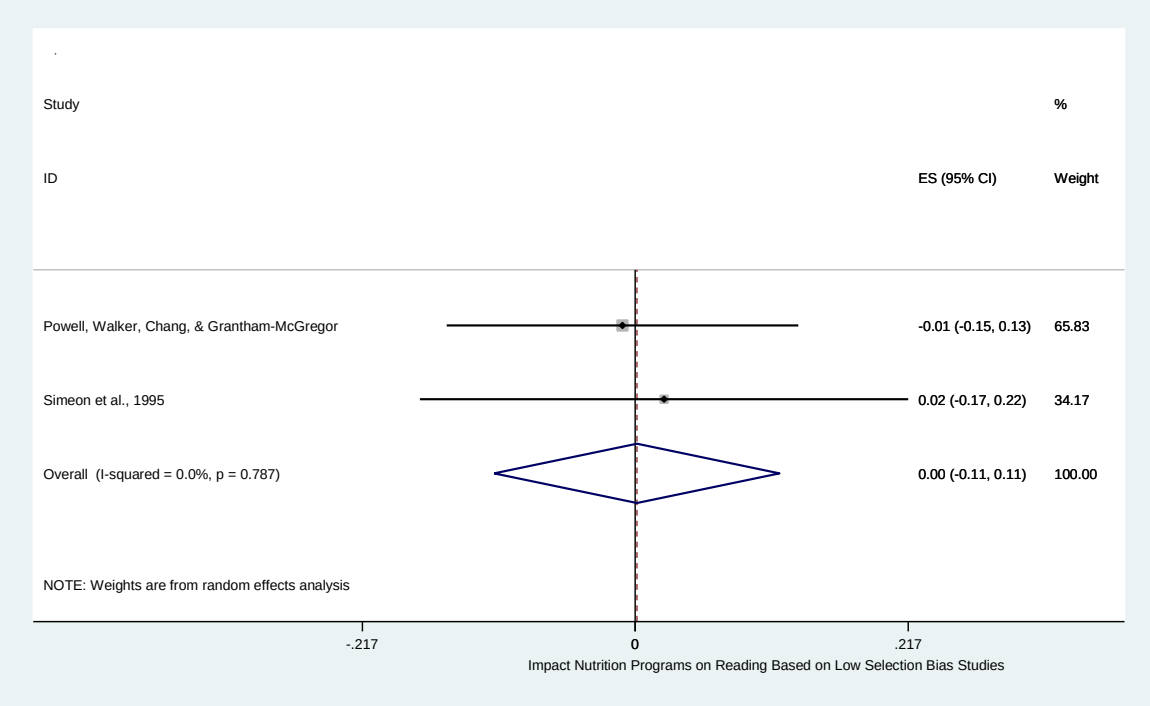
Pooled Results. We also present pooled results of the RCTs and quasi-experimental studies because the estimated effect sizes are similar. We again found no evidence of positive and statistically significant average effects of nutrition programs on early grade reading outcomes, but the results are close to statistically significant when we pool RCTs and quasi-experimental studies (SMD = 0.08, 95% CI = -0.02, 0.17; evidence from five studies). However, the positive results are driven by the study of Maluccio et al. (2008) in Guatemala and the study with a high risk of selection bias in Argentina. These findings indicate that nutrition programs may be effective in improving early grade reading outcomes, but only in contexts with high rates of malnutrition, such as Guatemala. We present the results of the pooled meta-analysis in Figure 13.

Figure 13. Impact of Nutrition Programs on Reading Outcomes in the LAC Region Based on Randomized Controlled Trials and Quasi-Experimental Studies



In any case, we should be cautious when interpreting our results because the effects of several included studies with an emphasis on nutrition on reading outcomes may present underestimates of the impact of these programs because of performance bias. For example, two of the studies in Jamaica are likely to underestimate the impact of nutrition programs on reading outcomes for this reason. We present a separate meta-analysis for these studies in Figure 14. The results show a difference between beneficiaries and no beneficiaries that is close to 0. This finding could well be explained by bias from spillovers or contamination. In that case, nutrition programs may be a promising approach to improve early grade reading outcomes but mostly in regions with high rates of malnutrition.

Figure 14. Impact of Nutrition Programs on Reading Outcomes in the LAC Region Based on Randomized Controlled Trials With a High Risk of Performance Bias



School Governance

Of the 24 included studies, two estimated the impact of a school governance program on reading outcomes. We used a narrative synthesis as opposed to a meta-analysis for school governance programs because of the small number of rigorous studies that focus on this topic. The evaluations that focus on school governance programs are summarized in Table 9.

Table 9. Primary Studies That Focus on the Impact of School Governance Programs

Study	Definition of variable	Evaluation design	Country
Bando (2010)	Language test score	OLS regression analysis	Mexico
Lockheed et al. (2010)	Early literacy outcome	Propensity score matching	Jamaica

Quasi-Experimental Studies. We included two quasi-experimental studies that focused on school governance and its impact on early grade reading outcomes. The first study focused on the impact of a cash transfer that is complemented by a matching grant as well as more responsibility for parents in decision making in primary schools in Mexico. Specifically, parents are given information and decision-making power to spend the matching grant. This process can increase school accountability, which can in turn result in improvements in the quality of education and learning outcomes. The second evaluation focused on the impact of a school improvement plan that was accompanied by increases in school inputs for primary schools in Jamaica. These school inputs included teacher training elements, parent education, and school feeding programs, reading materials, and summer courses in math and reading. Essentially, the program resulted in changes

in the implementation fidelity of other interventions. However, in contrast to the previously discussed evaluation studies, these activities are the results of changes in school governance as opposed to individual programs. Thus, we consider this study an evaluation of a school governance program and not part of any of the other program categories.

The two quasi-experimental studies found mixed evidence that school governance programs had positive effects on early grade reading outcomes in the LAC region. The matching grant program had positive effects on early grade reading outcomes in Mexico. However, the study in Jamaica did not find evidence that the school improvement program had positive effects on reading outcomes in Grade 4 (we included this study because students who were in Grade 4 during the endline survey were in early grades during the start of the program). The lack of positive impacts could be explained by the small differences in the school inputs between treatment and comparison schools even after the positive effects on school inputs.

However, we should exercise caution when interpreting these results. Both studies suffer from a medium risk of selection bias and are not able to convincingly demonstrate that their identification strategies enable the estimation of causal effects of school governance programs. Hence, the included evaluations of school governance programs do not present convincing evidence on the impact of these programs on early grade reading outcomes.

Preschool

Of the 24 included studies, two estimated the impact of participation in preschool on reading outcomes. We focused on a narrative synthesis as opposed to a meta-analysis for participation in preschool because of the small number of rigorous studies that focus on this topic. These evaluations are summarized in Table 10.

Table 10. Primary Studies That Focus on the Impact of Preschool

Study	Definition of variable	Evaluation design	Country
Campos, Esposito, et al. (2011)	Language test score	Hierarchical regression analysis	Brazil
Felicio, Terra, & Zoghbi, (2011)	Literacy score	Propensity score matching	Brazil

Quasi-Experimental Studies.

We included two quasi-experimental studies that focused on preschool and its impact on early grade reading outcomes in Brazil. Campos et al. (2011) argue that participation in preschool led to an improvement in language assessment scores for children in six Brazilian state capitals. They used hierarchical multivariate regression analysis to demonstrate the positive effects. Similarly, De Felicio et al. (2012) found that participating in early childhood education had positive effects on the literacy scores of children in second grade. They used propensity score matching to identify these impacts.

Although Campos et al. (2011) and De Felicio et al. (2012) make valid attempts to identify the impact of participation in preschool on early grade reading outcomes in Brazil, the two studies both suffer from risk of selection bias. We rated the study of De Felicio et al. (2012) as having a

medium risk of selection bias and the study of Campos et al. (2011) as having a high risk of selection bias. Thus, caution should be exercised when interpreting our results. Previous evidence suggests that participation in preschool can have a wide range of positive effects on children in low- and middle-income countries (Martinez et al., 2012). However, the studies of De Felicio et al. (2012) and Campos et al. (2011) are likely to suffer from bias due to selection on unobservables. Hence, these studies do not present convincing evidence that participation in preschool leads to improvements in early grade reading outcomes. It is possible that participation in preschool has these effects in the LAC region, but more rigorous research is needed to demonstrate these effects. For example, preschool may only be effective when the education is of sufficient quality.

Teacher Practices

Of the 25 included studies, six estimated the impact of the adoption of distinct teacher practices, such as the explicit instruction of new words, shared story book reading, and read-alouds. We used a narrative synthesis as opposed to a meta-analysis for teacher practices because the teacher practices that are discussed are very dissimilar. Therefore, we do not expect that a pooled effect size of these teacher practices would present any meaningful information. The evaluations that focus on teacher practices are summarized in Table 11.

Table 11. Primary Studies That Focus on the Impact of Teacher Practices

Study	Definition of variable	Evaluation design	Country
Larrain, Strasser, & Lissi, (2012) Experiment 1	Vocabulary acquisition	RCT	Chile
Larrain, Strasser, & Lissi, (2012) Experiment 2	Vocabulary acquisition	RCT	Chile
Cardoso-Martin et al. (2011) Experiment 1	Letter naming Decoding	RCT	Brazil
Cardoso-Martin et al. (2011) Experiment 2	Letter naming Decoding	RCT	Brazil
Vivas (1996) Experiment 1	Language comprehension Expressive language	RCT	Venezuela

Randomized Controlled Trials.

We included five RCTs that focused on the effects of specific teacher practices on early grade reading outcomes in the LAC region. These evaluations focused on distinct practices, such as the explicit instruction of new words, complex word elaboration during shared story book reading, and letter name teaching as opposed to only teaching the shapes of letters. The specifics of these tasks enabled researchers to examine how reading outcomes change in great detail. Researchers usually make use of this opportunity by estimating the impact of these practices on various reading constructs, such as letter recognition and vocabulary acquisition. Although the sample sizes for the included studies was small ($n<100$ in the majority of the studies), researchers nonetheless found statistically significant effects in the majority of the studies. However, these statistically significant effects are unlikely to be an indication of the effectiveness of teacher practices. Instead, evidence

suggests that the statistically significant effects are associated with publication bias. Evidence indicates that published studies with small sample sizes are disproportionally affected by publication bias (Borenstein et al., 2009).

Although the results of the studies are likely biased due to publication bias, the included studies on teacher practices present some interesting findings about how specific teacher practices can influence reading outcomes. These findings can serve as hypotheses for larger-scale research on which teacher practices are most effective in improving early grade reading outcomes. First, Larrain et al. (2012) presented evidence that word elaboration during shared story book reading has a positive effect on vocabulary acquisition. Larrain and colleagues (2012) also suggest that using simpler definitions of words is more effective in improving vocabulary acquisition than using complex definitions. In addition, Cardoso-Martins et al. (2011) found that teaching the names of letters is more effective than merely teaching the shapes of letters. They also present evidence that training children in phonological awareness can improve the learning of letter sounds (Cardoso Martins et al., (2011). Neugebauer and Currie-Rubin (2009) present some experimental evidence that reading aloud can improve reading outcomes in Peru. Finally, Vivas (1996) demonstrates that listening to teachers reading stories aloud results in improvements in language comprehension and expressive language first grade children.

The results of the studies with an emphasis on specific teacher practices should merely be interpreted as interesting hypotheses for larger-scale quantitative research for two reasons. First, as discussed above, there is major evidence for publication bias, which may invalidate the results of the studies because they are likely not replicable. Second, each of the included quantitative intervention studies with a focus on specific teacher practices suffers from a medium or high risk of selection bias. Each of these studies had a sample size that was too small to ensure equivalence in observable and unobservable characteristics between the treatment and the control groups. In addition, several of these studies made methodologically inappropriate choices in the design or analysis of the results. For example, Cardoso-Martins et al. (2011) switched treatment students to the control group because the students self-selected in the control group. These kinds of choices can result in a considerable risk of selection bias, particularly with small sample sizes. Thus, we do not recommend that policy makers or practitioners base their decisions on the findings of small-scale quantitative intervention studies with a focus on specific teacher practices. However, it would be interesting to test the effectiveness of specific teacher practices on a larger scale.

Parental Involvement

Of the 24 included studies, three estimated the impact of parental involvement with the aim of improving early grade reading outcomes. We used a narrative synthesis as opposed to a meta-analysis for parental involvement because of the small number of rigorous studies that focus on this topic. The evaluations that focus on parental involvement are summarized in Table 12.

Table 12. Primary Studies That Focus on the Impact of Parental Involvement

Study	Definition of variable	Evaluation design	Country
Tapia & Benitez, (2013)	Reading practices	RCT	Mexico
Vivas (1996) Experiment 2	Language comprehension Expressive language	RCT	Venezuela
Murad & Topping, 2000	Reading Practices Reading Comprehension Reading Fluency	RCT	Brazil

Randomized Controlled Trials. We included three studies that focused on the effects of programs that involve parents on early grade reading outcomes. Both of these studies were RCTs with a small sample size and challenges in the implementation of the randomization. Tapia and Benitez (2013) found that teaching mothers about joint reading of stories and puppet play had the potential to improve their literacy practices with their children. In addition, Vivas (1996) presents evidence that listening to stories read aloud by parents results in improvements in language comprehension and expressive language in first grade children. Finally, Murad & Topping (2000) found positive effects of paired reading with parents on children's reading comprehension and fluency.

However, similar to the studies with an emphasis on specific teacher practices, it is likely that the studies with a focus on parental involvement suffer from publication bias. The studies show positive and statistically significant results despite being underpowered to demonstrate these effects. Thus, although the studies by Tapia and Benitez (2013), Vivas (1996) and Murad & Topping (2000) show interesting hypotheses that need to be tested in larger scale studies, we do not recommend that policy makers use these studies to inform their decisions.

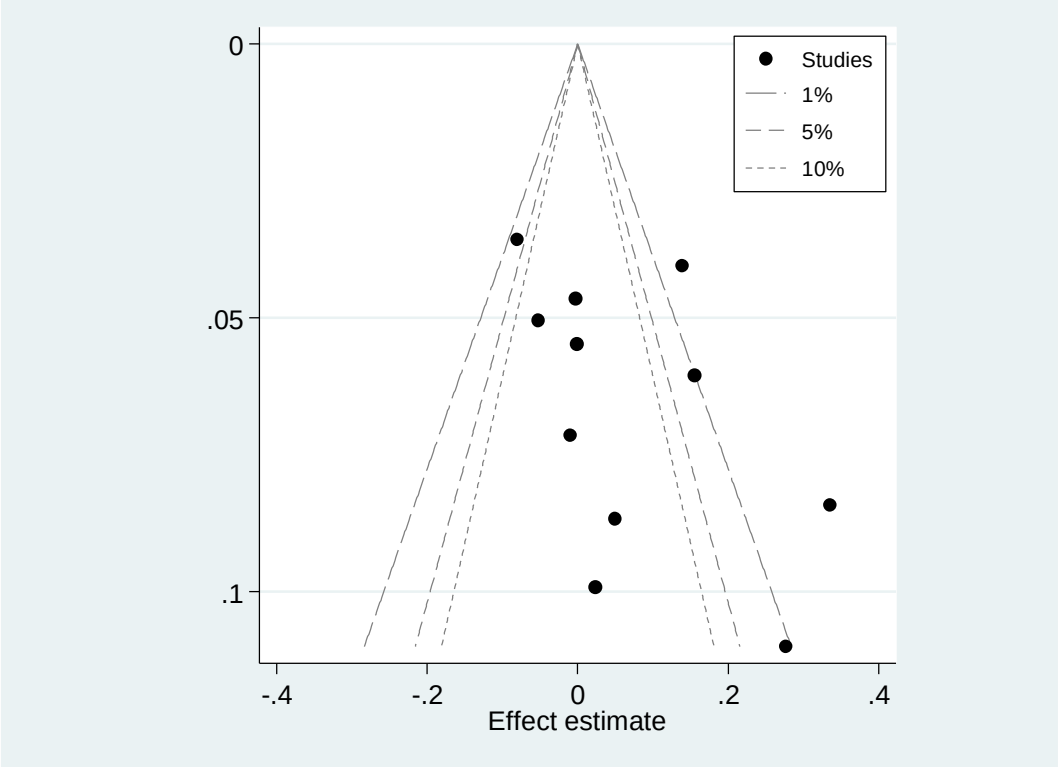
Publication Bias

It is likely that the included studies with a small sample size present a biased overview of the impact of specific teacher practices and parental involvement on early grade reading outcomes because of publication bias. As discussed in the previous section, publication bias is likely if studies with a small sample size consistently show positive and statistically significant effects on early grade reading outcomes (Borenstein et al., 2009). The unusual high statistical significance in the studies with a smaller sample size is a strong indication for publication bias in studies with a focus on teacher practices and parental involvement.

We also find some indications for publication bias in the studies we were able to include in our meta-analysis. We relied on funnel plots and the Egger test to examine the possibility of publication bias in these studies. The idea underlying funnel plots is that publication bias is most likely when effect sizes do not follow a normal distribution. Although we cannot reject the null hypothesis of no normal distribution, the funnel plot indicates that the effect sizes in the studies we included in the meta-analysis do not follow a full normal distribution. We present the funnel plot in Figure 15 below. We formally tested for publication bias by applying the Egger test. This test did not indicate formal evidence for publication bias in the studies that we included in the meta-analysis (β=2.91, S.E.=1.74; p=0.13). However, although the publication bias is not statistically significant, there is nonetheless some indication for publication bias in the studies we

included in our meta-analysis. We have to remain careful in interpreting this result, however, because tests for publication bias are only indicative of publication bias. There may be other explanations for the non-normal distribution. Nonetheless, our results suggest that publication bias may be present in our larger-scale studies as well.

Figure 15. Funnel Plot to Test for Publication Bias



Evidence-Gap Map of Quantitative Intervention Studies

An analysis of the quantitative intervention studies indicates that impact evaluations with an emphasis on early grade reading outcomes only focus on a small portion of the intervention types that can influence early grade reading outcomes. We only found three topic areas with more than two impact evaluations that focus on early grade reading outcomes: (1) teacher training, (2) nutrition interventions, and (3) ICT programs. This finding indicates that there is only a small amount of evidence for what works (or does not work) to improve early grade reading outcomes in the LAC region. The evaluations with an emphasis on teacher training, nutrition, and ICT also only focus on a small number of countries. The evidence base for these interventions exclusively focuses on Chile (3 evaluations), Peru (2 evaluations), Jamaica (2 evaluations), and Argentina, Guyana, Uruguay, and Colombia.

The included impact evaluations predominantly focus on high-income and upper-middle income countries in the LAC region. Of the included evaluations, seven focus on high-income economies such as Chile and Argentina, 11 focus on upper-middle income economies, two focus on lower middle-income economies, and none of the impact evaluations focus on low-income economies. It is likely that the limited research capacity and funding in low-income and lower-middle-income

economies constrains the ability of researchers in those countries to conduct rigorous impact evaluations. This hypothesis is exemplified by the fact that the two included impact evaluations in lower-middle-income economies were conducted by researchers based in the United States, while the majority of the impact evaluations in high-income economies were conducted by local researchers.

Although the majority of the included impact evaluations used an RCT, only eight of the studies were rated as having a low risk of selection bias. Of the eight studies with a low risk of selection bias, two focus on child nutrition, three focus on ICT, one focuses on parental and community participation, one focuses on teacher practices for reading, and two focus on teacher training. These data show that there is little strong evidence regarding the impact of development programs on early grade reading outcomes.

Several studies with a low risk of selection bias also suffer from a medium or high risk of performance bias. Of the eight studies with a low risk of selection bias, only three studies qualify as studies with a low risk of performance bias. These studies are the only included studies that can make credible causal claims about the impact on early grade reading outcomes without significant risks of spillovers and contamination. These three studies focus on a teacher training program in Chile, the distribution of one laptop per child in Peru, and the provision of computers for computer-aided instruction in Colombia. This finding indicates that there is a gap in the evidence on what works to improve reading outcomes in the LAC region, because this question can only be credibly addressed in studies that qualify as studies with low risk of selection bias and low risk of performance bias.

The general gaps in the evidence are presented in Tables 13–16. Table 13 shows the included evaluations by country. Table 14 shows the included evaluations by study design. Table 15 shows the included evaluations by country type. Table 16 shows the evaluations by intervention type. Each of these figures also show the outcome measures that were included in the evaluations. We grouped outcome measures that were associated with reading outcomes without distinguishing between different types of reading constructs. Thus, we pooled studies that measure reading comprehension with studies that use a standardize language assessment and studies that measure decoding. It is important to note, however, that the majority of the studies only included one measure of early grade reading outcomes without distinguishing between decoding, phonemics, reading comprehension, and so forth. This lack of emphasis on reading subskills is by itself a gap in the literature on what works to improve reading outcomes.

Table 13. Evaluations by Country

Gap Map of Country of Intervention and Outcome Measures*			
	Literacy skills	**Student reading practices**	**Parental reading practices**
Argentina	Adrogue & Orlicki (2013)		
Brazil	Felicio, Terra, & Zoghbi (2011); Cardoso-Martin et al. (2011); Campos et al. (2011); Murad & Topping (2000		
Chile	Larrain, Strasser, & Lissi (2012); Gomez Franco (2014); Pallante & Kim (2013); Yoshikawa et al. (2015); Mendive et al. (2016)		
Colombia	Osorio & Linden (2009)		
Guatemala	Maluccio et al. (2009)		
Guyana	Ismail et al. (2012)		
Jamaica	Lockheed et al. (2010); Powell et al. (1998); Simeon et al. (1995)		
Mexico	Bando (2010)		Tapia & Benitez (2013)
Peru	Cristia et al. (2012)	Beuermann et al. (2013)	
Uruguay	Ferrando et al. (2011)		
Venezuela	Vivas (1996)		

Table 14. Evaluations by Study Design

	Literacy skills	Student reading practices	Parental reading practices
RCT	Cardoso-Martin et al. (2011); Larrain, Strasser, & Lissi (2012); Gomez Franco (2014); Pallante & Kim (2013); Yoshikawa et al. (2015); Mendive et al. (2016); Osorio & Linden (2009); Cristia et al. (2012); Powell et al. (1998); Simeon et al. (1995); Maluccio et al. (2009); Murad & Topping (2000)	Beuermann et al. (2013)	Tapia & Benitez (2013)
Quasi-Experimental	Adrogue & Orlicki (2013); Felicio, Terra, & Zoghbi, (2011); Bando (2010); Ferrando et al. (2011); Lockheed et al. (2010);Ismail et al. (2012)		
Nonexperimental	Campos et al. (2011)		

Table 15. Evaluations by Country Type

	Literacy skills	Student reading practices	Parental reading practices
High income	Larrain, Strasser, & Lissi (2012); Gomez Franco (2014); Pallante & Kim (2013); Yoshikawa et al. (2015); Mendive et al. (2016); Gomez Franco (2014);Ferrando et al. (2011)		
Upper middle income	Adrogue & Orlicki (2013); Felicio, Terra, & Zoghbi (2011); Cardoso-Martin et al. (2011); Campos et al. (2011); Osorio & Linden (2009); Bando (2010); Vivas (1996); Cristia et al. (2012); Powell et al. (1998); Simeon et al. (1995); Murad & Topping (2000); Lockheed et al. (2010)	Beuermann et al. (2013)	Tapia & Benitez (2013)
Lower middle income	Maluccio et al. (2009); Ismail et al. (2012)		

Table 16. Evaluations by Intervention Type:

	Literacy skills	Student reading practices	Parental reading practices
Child nutrition	Powell et al. (1998); Simeon et al. (1995); Maluccio et al. (2009); Ismail et al. (2012); Adrogue & Orlicki (2013)		
ICT	Osorio & Linden (2009); Cristia et al. (2012); Ferrando et al. (2011)	Beuermann et al. (2013)	
Parental and community participation	Vivas (1996); Murad & Topping (2000)		Tapia & Benitez (2013)
Preschool	Felicio, Terra, & Zoghbi (2011); Campos et al. (2011)		
Teaching practices for reading (Reading out loud, etc. Must be reading specific)	Larrain, Strasser, & Lissi (2012); Cardoso-Martin et al. (2011); Vivas (1996)		
Teacher training	Gomez Franco (2014); Pallante & Kim (2013); Yoshikawa et al. (2015); Mendive et al. (2016); Gomez Franco (2014)		
School governance	Bando (2010); Lockheed et al. (2010)		

We also present evidence-gap maps that only focus on study designs that enable credible causal claims about the impact of development programs on early grade reading outcomes. These gap maps only include the studies that can be considered as having a low risk of selection bias and low risk of performance bias. We present these gap maps in Tables 17–20. The tables show that all studies that are able to demonstrate causal claims about the impact of development programs on early grade reading outcomes are cluster RCTs that focus on high-income or upper-middle-income economies.

Table 17. Evaluations by Country

	Literacy skills	Student reading practices	Parental reading practices
Chile	Yoshikawa et al. (2015)		
Colombia	Osorio & Linden (2009)		
Peru	Cristia et al. (2012)		

Table 18. Evaluations by Study Design

Gap Map of Study Design and Outcome Measures*			
	Literacy skills	**Student reading practices**	**Parental reading practices**
RCT	Yoshikawa et al. (2015); Osorio & Linden (2009); Cristia et al. (2012)		

Table 19. Evaluations by Country Type

	Literacy Skills	**Student reading practices**	**Parental reading practices**
High income	Yoshikawa et al. (2015)		
Upper middle income	Osorio & Linden (2009); Cristia et al. (2012)		

Table 20. Evaluations by Intervention Type

	Literacy skills	**Student reading practices**	**Parental reading practices**
ICT	Osorio & Linden (2009); Cristia et al. (2012)		
Teacher training	Yoshikawa et al. (2015)		

We also found gaps in the evidence base on what does not work to improve early grade reading outcomes in the LAC region. Specifically, our studies indicate that there is a large risk of publication bias in the evidence base on what works to improve early grade reading outcomes in the LAC region. We encountered a large percentage of studies with a small sample size that nonetheless found statistically significant effects on early grade reading outcomes. Our formal tests for publication bias indicate that it is very plausible that other studies with a small sample size that did not find these statistically significant effects were either not published or never finished by the authors because the authors did not find evidence for statistically significant effects. This finding is worrisome because it suggests that published studies with a focus on the effects of development programs on early grade reading outcomes generally tend to overestimate the effectiveness of these programs. This tendency complicates the ability of donors and governments to make evidence-based policy decisions.

Quantitative Nonintervention Research

It is important to include quantitative non-intervention studies in our review because these studies allow for learning about the predictors of early grade reading outcomes and the mechanisms that explain changes in early grade reading outcomes. Systematic reviews typically do not include quantitative nonintervention studies, however, because often these studies are not able to address counterfactual questions. We considered it important to include these studies, however, because they often examine the specifics of reading acquisition mechanisms and trajectories. In addition, these studies are able to uncover predictors of reading success, as part of the larger story of evidence of EGR development in the LAC region. In particular, we believe these studies can guide curricular and standards development, entangle specific aspects—and paths through which—a "bundled" EGR program may impact reading and help develop more targeted, language- and country-specific reading measures.

The quantitative nonintervention studies comprised the largest number of studies in the systematic review. The review included 62 articles from the following countries: Brazil (N = 18), Mexico (N = 13), Chile (N = 10), Argentina (N = 6), Peru (N = 4), Guatemala (N = 3), Cuba (N = 2), Puerto Rico (N = 1), Colombia (N = 1), and Costa Rica (N = 1). We also included two studies that involved cross-country comparisons. The included studies were mostly from psychology and linguistics disciplines and covered a range of topics on predictors of reading skill development in the LAC region. Next we present a gap map (Table 21) with only the medium and high quality evidence which shows the various gaps in the quantitative nonintervention research in the LAC region by topic and country. The gap-map demonstrates that there is no quantitative non-intervention research related to curriculums, ICT, reading in bilingual or multilingual contexts, or school governance.

Table 21. Quantitative Nonintervention Gap Map

Topics	Argentina	Brazil	Chile	Colombia	Cuba	Guatemala	Mexico	Peru	Puerto Rico
Assessment		Dias et al. (2006); Athayde et al. (2014); Capovilla et al. (2004)			Reigosa Crespo et al. (2013)	Salazar et al. (1996)			
Child nutrition						Hoddinott et al. (2013)			
Curriculum									
Disabilities		Bandini et al. (2006); Cardoso-Martins & Da Silva (2010).							
General pedagogical approaches			Pino & Bravo (2005); Bravo et al. (2002)						
ICT									
Oral language development								Abadzi et al. (2005)	Páez et al. (2007)
Parental and community participation		Fuller et al. (1999)							
Preliteracy/ Emergent literacy		Kessler et al. (2013)	Guardia (2003)						

Topics	Argentina	Brazil	Chile	Colombia	Cuba	Guatemala	Mexico	Peru	Puerto Rico
Poverty			Guardia (2003); Bizama et al. (2011)	Silva et al. (2013)			Cervini (2015)		
Preschool		Oliveira (1996)	Pino et al. (2005)				Benítez et al. (2007)	Castro et al. (2002)	
Reading in bilingual/ Multilingual contexts									
Reading skills	Plana & Fumagalli (2013)	Correa & Dockrell (2007); Medeiros et al. (2011)	Guardia (2011); Bravo et al. (2002); Villalon & San Francisco (2001)				Goldenberg et al., (2014)		
Teaching practices for reading	Manrique et al. (1994)		Muñoz (2011)				Reynoso-Alcántara et al. (2010); Goldenberg et al., (2014)		
Teacher quality/ Training		Fuller et al. (1999)							
Writing		Correa & Dockrell (2007)	Villalon & San Francisco (2001)						

In the following section, we discuss the quality criteria used to determine which studies to include in the review, the main findings that emerged from the studies, and implications for stakeholders and relevance to the field at large.

Quality Criteria

All non-intervention studies were rated by reviewers on a range of questions that were pooled together to target the following categories of quality: the outcome measures, sample, data collection, data analyses, and external validity. In the following section, we first describe in general how the whole set of studies were reviewer per category; and then in the second part, we present the reviewers' ratings for each study on each category.

Outcome Measures

Our most important category was whether or not reading, writing, or some reading- or writing-related subskill was measured. Two main questions were used to determine whether a study was included or not:

1. Did the outcome measure include some measure of reading or a reading subskill (e.g., fluency, phonological awareness, language decoding, letter knowledge, comprehension, etc.)?
2. If the study did not include a measurement of reading or a reading subskill, was literacy measured in a different manner?

In the sample, 58 of 64 studies had an outcome measure of reading or a reading subskill. In general, phonological awareness and reading were measured. Reading measures ranged from word level reading to reading connected text. One example of a study that focused on the essential components of reading and included writing was Plana and Fumagalli (2013). In contrast, some studies focused only on decoding (Jaichenco & Wilson, 2013). One study in the sample measured reading in a different manner than through phonological awareness or reading comprehension. Silva, Strasser, and Cain (2014) measured students' narrative skills using a wordless picture book that students used to construct a story.

The majority of the studies used reading assessment tests to measure reading outcomes, which reduces the risk of measurement error. Only 6 of 64 studies in the sample reported information on self-reports. These involved student (Cervini, 2015), parent (Salazar-Reyes & Vega-Perez, 2013), or teacher surveys (Janus, 2011).

We were also interested in understanding whether the studies provided information on data collector training in order to determine, to the extent reported, whether there were any concerns regarding the independence of the observers. We found that only 13 of 64 studies provided information on training of test administrators. Test administrators consisted of the study author (De Abreu & Cardoso-Martins, 1998) and graduate students (Benitez, Vargas, Hernandez, Sanchez, & Garcia, 2007). In one study, research assistants were trained over a week-long period on how to record their classroom observations. They then practiced by observing videotaped and live classrooms in Northeast Brazil. Following this training, pairs of observers were sent to 17 different classrooms in a school to obtain interrater reliability (Fuller et al., 1999). The studies demonstrate a wide range of variability when it comes to data collector training procedures and the degree to which such procedures are reported.

Sample

We were also interested in ascertaining whether the sample selection criteria were presented and justified in each study. We assessed whether the sample selection criteria were provided to determine whether the sample was appropriate for addressing the research question and to assess the generalizability of the results. We found that 45 of 64 studies provided sample selection criteria or justification of the sample selection process. Samples were generally described by age, grade, gender, economic level, country, and geographical region. In some cases, samples were also described as attending private or public schools (Jiménez, Puente, Alvarado, & Arrebillaga, 2009). Some studies excluded students with visual or hearing impairments (Salles & Parente, 2002), while others included students with hearing impairments (Bandini, Oliveira, & Souza, 2006). One study included students from 16 Latin American countries for a total sample of 90,251 students (Torrecilla & Carrasco, 2014). This study examined the effect of child labor on third and sixth grade students' academic achievement in math and reading. Another study compared students from Latin America to students in the United States (Treiman, Kessler, & Pollo, 2006).

Data Collection

We were also interested in determining the quality of various aspects of data collection, including training test administrators, data collection procedures, and whether or not the study took into consideration potential data collection implementation failures. Given that we had to rely upon study authors to report this information, we were cautious in interpreting these results. In other words, simply because it was not reported does not mean it was not done.

In the sample, 31 studies reported on data collection procedures. These ranged from individual to group administration of tests in the classroom or another room in the school. Locations were generally described as quiet. One study reported that children were individually tested in a single session in a quiet room in the school (Treiman, Kessler, & Pollo, 2006). Another study reported that the students were tested using a web-based assessment (Rosas et al., 2015). Nearly half of the studies in the sample did not report the data collection procedures.

Only 10 studies in the sample reported considering data collection failures, for a number of reasons, including priming effects and blinding (Silva et al., 2014) and inability to locate all of the participants (Castro et al., 2002). Another reason given for potential data collection errors was the cultural and linguistic differences between the test administrator and the students (Kudo & Bazan, 2009) and lack of cultural appropriateness (Castro et al., 2002). Castro et al. (2002) used a test that was translated and previously used in a United States study. The researchers concluded that it might have lacked cultural appropriateness.

We were also interested in determining whether authors acknowledged that monitoring activities may influence participant behavior. Only nine studies in the sample mentioned that monitoring can influence behavior. Monitoring behavior was not a factor across the studies. The focus of the studies was test performance. Students were assessed either orally or in a written test. In general, no information was provided regarding the behavior of the child while reading. The focus was on the accuracy of test responses, not on the effects of being administered an oral assessment or the effect of students' behavior due to testing.

Analysis

The analysis section for each study was important in determining the quality of the entire study. We asked the following questions to determine the quality of the analysis section:

1. Is there a description of the analytic method(s) used?

   The majority of the studies in the sample, 55 of 64, gave a strong description of the analysis methods used. Some studies provided ample description of the statistical analyses conducted (Paez, Tabors, & Lopez, 2007) while others gave brief descriptions and used simple analyses such as histograms (Bandini, et al., 2006). Two studies did not provide a description of the analysis (Massone & Baez, 2009; Melchiori, de Souza, & de Rose, 2012).

2. Does the study justify the analysis method (is the analysis method appropriate for the research question/objective)?

   In the sample of studies, 44 of 64 studies used analysis methods that were appropriate for the research question or study objective. In some cases, the analysis method was considered to be too simplistic and did not necessarily yield empirical information. For example, Dias et al. (2006) used T-tests for analyses and Morales, Van de Vijver, & Potinga, (2013) used differential item functioning.

3. Were any participants not included in the analysis? If so, is there justification for why?

   In 43 of 64 studies in the sample, all students were tested. Of the studies that excluded students from the sample, reasons provided were that they were absent (Cardoso-Martins, & Da Silva, 1999), researcher error (De Abreu & Cardoso-Martins, 1998), or because of age (Rindermann, Baumeister, & Groper, 2014). Ten studies in the sample did not specify this information in their report. The absence of these students may have resulted in a bias in the empirical findings.

4. Was there data reported on covariates?

   Information on covariates was reported in 35 of 64 studies in the sample. Covariates centered on similar characteristics mentioned above for sample descriptions (e.g., age, grade, gender, economic level, country, and geographical region). However, some studies included covariates such as parent's educational levels (Hoddinott, et al., 2013; Muñoz, 2002) and sociocultural characteristics influencing students (Iparraguirre, 2014).

5. Are there appropriate reliability scores for all tests?

   In the sample, 18 of 64 studies reported reliability scores for the tests. Among the studies reporting test scores, Cronbach's Alpha was commonly used to calculate reliability scores (Jimenez et al., 2009; Paez et al., 2007). Those studies with tests with reasonable reliability scores were deemed high quality.

External Validity

In the area of external validity, our goal was to determine whether authors generalized their findings only to the relevant population of study. In the sample, 47 of the 64 studies generalized the study outcomes to the population in the study. Several studies generalized the study findings to a different grade level or age group (Ramírez, Verdugo, & Sánchez, 2000), another country

(Manrique & Signorini, 1994), or to the population in the study despite a small sample (Bandini et al., 2006). Still, others generalized to the entire population in the country (De Abreu, & Cardoso-Martins, 1998) and across countries (Abadzi, et al., 2005). As such, most studies generalized their findings to a relevant population.

In the second part of the analysis a quality rating of "High", "Medium", and "Low" was assigned for each study on each category. The reviewers assigned the quality as they were answering the questions described above. If the answer to the question was "Yes" and the reviewer could identify portions of the full-text study that could justify their answer, the study was rated as "High", and vice versa for "Low". Reviewers rated studies as "Medium" on categories that were present, but were not strongly backed up in the study.

Two important points emerged in this part of the analyses. First, the notion of an appropriate "theoretical framework" may have been conceptualized slightly different amongst the reviewers from different disciplinary backgrounds, and therefore, studies with Medium or Low quality theoretical frameworks were re-checked by a second reviewer. Second, in terms of quality of data collection procedures, the procedures under which data collection took place (i.e. whether it is was in a quiet room, whether testing was counterbalanced, whether fatigue effects were taken into consideration etc.) were of more importance to these kinds of nonintervention studies, as opposed to observer bias, because there is a lower likelihood of bias due to the fact there are no programs to have any vested interest in.

Table 22 shows our quality ratings for each study on the following six criteria: 1) theoretical framework, 2) outcome measure, 3) sample selection, 4) data collection procedures, 5) analysis in light of the research question, and 6) external validity.

Table 22. Quality Ratings for Quantitative Nonintervention Studies

Study	Theoretical framework explaining the study's motivation and findings	Quality of outcome measure	Sample selection quality	Quality of data collection procedures	Quality and relevance of analysis, given the research question	External validity
Guardia, 2003	High	High	Low	High	High	High
Bizama et al., 2011	High	High	Low	Low	High	High
Muñoz, 2011	High	High	High	Low	High	Medium
Bandini et al., 2006	Low	Medium	High	Low	Medium	Low
Barrera & Maluf, 2003	Low	High	High	Low	High	High
Cardoso-Martins & Da Silva, 2010	Low	High	High	High	High	High
Cardoso-Martins, 2008	High	High	Medium	Medium	Medium	Medium
Cervini, 2015	High	Medium	High	Low	High	High
Giacomoni et al. 2015	Low	Low	Low	Low	High	Low
Matute et al., 2012	Low	Low	Medium	Medium	High	High
Torrecilla & Carrasco, 2014	High	Low	High	Low	High	High
De Abreu & Cardoso-Martins, 1998	Low	High	High	Medium	High	Medium
Massone & Baez, 2009	Low	Medium	Low	Low	Low	Low
Dias et al., 2006	Low	Medium	Low	Low	Low	Low
Páez, Tabors & López, 2007	High	High	High	High	Medium	Medium
Jaichenco & Wilson, 2013)	High	High	High	High	High	Medium
Iparraguirre, 2014	High	Low	Low	Low	Low	Medium
Medeiros et al., 2011	Medium	Medium	Low	Medium	Medium	Medium
Jiménez et al., 2009E	High	Medium	High	High	High	Medium
Gómez-Pérez et. al, 2011	High	High	High	High	High	High

Study	Theoretical framework explaining the study's motivation and findings	Quality of outcome measure	Sample selection quality	Quality of data collection procedures	Quality and relevance of analysis, given the research question	External validity
Athayde et al., 2014	Low	Medium	Low	Low	High	Medium
Francis, 1999	Low	Medium	Low	Low	Medium	High
Salles & Parente, 2002	High	High	High	High	Medium	High
Goldenberg et al., 2014	High	High	High	Low	High	Medium
Capovilla, Capovilla, & Suiter 2004	Low	Medium	Low	Low	Low	Low
Guevara et al., 2008	High	High	High	High	High	Medium
Capovilla, Gutschow & Capovilla, 2004	Medium	Medium	Low	Low	Low	Low
Benítez et al., 2007	High	High	High	High	High	High
Silva et al., 2013	High	High	High	Low	High	High
Moneda et al., 2009	High	High	High	High	low	medium
Plana & Fumagalli, 2013	High	High	High	High	High	High
Fuller et al., 2009	Low	Low	High	High	Low	Medium
Janus, 2011	Low	High	High	Low	High	Low
Cueto & Diaz, 1999	High	High	High	Low	High	Medium
Kim & Pallante, 2012	Low	High	High	Low	High	High
Bravo, Villalón, & Orellana, 2002	High	High	High	High	High	High
Favila, et al., 1999	High	High	High	Medium	High	Medium
Pino & Bravo, 2005	Medium	High	High	High	High	High
Querejeta et al., 2013	High	High	High	High	High	Medium
Manrique et al., 1994	Low	High	Medium	Medium	High	Medium
Kudo & Bazan, 2009	Medium	Medium	High	Medium	High	High
Melchiori, 2012	Low	Medium	Low	Low	Low	Medium

Study	Theoretical framework explaining the study's motivation and findings	Quality of outcome measure	Sample selection quality	Quality of data collection procedures	Quality and relevance of analysis, given the research question	External validity
Abadzi et al., 2005	Low	Medium	Low	Low	Low	Low
Morales et al., 2013	Low	Medium	Medium	Low	Low	Medium
Reigosa-Crespo et al., 2013	High	High	High	High	Medium	Low
Oliveira, 1996	Low	High	High	High	High	Medium
Castro et al., 2002	Medium	High	Medium	Medium	Medium	High
Ramírez, Verdugo, & Sánchez, 2000	High	High	Low	High	High	Low
Reynoso-Alcántara et al., 2010	High	High	High	High	High	Medium
Salazar, Amon & Ortiz, 1996	High	High	High	High	High	Medium
Rosas et al., 2015	High	High	High	Low	High	High
Salazar-Reyes & Vega-Perez, 2013R	High	High	High	High	High	Medium
Rindermann, Baumeister, & Gröper, 2014	Low	Low	Low	Low	Medium	High
Silva et al., 2014	Low	Low	High	High	High	High
Reigosa-Crespo et al., 2013	High	High	High	Medium	High	Medium
Rego, 1997	Medium	High	High	Low	Medium	Medium
Treiman, Kessler, & Pollo, 2006	High	Medium	Medium	High	Medium	Medium
Kessler, Treiman & Cardoso-Martins, 2013	High	High	Medium	High	High	High
Correa & Dockrell, 2007	High	High	Medium	Medium	Medium	Low
Villalon & San Francisco, 2001	High	High	Medium	High	Medium	High
Hoddinott et al., 2013	Medium	High	Medium	Medium	High	High

Synthesis of Quantitative Nonintervention Studies

Multiple themes emerged from the corpus of nonintervention quantitative studies. These included preschool programs; preliteracy/emergent literacy; individual differences in reading skills; poverty; disability; and assessment validation. Although some themes were interrelated, others were multidimensional, cutting across different themes. For example, one study measured phonological awareness (PA) but also examined quality of the preschool program (Pino & Bravo, 2005). Another study investigated the factors that were associated with student reading ability and found that school-level factors (e.g., teacher quality and student abilities) predicted 40% of students' academic performance, while the authors reported that home factors (e.g., poverty) account for more variance in school performance (Ramírez, Verdugo, & Sánchez, 2000).

Preschool

In the sample, 17 of the 62 studies focused on the overarching theme of preschool programs including the importance of preschool (7 studies) and the quality of preschool programs (10 studies) (Pino et al., 2005). Studies featuring the importance of preschool ranged from those finding a correlation between literacy and other measures of cognitive development (comparing cognitive) and more years of preschool related to better academic outcomes (Benítez et al., 2007; Oliveira, 1996; Castro et al., 2002). The studies with an emphasis on the quality of preschool included studies related to programming and pedagogical practices (Bravo et al., 2002; Pino & Bravo, 2005) to type of school as measured by rigor of preschool program (Gómez-Pérez et al., 2011). Studies described characteristics of preschools in low socioeconomic areas (Silva et al., 2013), including teacher quality and materials used (Oyarce & Mujica, 2001), and teacher quality and parent education levels (Fuller et al., 1999).

Preliteracy/Emergent Literacy

Several studies focused on preliteracy skills and the importance of early exposure to print (Guardia, 2003; Kessler et al., 2013) and oral language development (Páez, Tabors & López, 2007) to reading acquisition. This finding is supported by other studies that linked oral language to reading and writing ability (Correa & Dockrell, 2007) and to the writing ability as a product of the sociocultural background of the student (Rebeiro et al., 2014). These findings suggest that students' reading and writing abilities are directly related to the level of oral language they have at school entry and the linguistic influences they have had before entering school. From these studies we find that the quality of the preschool program, the quality of the teachers, and the materials used are all associated with student achievement.

Reading Skills

Of the 62 studies, 22 studies involved a measure of one or more reading skills (e.g., phonological awareness, phonics, decoding, comprehension, vocabulary). Of these, 10 studies focused on some element of phonics and the alphabetic principle, including letter-sound correspondence rules, letter recognition, and word level reading. Study findings support the idea that students with better letter recognition skills can read better (De Abreu & Cardoso-Martins, 1998; Medeiros et al., 2011; Guardia, 2003). Taken together, these studies found that explicit teaching of letter-sound correspondence is associated with children's decoding skills (i.e., the connection between sounds

and symbols). An additional 9 of the 22 reading skills studies found a strong correlation between phonological awareness and reading ability (Bravo et al., 2002; Plana & Fumagalli, 2013). Several studies found that teaching PA and phonics is associated with student decoding skills (Reynoso-Alcántara et al., 2010; Manrique & Signorini, 1994). One study from Chile found that rapid letter naming and phonological awareness were the strongest predictors of reading ability even for children from low socioeconomic homes who had less exposure to print at home (Guardia, 2003).

Another study from Chile found that, although some students with strong PA skills become strong readers, some do not because other factors interact with reading such as the instructional methodology and student motivation (Muñoz, 2002). A third study from Chile found that PA, phonics, reading, and writing are all significantly correlated, supporting the belief that these skills may be interrelated (Villalon & San Francisco, 2001). The last four studies of reading skills centered on decoding and comprehension. Three of these studies investigated finding a relationship between fluency and comprehension (Kudo & Bazan, 2009; Abadzi et al., 2005) while one found a relationship between numerical fluency and reading fluency (Reigosa-Crespo, et al., 2013). All these studies are correlational and cannot be interpreted as causal evidence.

Although the studies consistently provided evidence for significant associations between phonemic awareness and early word reading skills, one study suggested that phonemic awareness-focused instruction may not be as useful for Spanish-instructed children as a *teaching approach*, as compared with English-instructed children (Goldenberg et al., 2014). When tested on phonemic awareness, Mexican students performed worse than students in the United States, although both groups were instructed in Spanish. The researcher suggests that this is a product of strong phonemic awareness instruction in the United States, after controlling for various other factors including parental education. Interestingly, children in the United States performed better on Spanish phonemic awareness, even though they were only provided phonemic awareness training in English, providing strong support for cross-linguistic transfer. Despite this advantage in phonemic awareness, however, the Mexican children outperformed the other students in later and repeated measures of reading, suggesting that phonemic awareness may not be as necessary for sustained teaching when learning a transparent orthography such as Spanish (Goldenberg et al., 2014).

Multiple researchers stated that there is a zone of proximal development for students to benefit from phonological awareness and early exposure to print to learn to read efficiently (Bravo et al., 2002; Guardia, 2003).

The findings indicate that teaching phonemic awareness, phonics, fluency, and comprehension is associated with reading ability, but it is unclear whether this relationship is causal, and for how long such teaching is likely to impact reading outcomes. Thus, there may be a positive effect of teaching these abilities on reading comprehension, but there are several confounding factors that could bias the relationship. Neither the quantitative intervention nor the quantitative nonintervention studies are able to provide conclusive evidence on the effects of teaching phonemic awareness, phonics, fluency, and comprehension on reading ability. This is an important gap in the literature on early grade reading in the LAC region.

Poverty

Of the 62 studies, six present an association between poverty and associated factors and the ability to read. One study (Guardia, 2003) from Chile found that young children have a natural disposition for development of psycholinguistic and cognitive abilities that support reading acquisition, but these children need a print-rich environment to benefit from being read to by parents. The authors suggest that there is a "zone of proximal" development for reading acquisition enhanced by explicit and systematic instruction in phonological awareness and, in particular, rapid letter naming that supports early reading ability. Children from impoverished homes are less likely to have either of these present in their homes. Similarly, another study from Chile (Bizama, Gutierrez, & Saez, 2011) found that poverty is adversely related to children's academic performance in reading, highlighting the educational inequalities that poverty creates. Two studies investigated the effect of child labor on reading achievement. Students who work more hours have the lowest student achievement (Cervini, 2015) and those who get paid to work tend to have worse academic outcomes than those paid in kind (Torrecilla & Carrasco, 2014). One study from Guatemala focused on the predictive effects of child nutrition on growth and cognitive achievement as well as later adult outcomes (e.g., wages for men, family formation, reproduction, and poverty) (Hoddinott et al., 2013). Taken together, these studies demonstrate the apparently long-lasting associations of poverty and school achievement and later life choices, especially through the relationships they have with access to educational resources.

In all, these studies indicate that poverty and reading ability are negatively correlated, which is supported by some of the quantitative intervention research. Both the quantitative intervention research and the quantitative nonintervention research suggest that poverty and associated factors, such as nutrition and child labor, are negatively associated with early grade reading outcomes. However, the evidence is less clear on the direction of these effects. Although poverty and reading ability are negatively correlated, the quantitative intervention studies only find evidence for a positive effect of nutrition programs in countries where the incidence of stunting and wasting is very high. In other contexts it remains unclear whether confounding factors bias the relationship between poverty and early grade reading outcomes.

Disability

Three of the 62 studies in the sample investigated reading ability in students with disabilities. One study from Brazil investigated reading ability in children with hyperlexia and found that these students showed a discrepancy between word decoding and reading comprehension and that these traits are also found in preschool-aged students (Cardoso-Martins, & Da Silva, 2010). Another study from Brazil compared the differences between how deaf children interpret illustrated text and construct writing to that of hearing children and found differences in the two groups. Bandini et al. (2006) studied how children who are deaf learn to read and found that the students who signed followed the alphabetic principle and used a pattern similar to non-deaf children.

Assessment Validation

Of the 62 studies, nine studies involved a form of assessment validation. For example, three studies (Athayde et al., 2014; Dias, Enumo & Turini, 2006) assessed the Teste do Desempenho Escolar (TDE) instrument that is widely used in Brazil. Study findings differed, with one study finding

that discrimination power of the writing subtest could not distinguish between students of similar grades (e.g., 3/4 and 5/6) (Athayde et al., 2014) and another finding only differences between fifth and sixth grade results (Dias, Enumo & Turini, 2006). Similarly, Athayde et al. (2014) found that the TDE test could only discriminate between scores of students in Grades 1–3 but not 4–6. These results indicate that the TDE test may be best when administered on the early grades (e.g., 1–3). Another study measured the predictive validity of the ABC test (Salazar, Amon, & Ortiz de Urdiales, 1996) and found that the test, although widely used, does not predict future reading ability in oral reading fluency or comprehension.

Several other assessments were also validated, with the TECOLESI test demonstrating strong correlations between phonological awareness and memory with reading ability (Capovilla, Capovilla & Suiter. 2004) The SAL test, a computer-based video game, also correlated with reading ability and was also described as able to reveal cognitive processing deficits in children (Reigosa et al., 2013).

Taken together, this set of studies on assessment validation provide a basis for thinking about how we define and assess reading outcomes in further research on early grade reading in the LAC region.

Qualitative Intervention and Nonintervention Research

The review of qualitative research on EGR interventions in the LAC region included eight articles from Argentina, Brazil, the Caribbean, Colombia, Jamaica, and Peru. Of these eight articles, only four were considered high quality and included in the synthesis of the findings. These four articles focus on bilingual/multilingual education in Peru (Neugebauer & Currie-Rubin, 2009), curriculum in Jamaica (Roofe, 2014), parental and community participation in Argentina (Stein & Rosemberg, 2012), and general pedagogical strategies in Colombia (González et al., 2013).

The review of the nonintervention qualitative research included 16 articles from Argentina, Brazil, the British West Indies, Colombia, Jamaica, Mexico, Puerto Rico, and Venezuela. Of these 16 articles, only 13 were considered high quality and included in the findings. These studies focused on: assessment in multiple countries (Leal Carretero & Suro Sanchez, 2012); pedagogical approaches in Brazil, Mexico, and Puerto Rico (da Cunha Lima Rosado & Holanda Campelo, 2011; Medina, 2013; Gómez Nashiki, 2008); parental and community participation in Jamaica and Puerto Rico (Kinkead-Clark, 2014; Volk & de Acosta, 2001, 2003); bilingual/multilingual education in Colombia (Guevara & Ordonez, 2012); reading skills in Argentina (Manrique & Borzone, 2010); teaching practices for reading in Jamaica, Mexico, and Argentina (Diuk, 2007; Jimenez et. al., 2003; Webster, 2009); and teacher training in the Caribbean (Warrican, Down, & Spencer-Ernandez, 2008).

Table 23 below shows a gap map of the qualitative intervention and nonintervention research. This map represents the gaps in research on specific topic areas as well as the gaps in research from particular countries. Only the medium- and high-quality studies are included. There are clear gaps in the qualitative research on the topics of: child nutrition, disabilities, ICT, preliteracy, preschool, poverty, school governance, and writing.

Table 23. Gap Map of Qualitative Intervention and Nonintervention Studies

Topics	Argentina	Brazil	Colombia	Jamaica	Mexico	Peru	Puerto Rico	Multiple
Assessment								Carretero & Sánchez, 2012
Child Nutrition								
Curriculum				Roofe, 2014				
Disabilities								
General Pedagogical Approaches		da Cunha Lima Rosado & Holanda Campelo, 2011	González et al., 2013		Gómez Nashiki, 2008		Medina & Costa, 2013	
ICT								
Parental and Community Participation	Stein & Rosemberg, 2012			Kinkead-Clark, 2014			Volk & de Acosta, 2001, 2003	
Poverty								
Preliteracy/Emergent Literacy								
Preschool								
Reading in Bilingual/ Multilingual Contexts			Guevara & Ordoñez, 2012			Neugebauer & Currie-Rubin, 2009		
Reading Skills	Manrique & Borzone, 2010							
School Governance								
Teaching Practices for Reading	Diuk, 2007			Webster, 2009	Jiménez, Smith, & Martínez-León, 2003			
Teacher Training								Warrican & Spencer-Ernandez, 2008
Writing								

Red text = Qualitative intervention studies
Blue text = Qualitative nonintervention

The following section describes the quality of qualitative intervention and nonintervention research using the narrative synthesis techniques outlined above. We then provide a synthesis of the study findings by topic area.

Research Design

In qualitative research, questions—and the responses they elicit—tend to be discursive and descriptive, and the analysis helps to provide an explanation and interpretation for quantitative results. In general, qualitative approaches allow researchers to explore and understand the experiences, opinions, and perspectives of informants in greater depth than that offered by quantitative approaches. In turn, the use of qualitative approaches requires sacrifices in terms of generalizability and comparability—areas in which quantitative methods excel because of their use of large and probabilistic samples. Finally, qualitative inquiry allows researchers to look at mechanisms of impact that address how participants interact with intervention activities, as well as external or contextual factors that influence program implementation.

With these goals in mind, researchers must identify which qualitative methods are most appropriate for a particular program. Methods include, but are not limited to, case studies, focus group discussions, observations, key informant interviews, and participatory assessments. Generally, when it comes to the strength of a statement of research, the desire to make the research questions generalizable helps guide not only the selected methods but also the estimate of appropriate sample size. Unlike quantitative research, qualitative sampling is less focused on large numbers and more focused on purpose and saturation. That is, given the burden of the research on its participants, it is important to sample participants until additional information no longer adds to the knowledge base for the study. A thorough and transparent description of the research design and a written justification of the methods and sampling strategy adds credence to the quality of a study. We discuss the quality of the qualitative intervention research in this section through a summary and analysis of the research designs, ethics and reflexivity, and the relevance of the research to the field.

Statement of Research

A clear statement of purpose forms the basis for how a researcher then decides on methods, measurement, and analysis of a problem (Ford, 2009). Our review assumes the purpose of the research, or problem statement, "may be phrased as statements of research purpose, as specific research questions, or as research hypotheses, depending on the purpose of the study and selected design" (McMillan, 2001, p. 86). A research statement serves to introduce the reader to the research, provide context, and create a framework in which to report results that in the end guide the entire exercise (Bryman, 2007). We rated the quality of the research statement on the following parameters:

Quality Review Criteria
• **Clear statement of research** – The goal of the research – Why it is important

Qualitative Intervention. Reviewers rated the clarity of the stated goals as "high" on six of the eight articles. For example, Castanheira, Neves, & Gouvêa, (2013) state, "The purpose of this article is to describe a read-aloud program that incorporated research-based practices in an indigenous setting and to report preliminary findings about the effectiveness of the program as a predictor of reading comprehension" (p. 297). Here both the goal and the methods by which the goal will be realized are clearly stated in the text. Successful research statements also justify goals by explaining their importance. In comparison, weak goals are not clearly articulated or contradict other portions of the text. For example, Mahurt (1993) and Caldera de Briceño, Escalante de Urrecheaga, & Terán de Serrentino, (2010) do not include an explanation of the programs they are evaluating anywhere in the text. Furthermore, there are other instances where the research methodology does not align with the stated goal (Caldera de Briceño, Escalante de Urrecheaga, & Terán de Serrentino, 2010).

Effective statements of importance not only explain why the research is necessary but also show why findings would be important within the research context as well as within the larger community of stakeholders. Neugebauer and Currie-Rubin (2009) successfully demonstrate the importance of their research in Peru through the following statement:

> *"The need for research focused on read-alouds in such communities is particularly compelling given the nature of read-aloud pedagogy (the integration of oral elaborations of text and vocabulary with written narratives) and the tradition of oral story telling that is central to many indigenous cultures. Given the strong emphasis in these communities on oral histories as a means of "communicat[ing] ideas and images" (Mello, 2001, p.1), read-alouds can extend the connection between oral narratives and written genres. Furthermore, this instructional format includes community experiences and simultaneously provides a wealth of language-rich pedagogy especially useful for bilingual populations" (p. 297).*

In this passage, Neugebauer and Currie-Rubin (2009) explain the relevance of the research for the local communities as well as how the research would be applicable to the larger field, particularly bilingual populations. Of the surveyed articles, the majority communicated the importance of their stated goals.

Qualitative Nonintervention. Similar to the qualitative intervention articles, nearly all qualitative nonintervention articles clearly stated the goal of the research. Reviewers rated the quality of 11 articles as "high," 4 articles as "medium," and 1 article as "low" quality on the clarity of the research goals. The articles where quality was rated high clearly stated the goal and wove the goal throughout the article. For example, one of the high-rated articles states, "The purpose of this article is to describe the instructional strategies and related activities that Ms. Bingham and I used as we integrated language arts and science into her literacy instruction" (Webster, 2009, p. 662). However, articles where quality was rated as low did not clearly state their goal or did not weave the goal throughout the article.

The majority of nonintervention articles also effectively communicated the importance of the stated goal. Reviewers rated 13 articles as either "high" or "medium" quality for demonstrating the importance of the research goals. Articles rated as high quality showed importance by highlighting gaps in the existing literature or situating the research within continuing challenges

to early grade reading. For example, Manrique and Borzone (2010) argue that their research in Argentina is necessary because the existing literature does not explain the difficulties that children from marginalized sectors have in processing process written narratives. Another strong article, "Teaching English to Very Young Learners," expressed the importance of the research by describing the continuing challenges facing second language acquisition in Colombia. The authors state, "In this present reality, the search for effective ways to teach an unfamiliar language to young children has become a priority, and it constitutes a challenge in diverse Colombian sociolinguistic contexts, where the use of a second international language is not necessary and, therefore, not naturally stimulated" (Guevara & Ordonez, 2012, p. 11). The authors communicate the necessity of the research by positioning it within the existing literature and socio-political context.

Methodology

The chosen method of qualitative inquiry defines results as "when a finding is extracted, the perspective or context that the study author intended for the finding is not lost but is embedded in the extraction" (The Joanna Briggs Institute, 2014, p. 18). Importantly, clear explanations of how research was conducted and what methods were used affects the way the reader will interpret the findings derived from that research. Strong research methodologies are guided by clearly presented research questions or hypotheses, an explanation of the research methods, and a justification for why particular methods are used. We assessed the quality of the papers' methodologies to the extent they were described using the criteria below:

Quality Review Criteria

- **Appropriateness of qualitative methodology**
  - Does the research interpret or illuminate the actions and/or subjective experiences of research participants?
- **Research design addresses the aims of the research**
  - Is the research guided by research questions or hypotheses?
  - Has the researcher justified the research design? (i.e., have they discussed how they decided which methods to use)?

Qualitative Intervention. Reviewers rated two qualitative intervention articles as "high," one article as "medium," and one as "low" quality on including research questions or a hypothesis, while four articles did not include research questions or a research hypothesis at all. In articles that included strong research questions or hypotheses, the research questions or hypotheses were explicitly stated in the text and guided the overall research. In comparison, low performing articles included research questions that were not well formulated or did not align with the data researchers collected. Furthermore, four articles did not present research questions or hypotheses, making it difficult to understand the direction of the research as well as the justification for the design, which compromises the quality of these studies.

The majority of included studies failed to explicitly convey the methodologies used in the research. One article scored high, two scored medium, two scored low quality, and three papers did not discuss methodology. Strong articles clearly articulated the methodology including the methods used, rationale for using particular methods, and an explanation of how the researchers used the methodologies. For example, Belintane (2010) stated that the intervention was based on the

weaknesses in student literacy that researchers discovered during the primary assessment. Surveyed research papers used a variety of methodologies including observations, case studies, qualitative interviews, and journaling. Overall, only one study altered the methods during the evaluation to reflect more of a case study format. The other studies (n = 7) do not report any modification to the methods.

All of the surveyed papers adequately justified the use of qualitative methods. Reviewers rated four articles as "high" and four articles as "medium" on appropriateness of qualitative methodology and research design. Compelling justifications explained how the research aimed to achieve its goals through an understanding of the subjective experiences of teachers and students. For example, Mahurt (1993) used a case study to provide intensive, in-depth exploration using a hermeneutic phenomenology theoretical framework. However, only three of the eight articles scored high on research methodology justification. The other articles did not explain how methodologies were used and why. For example, the article by Roofe (2014) does not explain why focus group discussion or semi-structured interviews were chosen or why the study was limited to only 11 teachers. Many articles either loosely articulated why certain methods were used over others or omitted important details about their methodology. Overall, justifications of research methodologies could be bolstered through a description of why certain methods were selected over others as well as how methods were used in relationship to each other.

In some instances, the use of qualitative methods may not have been appropriate for achieving the goal of the research. For example, Castanheira, Neves, & Gouvêa, (2013) examined the effectiveness of a read-aloud program. However, in general, qualitative methodologies are less appropriate for determining the effectiveness of development programs than quantitative or mixed-methods methodologies because qualitative methods do not allow for determining the causal effect of development programs.

Qualitative Nonintervention. More than half of qualitative nonintervention articles used clearly stated research questions to guide the text. For example, Webster (2009) states, "What is the influence of teacher read-alouds of informational texts on grade 1 students' science learning as revealed through their drawings and written retellings?" (p. 663). Other articles either included vague research questions embedded in the text, used exploratory research designs that do not necessarily require research questions, or did not include research questions.

Qualitative nonintervention articles successfully supported the use of qualitative methodologies but could provide greater detail to justify the use of specific methods. The majority of surveyed articles (n = 13 of 16) effectively used qualitative research to illuminate the actions and subjective experiences of the research participants. The articles included a variety of subjective experiences and perspectives including students' interactions, reactions to particular texts, and perspectives on curricula as well as teachers' actions, goals, reflections, and perspectives on curricula. However, a minority of articles (n = 8) explicitly stated the research methodologies used in their respective studies, and none of the surveyed articles discussed modifying their methods. Furthermore, eight articles either included an incomplete discussion or explanation of why particular methods were chosen (Kinkead-Clark, 2014), lacked theoretical support for the chosen design (Rosado & Campelo, 2011), or included no explanations of the methodological choices (Ribeiro & Souza, 2012). Similarly, only nine of the surveyed articles included justifications for why particular

methods were best positioned for particular goals and contexts, and none of the articles explained how researchers triangulated multiple methodologies.

Data

Describing methodologies also entails detailing the setting, justification, process, and the form of data collected. This information is particularly important to include as data in qualitative research, where context can determine participants' level of comfort with participation especially in cases where participant observation is part of the methodology. In addition, a discussion of saturation—the point at which researchers gain no new information from an additional data point—typically serves as the justification for a study's numerical sample. Reviewers accounted for the following elements when rating a study on data quality:

Quality Review Criteria

- **Was the data collected in a way that addressed the research issue?**
  - If the setting for data collection was justified
  - If it is clear how data were collected (e.g., focus group, semi-structured interview, etc.)
  - If the researcher has justified the methods chosen
  - If the researcher has made the methods explicit (e.g., for interview method, is there an indication of how interviews were conducted, did they use a topic guide?)
  - If methods were modified during the study. If so, has the researcher explained how and why?
  - If the form of data is clear (e.g., tape recordings, video material, notes, etc.)
  - If the researcher has discussed saturation of data

Qualitative Intervention. Evaluators rated six of the qualitative intervention articles as "high" and two as "medium" on presenting details of data collection. Articles rated medium did not present data collection protocols or articulate the length or timing of data collection. Although all articles touched on the data collection setting, only four described the data collection context. Many articles rated as low on this measure did not explain the importance of the site or include a justification for why a particular site is most relevant for the evaluation. Finally, none of the articles included a discussion of data saturation; this discussion may have helped the reader understand cases such as in the study of Roofe (2014) in Jamaica, which included only 11 interviews. This number of interviews could have been sufficient for the study, but a discussion of saturation or selection process would strengthen the article's scientific validity.

Qualitative Nonintervention. Qualitative nonintervention articles should clearly explain and justify the data collection site. Of the 16 articles reviewed, 10 effectively justified and explained the data collection site. For example, Kinkead-Clark (2014) selected the Turtle Islands because it is a diverse cultural setting that offers insight into the role of culture in literacy. Furthermore, the researcher was a teacher in the selected classroom, which allowed her to have increased access to the student participants (Kinkead-Clark, 2014). Articles that include weaker explanations of the data collection site often require additional support to justify the site selection (Warrican et al.,

2008) or lack sufficient detail (Gómez Nashiki, 2008; Rosado & Campelo, 2011). For instance, Rosado and Campelo (2011) state that data collection took place in a school because the research required the study to take place in a school. However, the researchers did not provide a justification for why the particular research schools were selected.

Similarly, the majority of surveyed articles successfully described the type and form of collected data. 11 of the 16 articles described the form of data and 11 also described how researchers collected the data. Although articles rated as "low" quality often lacked details of the data collection process, strong articles included clear descriptions of how researchers collected the data as well as the type of data collected. For instance, Volk & de Acosta (2001) state:

> *From January through to the end of the school year, we observed and audio taped in the classroom twice a month for the three-hour morning session and for about an hour after lunch; times when most literacy events occurred. We observed and taped in each home once a month for between two and four hours at a time. Observations and interviews were conducted in two of the churches and their Sunday schools; interview data were collected about the other church and Sunday school (p. 197).*

Finally, although many of the qualitative nonintervention articles effectively described the data researchers used as the foundation for the analysis and findings, none of the surveyed articles discussed data saturation.

Data Analysis

The objective of qualitative analysis is to ensure that data analysis is consistent with the goal of ensuring data trustworthiness and, thus, credibility of the findings. Qualitative data analysis is by nature not as systematic as quantitative data analysis, but it can become systematic through methodical coding using an iterative process that promotes consistency in all facets of data collection, analysis, and reporting. The process of deducting themes from the data may be subjective, which is why various quality checks and procedural considerations—including open coding, interrater reliability checks, and continuously defining codes and sub-codes among a team of researchers—are typically included in the analysis of qualitative data. Triangulation techniques (Denzin, 1978), including methodological triangulation (Guba & Lincoln, 2005; Lincoln & Guba, 1985) and triangulation among raters, also support efforts to promote the integrity of the overall research. Given the risk of subjectivity even while using these techniques, researchers can also lend credibility to findings by discussing contradictory data, potential biases, and contextual factors that may give the reader a more holistic view of the results. We reviewed the quality of qualitative data analysis for the included articles on the following criteria:

Quality Review Criteria

- **Was the data analysis sufficiently rigorous?**
  - if there is a thorough description of the analysis process
  - if thematic analysis is used. If so, is it clear how the categories/themes were derived from the data?

- if the researcher explains how the data presented were selected from the original sample to demonstrate the analysis process (e.g., I chose this because 90% of the participants said something similar)
- if sufficient data are presented to support the findings
- if contradictory data are taken into account
- whether the researcher critically examined their own role, potential bias, and influence during analysis and selection of data for presentation
- if the researcher considered contextual factors that may have influenced the research results (if you do a study in Peru, you must take into consideration context of Peru, Urban vs. Rural, etc.)

Qualitative Intervention. Out of eight articles, only two articles received high ratings for their description of the analysis process, while six did not discuss this process in detail. Four articles used thematic analysis and of these four, two used thematic analysis effectively—that is, the articles used themes to guide the analysis process and supported these themes with data. However, in articles with a low rating on this domain, themes were not clearly used to guide the analysis and overall conclusions (Caldera de Briceño, Escalante de Urrecheaga, & Terán de Serrentino, 2010), or there was no explanation of how themes emerged from the data. Four of the eight articles used sufficient data in their analysis; however "sufficient" is dependent on the parameters of the research study. For example, Mahurt (1993) used limited but sufficient data sources because the research aimed to look at the struggle of a single teacher trying to enact behavior change. Furthermore, only one article explained how researchers selected the data presented in the article from all of the collected data and only two articles included discussions of contradictory data. Contradictory, minority results are important to note to demonstrate that all findings are taken into account. Failing to report contradictory results may be an indication for a bias in the research findings. Three of the eight articles included a consideration of the context in their analysis. For example, the article "Orality, Literacy and Reading: Differences and Complexities Facing the Public School" highlights the importance of context through its description of other development programs in the area including the Ler e Escrivir project. Context is important to consider in this case because some of the changes described in the article could have been a result of the other intervention.

Qualitative Nonintervention. The qualitative nonintervention articles could improve the description and execution of the data analysis. Around half of the articles (n = 9) included a thorough description of the data analysis process. Thorough descriptions explicitly stated the relevant analytical process in sufficient detail for the reader to understand how researchers translated data into findings. For example, Leal Carretero & Suro Sanchez (2012) described their analysis by presenting a comparative table with the characteristics of the tests given to participants then followed up with a categorical analysis (pp. 738–739) in their study on literacy assessments from multiple countries. Although 12 articles reported using thematic analysis, only eight of those did so effectively. Furthermore, 15 of the surveyed articles used sufficient data in their analysis process but only five articles presented data to demonstrate the analysis process. Lower performing articles included analyses that are hard to follow (Medina & Costa, 2013) or lack sufficient detail (Gómez Nashiki, 2008; Guevara & Ordonez, 2012; Warrican et al., 2008).

The qualitative nonintervention articles failed to adequately report the limitations and context of the data used in the analysis. Nine articles included some mention of the research context.

However, only five articles included a discussion of how researchers' bias may have affected the data analysis process. These articles positioned the research within the analytical process, stating how their background may predisposition them to particular findings. One article included a weak discussion of researcher bias, and the remaining 11 articles did not discuss potential biases in the analysis process. Finally, none of the 16 articles presented any information regarding the consideration of contradictory data.

Statement of Findings

A primary goal of research is to translate data into accessible findings and practice among non-researcher audiences—especially in the case of the LAC Reads Capacity Program. Findings should explicitly relate back to the purpose of the study and directly answer the research questions, while also phrasing them in a way that makes the level of reliability and transferability explicit. We rated articles' statements of findings on these parameters:

Quality Review Criteria

- **Is there a clear statement of findings?**
  - if the findings are explicit
  - if there is adequate discussion of the evidence both for and against the researcher's interpretations
  - if the researcher has discussed the credibility of their findings (e.g., triangulation, respondent validation, more than one analyst)
  - if the findings are discussed in relation to the original research questions

Qualitative Intervention. The majority of the selected papers clearly presented findings, but they could have provided more information about how researchers arrived at the findings. Three articles discussed findings in relation to their original research questions or the findings were in direct conversation with them; three did not discuss their findings in terms of the research questions; and one was exploratory and therefore did not have research questions. The majority of articles did not include a discussion of triangulation, respondent validation, multiple analysts, or evidence against interpretations. Only one article included evidence that contradicted the findings of the research. Furthermore, two articles did discuss credibility; one article used the qualitative research to supplement the quantitative research findings (Neugebauer & Currie-Rubin, 2009) and another triangulated results through multiple qualitative methods (Mahurt, 1993).

Qualitative Nonintervention. The qualitative nonintervention articles successfully communicated findings but could bolster the credibility of findings through triangulation and the presentation of contradictory data. The reviewers rated seven articles as "high," five as "medium," and two as "low" on explicitly stating findings, and three articles did not include a clear statement of findings. Articles rated as high clearly articulated findings that linked to the research questions, theoretical framework, context, and analysis (Manrique & Borzone, 2010; Webster, 2009). Although a minority of the articles (n = 5) linked findings to the original research questions, this type of presentation improves the organization and flow of the text for the reader (Guevara & Ordonez, 2012; Medina & Costa, 2013; Volk & de Acosta, 2001; Webster, 2009). Only two articles discussed evidence against the findings and only three discussed triangulation. Articles rated as

high typically triangulated findings using multiple data sources (Medina & Costa, 2013), or multiple researchers (Jimenez et al., 2003). For example, Webster (2009) triangulates her findings between the students, the teacher, her observations, and observations of the assistant principal.

Ethics and Reflexivity

Reviewers assessed the quality of an article's transparency on ethics based on its described recruitment strategy, its recognition of potential bias in the researcher-participant relationship, and its attention to protection of human subjects in research.

Recruitment Strategy

In addition to the importance of considering saturation and strategy to the extent possible in qualitative sampling, researchers should explain the criteria they use for participant recruitment and how they select and exclude potential participants. Recruitment should explain why the informants are best suited to inform the research using explicitly stated criteria. In particular, in cases where the intervention research involves selecting some potential participants over others, the researcher must explain why the selected participants were appropriate over others who may have benefitted from the intervention. Finally, "convenience sampling" (Patton, 2015a, p. 209), or sampling participants based on who is easiest to access, poses threats to study validity by ignoring inclusion criteria. Recruiting participants who are the most convenient to reach also diminishes the quality of the study, as the information from convenient samples is not likely to be the most valuable information (Patton, 2015a).

Quality Review Criteria

- **Appropriate recruitment strategy**
  - if the researcher has explained how the participants were selected
  - if they explained why the participants they selected were the most appropriate to provide access to the type of knowledge sought by the study

Qualitative Intervention. The qualitative intervention articles included limited information on recruitment strategies. Only four articles described how participants were selected. For example, Mahurt (1993) clearly states that participant selection was based on the following criteria:

> *(a) a teacher who had made a recent decision to change to whole language; (b) a teacher whose decision to change was based on personal factors and not influences from graduate courses or mandates from the school district or administrator; (c) a teacher who seemed interested enough in whole language instruction to continue for at least two years (p. 8).*

Furthermore, only three articles explained why researchers selected certain participants over other individuals.

Qualitative Nonintervention. Nine of the qualitative nonintervention articles included an explanation of how researchers selected participants. Volk & de Acosta (2003) explained that they

chose to include three children in their study in Puerto Rico to balance the need for rich description of a variety of literacy experiences with the constraints of equipment and time. Furthermore, the researchers selected participants in consultation with their teacher and based on information from observations, an assessment conducted by the teacher, and an informal reading assessment. Thus, the researchers demonstrated the process used for selection as well as what type of criteria were involved. However, the majority of articles included an insufficient explanation of the method used to identify the study population (e.g., Kinkead-Clark, 2014; Rosado & Campelo, 2011). Furthermore, the majority of articles (n = 11) did not include an explanation of why particular participants were chose over other participants.

Research-Participant Relationship

Researchers' positionality may affect the formulation of research questions or the interpretation of data. This does not necessarily affect the research, but outlining and recognizing the potential for the researcher to influence the process reflects an attempt at neutrality. Again, given the risk of subjectivity throughout qualitative research, "the issue in fieldwork is avoiding judgment so as to be open to deep and meaningful understanding of another" (Patton, 2015a, p. 57). In addition, a researcher has the potential to influence respondents during data collection by asking leading questions or by holding the interview in a place that does not allow the participant to feel comfortable (Phillips & White, 2012). Ways to mitigate these potential biases include holding the interview in a space that is familiar for the participants, grouping participants with familiar counterparts, and having a local researcher conduct the interview so the researcher is familiar with the context and can use the local dialect. We evaluated the assessment of researcher-participant bias using the following criteria:

Quality Review Criteria

- **Has the relationship between the researcher and participants been adequately considered?**
  - Consider if the researcher critically examined their own role, potential bias, and influence during:
    - » Formulation of research questions and research instruments (e.g., asking leading questions)
    - » Data collection, including sample recruitment and choice of location

Qualitative Intervention. Only one article included a discussion of subjectivity and positionality in the formulation of research questions. The remaining articles did not acknowledge how researchers' bias may affect the formulation of research questions or instruments or how researchers' involvement in "interpreting" questions for participants may have led the participants to a certain answer. Further, only one article mentioned the potential for researcher bias in the data collection process.

Qualitative Nonintervention. The majority of articles that touched on potential biases focused on how researchers influenced the site selection, while a small number of articles discussed researchers' bias in the sampling and recruitment of participants (Jimenez et al., 2003; Kinkead-Clark, 2014; Medina & Costa, 2013; Webster, 2009). Only six of the articles discussed the

researchers' bias in the data collection process. Bias can influence a number of factors during data collection including the sampling, recruitment, and site selection. Seven of the articles included a discussion of the researchers' bias in the formulation of research questions. In "Teaching English to Very Young Learners," the researchers disagreed with the school's early introduction of English as a second language, a concept which they are aiming to better understand. This bias was crucial to present within the text as the authors cannot fully remove this bias from their analysis. However, many articles did not present any information about how the researchers' bias may have affected the various research components. Finally, the majority of articles did not mention any bias in the data analysis process and only four included a discussion of subjectivity or positionality.

Ethics

Ethics review committees review research to ensure the protection of human subjects. The effort to enforce ethical review in all aspects of research stems from human rights violations in earlier research experiments such as research conducted by the Nazis during World War II or the Tuskegee Syphilis Study of 1932. Ethical committees (sometimes known as Institutional Review Boards) standardize the requirements that "fully inform and protect" participants and serve as an "affirmation of our commitment to treat all people with respect" (Patton, 2015a, p. 341). In the United States, for example, all social research is required to undergo review by an official ethical committee. Although there is no overarching ethical review board covering the entire LAC region, individual institutions, universities, and publications have their own ethical review boards and ethics codes with similar standards that researchers should follow. As a standard protection for human subjects, the CASP qualitative research checklist recommends assessing ethics on the following dimensions:

Quality Review Criteria
• **Have ethical issues been taken into consideration?** – if there are sufficient details of how the research was explained to participants for the reader to assess whether ethical standards were maintained – if the researcher has discussed issues raised by the study on sensitive issues (e.g., issues around informed consent or confidentiality or how they have handled the effects of the study on the participants during and after the study) – if approval has been sought from an ethics committee

Qualitative Intervention. None of the articles included a description of how researchers explained the study to participants, any reference to working with an IRB or seeking ethical approval, or a discussion of sensitive issues raised by the study. Ethical standards serve the critical role of protecting informants, particularly vulnerable informants such as children. We recognize, however, that reporting standards vary greatly by field such that an economics journal, for example, might not require any mention of ethical procedures whereas a medical journal would surely require it. Thus, although several of the studies do not report on seeking ethical approval, this does not necessarily mean that they did not obtain it.

Qualitative Nonintervention. As with the intervention articles, qualitative nonintervention articles included only limited discussions of ethical issues related to the research. Only two of the

surveyed articles mentioned obtaining consent from participants and only one article mentioned conducting research through an IRB. The vast majority of articles made no reference to ethical approval or issues of consent. Furthermore, none of the articles included a discussion of how researchers dealt with sensitive issues or took precautions to ensure the well-being and security of participants. Most of these studies did not cover data that would be considered highly sensitive, although many did work with children, who are considered a vulnerable population. Because most of the reviewed articles did not report on how ethical issues were addressed, it is difficult to say whether or not researchers took into account ethical considerations and to what extent. These procedures are sometimes not reported on in publications because they are so standard that it is assumed that one has completed them. In addition, researchers would not have been required to undergo IRB approvals for some of these studies as they made use of publicly available secondary data sets.

Relevance to the Field

Finally, raters reviewed qualitative intervention and nonintervention articles on two dimensions that describe how the study is relevant to the field. First, the study should add new information to the existing body of literature and policy documents. Second, the research should discuss the potential for study replicability and how replicability could add to the current findings—that is, how the study could be improved or adapted to further gain insight on the topic of research.

Situating the Research

Situating research within the existing body of literature entails discussing existing knowledge, understanding, and views on the topic. Researchers should describe how the research question adds to what exists and why it is necessary to understand. We assessed this discussion based on the following criteria:

Quality Review Criteria

- **How valuable is the research?**
  - if the researcher discusses the contribution the study makes to existing knowledge or understanding (e.g., do they consider the findings in relation to current policy or relevant research-based literature?)
  - if they identify new areas where research is necessary
  - if the researchers have discussed whether or how the findings can be transferred to other populations or considered other ways the research may be used

Qualitative Intervention. Reviewers rated four of the eight qualitative intervention articles as "high" on communicating the value of the research. The other four articles did not effectively contextualize findings within the existing literature or explicitly state the relevance to readers or the larger field. For example, although Stein and Rosemberg (2012) do not discuss how the study contributes to existing knowledge or understanding, they do discuss how this research could speak to existing theory around students' learning to write in English. Another way to communicate relevance is through a discussion of how the research can be applied in other contexts. Only two articles included this discussion, while six did not.

Finally, the majority of articles did not identify areas for further research. The two articles that effectively communicated areas for new research suggested expanding the current study (Mahurt, 1993) and continuing research on read-aloud efficacy in international contexts (Neugebauer & Currie-Rubin, 2009). However, four articles did not discuss areas for further research and two articles discuss additional research topics in an unclear manner.

Qualitative Nonintervention. Overall, the qualitative nonintervention articles consistently situated the research within the existing literature and intellectual field. The articles discussed the contribution to existing knowledge, identification of areas for further research, and how the findings could be used. Articles contribute to existing knowledge by supporting existing claims, expanding on existing research, or filling in gaps in the current literature. Eleven articles discussed how the findings contributed to existing knowledge, including both existing literature and education policies. For example, Volk & de Acosta (2001) state,

> *Previous research has emphasized matches and mismatches between teaching and learning practices in homes and classrooms. Often, mismatches are identified as causes or correlates of the low achievement levels of children who come from diverse cultures. But while continuity is an admirable goal, the complex and shifting relationships between literacy practices in these three homes and in this bilingual classroom suggest that an analysis limited to matches and mismatches is oversimplified and misleading. A broader view of literacy that encompasses many literacies that are similar in some ways and different in others may be more appropriate and, ultimately, more useful for teachers (p. 220).*

Similarly, Gómez Nashiki (2008) suggests that it would be helpful to encourage the formation of a network in charge of analysis and reflection on reading within Colima state in Mexico. In contrast, very few articles suggested areas for further research. In fact, the majority of articles (n = 12) did not include any mention of areas for further research.

Replicability

We assessed replicability based on two dimensions: first, whether stakeholders could replicate the program; and second, whether researchers provided sufficient information for other researchers to replicate the study in different contexts. Typically, systematic reviews with an emphasis on qualitative research assess replicability only on the research design dimension; however, given the context of our review and the end users, we also assessed replicability of the program so that stakeholders could independently consider whether example programs may fit their particular context and adapt the program to improve implementation. For example, length of trainings, monitoring tools, and training materials could be adapted to other contexts if included. We used the following criteria to assess replicability:

Quality Review Criteria

- **Information for stakeholders to assess replicability**
  - Does the paper provide adequate details on the design and implementation of the intervention to enable replication, such as:
    - » Length of training
    - » Monitoring tools

» Training materials

Qualitative Intervention. For the most part, we found that studies did not include enough information on the program being evaluated for a practitioner to be able to adapt the program to their context. For example, Belintane (2010) described some of the intervention techniques but did not provide details about how these elements were implemented into specific aspects of the classroom. In addition, the author did not provide tools for how to implement or monitor the techniques. Two articles provided enough information to repeat the described studies. Neugebauer and Currie-Rubin (2009) explained exactly how each of the seven techniques described in their article were used and could be easily adapted and used in the classroom. Furthermore, González et al. (2013) provided descriptions of the types of collaborative learning strategies researchers implemented in the study classroom; however, there were no explicit statements about the length of the training, the tools or instructional methods used, or the training materials for teachers to be able to implement the methods.

Similarly, a study's replicability depends on whether the researcher includes adequate details on the study design, including much of the quality criteria we previously discussed. Based on our assessment of the prior dimensions of the quality review, the majority of articles did not include enough information to easily replicate the studies that were discussed. Many articles were strong in some dimensions of quality, but these same articles excluded other elements that would be essential for replication. For example, four of the articles do not present methodological protocols, explanations of how methods were actually implemented, nor training materials (Caldera de Briceño, Escalante de Urrecheaga & Terán de Serrentino, 2010; Castanheira, Neves, & Gouvêa, 2013; Mahurt, 1993; Stein & Rosemberg, 2012).

Qualitative Nonintervention. None of the qualitative nonintervention articles discussed how findings can be transferred to other populations or used in other ways. Reviewers rated two articles as "low," three as "not applicable" (as they were ethnographic studies), and 11 articles did not include any information about the transferability of findings. Volk & de Acosta (2001) discussed how findings could be used to improve teacher practices, Jimenez et al., (2003) discussed the implications of the research, and Guevara and Ordonez (2012) discussed how their findings might be relatable to similar contexts. For example, Guevara and Ordonez (2012) offered the following advice for bilingual schools in other monolingual contexts:

> *It is also essential that children always understand what they are doing and saying in the foreign language and that they also do it in Spanish. The effective, conscious use of the students' knowledge of their first language is a must in helping our monolingual children become good consecutive bilinguals; and a truly bilingual curriculum may be a much better way than what we know as bilingual education to work towards bilingualism at school in monolingual environments (p. 22).*

Examples of how the findings can be applied in different contexts help make the findings relevant to practitioners in the region.

Synthesis of Qualitative Studies

In this section, we present the results of studies that were rated as high and medium quality by EGR topic area. The syntheses include both qualitative intervention and nonintervention research and highlight the main findings across the articles within our topic areas. Many of the topic areas did not have studies that made the final cut for inclusion in the review which can be seen in the gap map (Table 23).

Assessment

One qualitative nonintervention article focused on literacy assessments from multiple countries (Leal Carretero & Suro Sánchez, 2012). Researchers analyzed 21 different tests with measures of phonological awareness that were gathered through a detailed literature search with specific inclusion criteria. Tests had to target Spanish-speaking preschool children and include specific questions focusing on phonemic awareness. The researchers found 26 unique tasks among the 21 tests that measured phonological awareness. Among the 26 tasks, nine were productive tasks such as repeating syllables or constructing words from a sequence of word segments. Nine tasks involved implicit categorization such as identifying the number of syllables in a word or the number of words in a sentence. The remaining eight tasks involved explicit categorization such as categorizing the words with the same syllable or categorizing words with the same ending. The fact that there was so little coherence among the 26 tests and such a wide variety of tasks indicates that there is little consensus as to which tasks most accurately measure phonological awareness. In addition, many of these tasks were very prone to errors and often did not even measure phonological awareness because of the way that the tasks were worded. Tests did not measure syllable structure, or subsegmental, melodic, metrical, or intonation awareness, all of which could be useful measures of phonological awareness. Findings from this review of literacy assessments on phonological awareness indicate that:

- Phonological awareness tests should systematically include tasks that measure students' awareness of syllable structure (i.e., each syllable has a hierarchical organization formed around a core vowel).
- Current testing may be enriched by adding tasks for metrical awareness.
- Tests could be enriched by adding tasks for either intonation awareness or melodic awareness.
- Tests might be enriched by adding tasks for subsegmental awareness since a segment is not indivisible but instead has distinctive sound features.

Curriculum

The team found only one qualitative intervention article on curriculum. It focused on the implementation of Jamaica's revised primary curriculum in 2014. Although the article is specific to the Jamaican context and has some gaps in information about the data collection methods, the authors recommended some principles that could be applied to a wide range of contexts. For example, the authors pointed to a need for alignment between pedagogical and assessment practices for new curriculum; a rigorous implementation plan for training teachers and principals

who will use the curriculum; a monitoring and evaluation (M&E) system to hold individuals accountable; and finally, training materials that provide sample lesson plans and examples of how users can adapt curriculum to suit their contextual needs. However, although the curriculum aims to emphasize literacy development as a "key indicator of improved quality education," the authors determined that parts of the curriculum "disadvantaged students with low ability levels" in literacy development, as well as students from rural areas on topics for writing activities (p. 4). This finding is consistent with the theme we identified elsewhere in qualitative and quantitative studies: that poverty is a strong contextual factor in explaining student learning.

General Pedagogical Approaches

The team included one qualitative intervention article and four qualitative nonintervention articles that discussed general pedagogical approaches (i.e., approaches which were not specific reading approaches). Most of the approaches across articles centered on context and environment—that is, how students interact and are involved in the construction of their own learning. For the most part, the articles presented strong methodologies that link their conclusions to the data. Therefore, much of the information in the pedagogical articles could be reliably adapted to fit other contexts based on need.

Qualitative Intervention. The qualitative intervention article on collaborative learning approaches in Colombia received high ratings on most quality criteria. González et al. (2013) examined how the use of collaborative work in the classroom can aid in the development of students' writing skills.

The study observed students using three collaborative learning strategies that could be adapted to other contexts. The first activity entailed students outlining the task, preparing individually assigned parts, and then coming together to revise the whole document with other students. Teachers observed that students allowed group-level decisions to prevail over their own interests. In the second activity, students played specific roles in the writing process based on their abilities (i.e., writer, idea proposer, leader, compiler, editor). The authors noted that students comprehended "the relevance and importance of their contributions to the initial task," which enabled students to rely on their peers to support their roles (p. 23). In the third strategy, students worked together on the entire development of the document, which allowed the interactions to be more natural and also allowed students to freely use language to communicate ideas.

The authors conclude that collaborative learning approaches are "an opportunity for students to help each other to construct meaning and knowledge, as they work on tasks that demand analyzing, planning, acting, and reflecting on their work as a tool to measure their capacity to work with others" (p. 24). Specific teacher training materials or more specific information on how to implement these strategies and encourage collaborative work in the development of reading and writing skills would be a useful supplement. Teacher trainers should consider looking into how collaborative work could enhance reading and writing abilities in their contexts as students can support each other in the learning process. Researchers could implement quantitatively oriented studies to understand how this strategy might be effective in other contexts (such as poorer schools).

Qualitative Nonintervention. One qualitative nonintervention article that discussed pedagogical approaches encouraged reflection and questioning among children about their educational experiences in Brazil (da Cunha Lima Rosado & Costa Holanda, 2011). The authors argued that considering children's input on the learning process is essential because it contextualizes their reading experience. Understanding children's perspective on learning allows educators to control for those factors that can impact children's perspectives of themselves as learners, impact learning performance, and impact the motivation to learn. Awareness of the importance of young children's views is an issue that can be included in the teacher training process. NGOs can work in this area as well by developing projects that support the social-emotional aspect of learning, particularly motivation.

A second article (Gómez Nashiki, 2008) also argued for incorporating into the classroom aspects of student experience that students consider important. The article focused on strategies to increase the reading level among Mexican students in study areas by conducting a survey of youth about their reading preferences. The author lays out specific recommendations that came out of the survey—as well as a series of proposals from teachers—including: having students create a personal dictionary; having students make their own book; and to establish a reading club. This methodology is similar to others that advocate involving children in the design of classroom activities to contextualize their experiences. This research could be particularly useful for teacher strategies or as a model for teachers in other contexts to conduct their own student surveys and choose teaching practices based on the results.

The third article (Medina & Costa, 2013), about a study in Puerto Rico, discussed context through looking at "children's curricular engagement with the Spanish television genre of telenovelas in relation to classroom critical literacy and performative inquiry." Keeping with the theme of involving children in learning, this study was student led and negotiated. The authors argued that such a lens is important because processes are increasingly becoming globalized, and therefore it is critical to understand how these global processes are being embodied at the local level. Through methodologies such as observation and artefact collection, the authors found that "the idea of reading, writing, and producing across communities could also serve as a powerful lens for engaging in the creation of expansive classroom critical literacy pedagogies" (p. 187). However, the analytical frameworks the authors used in their write-up do not necessarily lend themselves to practical application, especially in contexts where the telenovela is not necessarily prevalent. Nonetheless, the context of globalization and "new ways of reading, interpreting, and producing as children navigate across local global spaces" speaks to the importance of context discussed in the other articles.

The fourth article (Ribeiro & Souza, 2012) discussed the importance of considering context in learning to read and other literacy practices in Brazil. This article was similar to the intervention articles that discussed importance of context in learning pedagogy. The study aimed to understand the impact of certain types of written material on children and found that children recognized maps, medicine labels, newspapers, storybooks, traffic signs, and comic strips with the greatest frequency, indicating that this type of written material speaks to experiences in their lives. However, the authors did not address the practical use of such strategies in the classroom.

This research provides insight into the genres of literature that children commonly recognize. The researchers recommend "considering the processes of literacy in both the pedagogical strategies

of early childhood education, and in speech therapy with students who have difficulties and/or disturbances in the acquisition of writing" (translated from Portuguese). This research also provides insight into the forms of the written word that children commonly recognize and how context impacts learning. In addition, the research argues that reading materials and pedagogies should include lived experiences, as children already come to school with a rich knowledge base that can be used to motivate interest in learning to read. This data could contribute to the development of reading materials that target the contexts in which children focus on the written word in their daily lives and to expand on such genres for pedagogy.

Parental and Community Participation

The team included two qualitative intervention articles and three qualitative nonintervention articles on parental and community participation that met the basic inclusion criteria. The articles argue that home and community contexts should be considered in children's literacy experiences.

Qualitative Intervention. Stein and Rosemberg (2012) discuss how living with extended families in Argentina may contribute to children's literacy development. Particularly, the authors argue that, "it is important to interweave early educational interventions with the funds of knowledge and interactional patterns that characterize children's culture." In this case, the culture meant that "the literacy situations took place within the framework of the interaction between the child and the diverse and multiple participants that comprise the collaboration networks where children and adults assume different roles." This theme of considering the importance of a child's context in his or her literacy experience was evidenced throughout articles across all categories in the review.

Qualitative Nonintervention. The three qualitative nonintervention studies also center on the idea that context and social experience drive a student's literacy experience. Kinkead-Clark (2014) studied immigrant kindergarten children in Jamaica using interviews, artefacts, and school and family observations and found that, "literacy serves a unique purpose to the family unit. Their experiences with literacies reflect their cultural identities and the value they place on its role in as an agent of change." Although this study heavily advocates for considering context when forming a student's classroom experience, the authors do not present specific strategies that could potentially be extrapolated.

Volk and de Acosta (2001 & 2003) conducted three ethnographic case studies of children in mainland Puerto Rico to understand "syncretism," or how students draw from the various contexts in which they interact to construct literacy events. The studies addressed communication within particular social cultural contexts, which is important for sensitizing education stakeholders on how dominant instructional narratives practices can drown out the phenomenology of children's experiences in the learning process—that is, the experiences that they bring to school and the content, form, and meaning of their communications.

Their findings indicate that the three children in Puerto Rico were able to reconstruct literacy lessons using stories, texts, and other tools from their own contexts—a finding supported by other literature. The study does not describe some essential elements of the research, such as the justification for the methodology or a discussion of the evidence against the researchers'

interpretations. In addition, the case study methodology does not allow for extrapolation of findings to other contexts (which the authors address).

The authors indicate that the study contains lessons for sensitizing preservice teachers to different cultures about which they are unfamiliar in teacher education, including through observations of "students literacy learning in homes and communities" (p. 40) and a discussion on how school literacies are often privileged while others are dismissed—including in teachers' own biases. Finally, the authors also recommend that teachers learn how to "co-construct syncretic literacy with children" (p. 40) and how to add to school-centered approaches by consulting families to help construct specific goals for their children appropriate to their skill levels and context.

Reading in Bilingual/Multilingual Contexts

Three qualitative articles focused on learning to read in bilingual/multilingual contexts. One qualitative intervention article (Neugebauer & Currie-Rubin, 2009) focused on using the read-aloud technique to develop Spanish vocabulary and comprehension skills in native Quechua speakers in Peru. Two qualitative nonintervention articles from Colombia discussed reading in bilingual/multilingual contexts, both with a focus on learning English in a dominantly Spanish speaking context. One of the two studies was not of sufficient quality to support the findings. The remaining study by Guevara and Ordoñez (2012) was of high quality and the findings are included below.

Study 1. Neugebauer and Currie-Rubin (2009) conducted a mixed-methods study with first-grade indigenous Quechua speakers in Calca, Peru. There were two control and two intervention classes with a total of 26 and 29 students respectively. While control classes continued business as usual, researchers trained intervention teachers in seven specific read-aloud techniques. Both groups of teachers were given a set of three books on which they were asked to focus their teaching during the normal 30-minute class period five times a week for 3 weeks. Students in the experimental group scored 30 more correct items on the vocabulary assessment than their peers in the control group after only 1 month of the intervention. These data seem to support the effectiveness of read-alouds and the specific read-aloud techniques for promoting vocabulary acquisition in second language learners—particularly those who come from an oral culture. However, as indicated in the section on quantitative intervention research, this study suffers from a high risk of selection-bias because of the small sample size. Thus, we should be cautious in interpreting these results.

Much of the research on the importance of read-alouds thus far has focused on learners of English as a second language in the United States. This research emphasizes providing definitions and contextual information about vocabulary and "actively involving students in word learning through talking about, comparing, analyzing and using the target words" (August, Carlo, Dressler, & Snow, 2005, p. 54). The study of Neugebauer and Currie-Rubin (2009) appears to be one of the few of its kind focusing on the topic of read-alouds for second language learners in the LAC context. In addition, the researchers argue that read-alouds are particularly effective as a pedagogical strategy for indigenous learners who come from a culture where oral traditions are strong as read-alouds combine oral discussion with written narratives.

Study 2. Guevara and Ordoñez (2012) conducted a qualitative study designed to evaluate a newly developed kindergarten curriculum focused on incorporating authentic communication experiences in order to improve language learning in a bilingual education program in Colombia.

The new curriculum focused on building connections between students' first language (Spanish) and English, finding authentic ways for students to practice oral English, as well as promoting interaction and cooperation between students. In order to determine the perception of teachers about the relationship between the curriculum and children's attitudes toward English and learning of English, researchers analyzed four teacher interviews and four classroom observations over the period of a year in addition to two classroom recordings done by the teachers. Researchers found that children:

- Developed positive attitudes towards the foreign language class
- Showed increased motivation and interest to use the foreign language (English)
- Participated more in class

Teachers reported that students showed great improvements in oral vocabulary because of the focus on expressive vocabulary through authentic performances as opposed to the previous focus on written language and receptive skills. One teacher noted,

> *[Before] there used to be a lot of desk work… Everything was focused on listening. You told the children something and they understood, but they couldn't express anything… So you thought, 'If they can listen to me and understand, why can´t they talk? p. 20*

The data showed that students produced a lot of language orally and learned to communicate in different daily situations using accurate structures and vocabulary. Although this study is not generalizable due to the small sample size and is not able to make causal claims, it presents a compelling case for incorporating authentic ways for children to practice foreign language skills (particularly in contexts where there is not a lot of exposure to the language outside of the academic context). Teachers most commonly incorporated games, role plays, songs, and stories and engaged the children in selecting topics and ideas that would be most relevant to them which in turn led to improved student attitudes and increased motivation and participation. The focus on oral English enabled students to "advance in their Spanish literacy process before a different reading system was introduced" (Guevara & Ordonez, 2012, pp. 16–17). This finding is supported by current research on learning multiple literacies in multilingual contexts indicating that making use of students' knowledge of their first language is key to developing literacy in a second or additional language (Cummins, 1979; Koda, 2008; Verhoeven, 1994). However, we need to be careful when interpreting this finding because we only found one study that supports this finding in our review.

Both of the high-quality articles on reading in multilingual contexts share some common themes. Both articles

- Recognize the importance of using and building on the first language in the development of literacy in the second language;
- Focus on building oral vocabulary in the second language to support reading comprehension; and
- Focus on connecting language learning to real life "authentic" experiences and building on what students know and the context they are familiar with in their daily life.

Reading Skills

Only two of the 26 qualitative articles specifically focused on reading skills. Both were nonintervention research articles. One focused on phonological skills and one focused on comprehension but the findings of the article on phonological skills will not be discussed as the article was not rated as high quality.

Language Comprehension. One study aimed to identify the comprehension difficulties faced by 4- and 5-year-old children from low-income populations during story reading at kindergarten, in Buenos Aires, Argentina (Manrique & Borzone, 2010). Researchers analyzed the teacher-student interactions during 26 story-reading settings in nine different kindergarten classrooms and identified three main types of difficulties children faced when trying to comprehend a text that was read to them: (1) illustration-level difficulties, (2) text-level difficulties, and (3) teacher-student interactions. Illustration-level difficulties often occurred when there was a disconnect between the pictures and the text being read, when pictures did not accurately represent the text, or when the pictures contained too much detail and therefore became distracting from the story. Text-level difficulties arose when a text included complex or abstract vocabulary that had not been adequately explained to the children, metaphors, or narrative structure. Teacher-student interactions that led to comprehension difficulties occurred when teachers focused only on explicit aspects of the text such as asking students "what colour was the umbrella" or "what was the boy's name," which caused students to focus on those very specific details as opposed to helping them get a better overall picture of what was happening in the story. In addition, researchers found that when teachers didn't express the emotions elicited by a story, children experienced a disconnect with the text.

The findings from this research indicate that for very young learners, there are specific text and picture factors as well as teacher interaction factors that can affect their comprehension of stories being read aloud to them. Specifically, the findings show the importance of: (a) coherence between the illustrations and text of a story and a need for illustrations that are simple and clearly representative of the text, (b) vocabulary that is understood by the students (or which is clearly explained in the context of the story) and the avoidance of metaphors and narrative structures, and (c) the ability of students to focus on the meaning of the story through more implicit questioning as well as embodying the emotion of the text. However, more research is needed on this theme because we only found one study that focuses on specific text and picture factors and the relationship between teachers and students.

Teaching Practices for Reading

There were three qualitative nonintervention articles that reported on teaching practices for reading.

Study 1 (Webster, 2009). In this study, the researcher worked with a single teacher and her class of 30 Grade 1 students in a rural primary school in Jamaica to determine the relationship between teacher read-alouds of informational texts and students' science learning (as revealed through vocabulary).

The study found that first graders used their own realities to make connections with informational text—that is, they draw on their background knowledge and experience to enhance their understanding of the text. A second finding is that directed look-backs—where the students and teacher go back through the pages of the story to find information—can enable students to gather important facts about the topic of the book and to internalize this technique as a useful literacy strategy. Finally, teacher read-alouds are associated with student content knowledge and expand student vocabulary about the story topic. The results of this research suggest that before, during, and post-reading activities led by the teacher may contribute to the success of read-alouds in developing students' vocabulary and comprehension skills. However, the study design does not allow for making causal claims about the impact of read-alouds.

Study 2 (Jimenez, Smith & Martinez-Leon, 2003). This study examined the language and literacy practices in two Mexican schools over a period of approximately 6 months in two preschool and two Grade 4 classrooms. Researchers conducted 34 classroom observations, interviews with teachers and school principals, and document analysis. In addition to identifying the literacy practices used by students and teachers, researchers sought to determine the ways in which spoken language, reading, and writing were viewed and regulated.

Researchers found that students were given considerable freedom in terms of their spoken language as evidenced by the high noise level in the classrooms and students interjecting while the teacher was talking and asking questions and talking openly with their classmates without any censure from the teacher. This freedom of oral expression contrasts with the emphasis on correct form in students' written work as evidenced by the focus on proper spelling, good handwriting, and general neatness. Reading seemed to fall in the middle depending on whether students were reading silently or aloud. When students read aloud, they were subjected to much more control by teachers as to their pronunciation and inflection and it was clear that their oral reading was expected to be fluent and flawless. However, when students were allowed time to read as they pleased, this was completely unregulated by teachers, and students could be seen reading silently, reading in groups, and informally discussing the text and illustrations.

It is difficult to extrapolate the findings of this study as the purpose was primarily to identify existing literacy practices in a specific location. Studying the regulation of different literacy practices by teachers could be a necessary first step in implementing changes to teaching practices in order to determine how literacy is currently taught as well as whether the emphasis is on different aspects of the literacy process.

Study 3 (Diuk, 2007). The aim of this study was to analyze the reading and spelling acquisition process of two first grade girls in Buenos Aires, Argentina. Reading tests were given to both girls at the beginning of the year focusing on skills such as the recognition of rhymes, initial sounds of words, letter knowledge, and the reading and writing of words. Researchers administered another reading test at the end of the first year (35 weeks of class) to see what changes had occurred in the girls' literacy skills. The girls were asked to self-report on strategies they used during the reading and writing of words. As in previous studies, this study found that the girls both relied on logographic strategies in the initial stages of literacy learning but slowly developed more analytical strategies. The authors suggested that poor reading levels of children in marginalized contexts may be the consequence of not providing them with adequate instructions on metaphonological strategies and explicit and systematic phonics. However, with a sample size of only two children,

this study cannot credibly make these claims but only suggest this as a possible avenue for future research.

Teacher Training

We included one qualitative intervention article and one qualitative nonintervention article that related to teacher training. The qualitative intervention article (Caldera de Briceño, Escalante de Urrecheaga & Terán de Serrentino, 2010) did not provide strong evidence for its conclusions. The nonintervention article (Warrican et al., 2008) discussed challenges exemplary teachers in the Caribbean faced in promoting literacy among students using a model shown to be effective in promoting literacy in students. Although the article does not provide an in-depth description of the program elements, the authors state that teachers receive training in a wide variety of teaching methods that contribute to their understanding of literacy development (e.g., phonological awareness, word recognition, and fluency) as well as differentiated instruction, student-centered activities, and the use of action research.

> *The mentoring, training and the collaboration that is fostered through working together on problems and finding solutions, result in a validation of the teachers that leaves them feeling cared for and special. Despite the often difficult circumstances under which they find themselves, these teachers are thus unlikely to experience the isolation that others in equally challenging situations experiences (p. 28).*

More generally, the training may have allowed the teachers "to acquire knowledge and skills that brought about noticeable changes in some classrooms;" however, more explicit linkages from specific project elements to specific outcomes would help to determine which elements are a priority and why. As with the articles on parental and community participation and reading materials, the teacher training article advocates encouraging teachers to create a highly contextual literacy environment for students.

Discussion

The purpose of this systematic review is ultimately to understand the existing evidence about successful reading-related programs, practices, and policies in the LAC region. The evidence can then be used to inform the ongoing work of practitioners, support evidence-based policy decisions, and provide direction for further research priorities. To triangulate the findings from the meta-analysis and synthesis of all research types, the team conducted a systematic qualitative evaluation review. Patton (2015a) defines this process as "selecting diverse evaluation studies already completed on a program or policy and synthesizing findings across those separate and diverse evaluations to reach conclusions about what is effective" (p. 270).

In this section, we seek to tie together the key findings from each of the four research types to present a coherent picture of the early grade reading evidence emerging from the LAC region and identify gaps in that evidence. LAC is a large region composed of more than 40 countries with multiple languages, diverse populations, and unique education systems. Thus, the findings we present here are not necessarily generalizable to the entire LAC region. Although our synthesis presents results for a wide variety of contexts, some countries are overrepresented, as shown in the

evidence-gap maps. Specifically, our results may be more externally valid for high-income and upper-middle-income economies because these contexts are overrepresented in our synthesis. We will take these findings into consideration in the interpretation and will examine differences in enabling factors across contexts.

Overall Synthesis

The purpose of this exercise is to understand what information exists about successful and unsuccessful strategies to improve early grade reading outcomes in the LAC region and how these strategies differ across contexts and why. The information can inform the ongoing work of policy makers and practitioners as well as priorities for further research. To achieve this goal, we triangulated the findings of the four different research types. To synthesize the findings from the meta-analysis and narrative synthesis, the team conducted a systematic qualitative evaluation review.

Overall, we only found a small number of studies that can make credible causal claims about the impact of development programs and strategies inside and outside the classroom on early grade reading outcomes. The majority of quantitative intervention studies suffer from either a medium or high risk of selection bias or a medium or high risk of performance bias, meaning the evidence is derived from methodologically weak studies. Furthermore, we found strong evidence for publication bias in the studies that focus on the effects of teacher practices and parental involvement on early grade reading outcomes in the LAC region; that is, there are likely to be a large number of additional studies that have not been published on similar topics because they did not find statistically significant effects. Findings about non-effective interventions are also important, and publishing only the results of programs that show positive and statistically significant effects on early grade reading outcomes impedes policy makers' ability to make evidence-informed decisions. Although the evidence is less clear, we also found some indications for publication bias in the other studies that we included in our meta-analysis. Thus, our synthesis suggests that the current evidence base on the impact of development programs on early grade reading outcomes is rather small.

Importantly, it is not unusual to only find a small evidence-base on the impact of development programs. Although the number of impact evaluations of development programs has increased dramatically in the last couple of years (Cameron, Mishra & Brown, 2015), the impact evaluation field in international development is still relatively young. Nonetheless, it is important to realize that at this moment there is only limited rigorous evidence on what works to improve early grading reading outcomes. However, each of our study types presents some inconclusive but promising evidence on the types of programs that may be effective in improving early grade reading outcomes.

Evidence on Improving Early Grade Reading Outcomes

Quantitative intervention studies present some examples of development programs that are likely to have positive effects on early grade reading outcomes in specific circumstances and contexts. Specifically, we found evidence that teacher training programs can positively affect early grade reading outcomes in high-income economies when they are well implemented and complemented by the sustained coaching of teachers. In addition, we found some evidence that nutrition programs

have positive effects on early grade reading outcomes in contexts where stunting and wasting are high, such as Guatemala. However, we also found evidence indicating that the distribution of laptops to children had adverse effects on early grade reading outcomes, particularly when the distribution of laptops is not complemented by additional programs.

For the effects of preschools, school governance, specific teacher practices, and parental involvement, we only found quantitative intervention evidence with a medium or high risk of bias. These programs could potentially positively affect early grade reading outcomes. However, the quantitative evidence for the effectiveness of these programs has not (yet) been rigorously established.

The findings for quantitative nonintervention studies complement the quantitative intervention studies by presenting directions on how programs can be improved to stimulate early grade reading programs. First, the findings indicate that phonemic awareness, phonics, fluency, and comprehension are associated with reading ability, which suggests that teacher training programs need to focus on these aspects to increase their effectiveness, albeit in a way that reflects that languages and scripts used in the LAC region. Furthermore, the research indicates that poverty and child labor are negatively correlated with early grade reading outcomes. This finding on the importance of poverty and socio-economic factors for early grade reading outcomes supports the quantitative intervention result that nutrition programs may be effective in improving early grade reading outcomes in low-income or lower-middle-income contexts with high stunting and wasting, such as Guatemala. Finally, the quantitative non-intervention studies show that the quality of preschool and the development of very early emergent literacy skills is positively associated with early grade reading outcomes. Triangulating this result with the quantitative findings on the impact of teacher training suggests that teacher training could possibly positively affect early grade reading outcomes through its influence on the quality of preschool. However, it remains important to combine these programs with sustained coaching of teachers to ensure sustainable implementation of the recommendations given in the teacher training.

Context and Experience Affect Student Learning.

Both qualitative and quantitative studies indicated that social learning and consideration of context is key to improving reading outcomes. This lends credence to the conceptual framework, which suggests that enabling factors and assumptions in part determine the potential for success of various programs or strategies. In addition, continuing to consider the context during program implementation in turn affects educational outcomes. The articles in our syntheses specifically discuss context and experience in terms of child labor, poverty and nutrition, importance of family, and considering student input in the learning process.

Poverty, Child Labor and Poor Nutrition Are Associated With Poor Early Grade Reading Outcomes.

Some of the included quantitative studies demonstrate the importance of accounting for socioeconomic enabling factors in the design of education programs and teaching strategies. There is some evidence that programs that aim to improve nutrition outcomes can positively influence early grade reading outcomes in contexts with high rates of malnutrition, such as Guatemala. In addition, seven quantitative nonintervention studies indicated that poverty and child labor are

associated with poor student reading outcomes. However, the quantitative intervention studies also show that nutrition and poverty may be less crucial in determining early grade reading outcomes in upper-middle-income or high-income contexts. Nonetheless, the findings of these studies may well underestimate the impact of nutrition programs because of risk of spillovers. Together, these findings indicate that socioeconomic factors are important predictors of early grade reading outcomes. However, programs that stimulate nutrition only appear to be effective in stimulating early grade reading outcomes in contexts where levels of stunting and wasting are high.

This quantitative finding is consistent with the qualitative evidence that suggests that education programs need to be tailored to the local contexts to maximize the effectiveness of early grade reading programs. The evidence indicates that experiential learning or considering children's inputs in the learning process may contribute to the tailoring of education programs to the local context. In addition, extended families and social networks can also contribute to stimulating early grade reading outcomes.

Collaborative Learning

The most frequently discussed topic in qualitative nonintervention articles is the use of social learning to improve early grade reading. For example, Gonzalez (2013) shows that collaborative learning approaches provide an opportunity for students to help each other to construct meaning and knowledge. Furthermore, Castanheria et al. (2013) argue that children's engagement in literacy events shows that there is both a collective and individual effort for students to position themselves as readers and writers in the classroom environment. Manrique and Borzone (2010) also discuss the potential for "interaction in a process of shared cognition" to help children from marginalized urban sectors to understand a text. However, it is unclear whether social learning has a positive effect on early grade reading outcomes. Deriving such conclusions requires more rigorous mixed-methods research with a focus on addressing counterfactual questions.

The use of ICT may contribute to social learning if it is used for computer-aided instruction, but our evidence also indicates that the distribution of laptops may have adverse effects if this effort is not complemented with additional interventions or programs. It is possible that computer-aided instruction contributes to social learning, while the individualized nature of learning through using laptops may have contributed to the adverse effects. However, more rigorous mixed-methods research is needed to assess whether ICT programs are indeed associated with reductions in social learning.

Teacher Training and Teacher Practices

The evidence shows that teacher training programs are likely to positively influence early grade reading outcomes in high-income economies such as Chile when they are well implemented and combined with coaching of teachers, but there is little evidence on what teacher practices are required to improve early grade reading outcomes. Nonetheless, Porras Gonzalez (2010) explore how playing games can possibly contribute to teacher practices and motivate children to learn. Furthermore, Warrican et al. (2008) show that exemplary teachers possess a caring attitude toward their students that contributes to teachers' promotion of literacy and can potentially improve student performance. These articles suggest that shifting teachers' practices and school ideologies

can potentially contribute to improving education systems. However, more rigorous mixed-methods research is needed to determine the causal mechanisms underlying these relationships.

Concrete Literacy Strategies

Across the four types of research, the programs and implementation techniques that aim to impact early grade reading focus on developing phonological awareness and using read-alouds. Both qualitative and quantitative intervention research focused on read-aloud interventions. In Jamaica, findings showed that read-alouds with informational texts can help children make connections with their own realities and increase their content knowledge and expand their vocabulary (Webster, 2009). Read-alouds were also used successfully in bilingual settings to support vocabulary acquisition in the second language (Neugebauer & Currie-Rubin, 2009). However, the quantitative intervention research indicates that studies with an emphasis on read-alouds have a high risk of selection bias. Furthermore, the risk of publication bias is substantial. It is likely that research that doesn't show statistically significant effects for read-aloud practices is not published. Thus, we may have an incomplete picture of the influence of read-aloud strategies on early grade reading outcomes. Again, more rigorous mixed-methods research is needed to determine the effects of read-alouds on early grade reading outcomes.

Nine of the 22 quantitative non-intervention reading skills studies found a strong connection between phonological awareness and reading ability suggesting the need to teach PA skills early on. Studies focused on the importance of phonological awareness and phonics to help students become strong decoders and one study also suggested that phonemic awareness is not as necessary when learning a transparent orthography such as Spanish (Goldenberg, 2014) although more research is needed for this finding to be conclusive. One quantitative intervention study also presents evidence that training children in phonological awareness can improve the learning of letter sounds (Cardoso Martins et al., (2011), but this study suffers from a high risk of selection-bias.

There was a clear lack of studies focusing on reading comprehension. This is interesting given the fact that comprehension is the ultimate goal of reading and is something that students in the LAC region struggle to master as evidenced by scores on national reading assessments. One qualitative article focused on comprehension in very young learners and indicated that specific text and picture factors as well as teacher interaction factors affect student comprehension of stories being read aloud to them. Only three of the quantitative nonintervention studies centered on comprehension and its relationship to fluency but most studies only discussed comprehension at the word level. The quantitative intervention research on comprehension was also quite sparse. Vivas (1996), indicated that listening to stories read aloud by parents results in improvements in language comprehension and Murad & Topping (2000) found positive effects of paired reading with parents on children's reading comprehension and fluency. However, these studies could be biased due to a high risk of selection-bias.

Quantitative nonintervention studies and qualitative intervention studies also provide evidence for a positive association between teaching phonemics, fluency, and reading comprehension. However, it is unclear whether the relationship is causal. Quantitative intervention studies do not present rigorous evidence for the positive effects of these trainings on early grade reading comprehension. Nonetheless, the quantitative nonintervention and intervention research suggests

some interesting hypotheses on what types of programs may be effective in improving reading comprehension, which could be tested in future rigorous mixed-methods research.

Impacts on Early Grade Reading Outcomes

Overall, we only found a small number of studies that can make credible claims about the impact of development programs on early grade reading outcomes. The majority of the studies suffer from either a medium or high risk of selection bias or a medium or high risk of performance bias. Furthermore, we found strong evidence for publication bias in the studies that focus on the effects of teacher practices and parental involvement on early grade reading outcomes in the LAC region. These findings suggest that policy makers and other key stakeholders currently do not have access to sufficient rigorous evidence for informing their policy decisions.

In general, the evidence base on what works to improve early grade reading outcomes in the LAC region is weak. We only found a small number of studies that are able to present credible estimates on the impact of development programs on early grade reading outcomes.

Strengths of the Review

This systematic review is strong because of its broad search and analysis approach, its use of novel computational techniques grounded in computer science, and its focus on high-quality research. The review includes articles in multiple languages, across more than 20 countries, and on a variety of topics within early grade reading. We included not only experimental or quasi-experimental quantitative intervention studies, but also quantitative nonintervention studies, qualitative intervention studies, and qualitative nonintervention studies. Typically, systematic reviews include only experimental and quasi-experimental quantitative studies; however, we adapted internationally recognized quality rating protocols to review nonexperimental studies so that this valuable body of evidence could be included. Finally, we used machine learning, a technique that enabled us to increase the comprehensiveness of our review. The team's use of these methods not only enhanced the breadth and accuracy of the review but may also enhance its usability for end users who can now use experimental, quasi-experimental, qualitative, and quantitative nonintervention research to inform their policy decisions.

This review also uses risk of bias assessments for different research types to determine the validity and reliability of the research on early grade reading in the LAC region. The inclusion of risk of bias assessments is an important strength of this review. It allows donors and policy makers to determine the quality of early grade reading research. Currently, the ability of policy makers to implement evidence-based policy is compromised by the difficulties they experience in determining the quality of research. The use of risk of bias assessments enables us to assess the potential biases in the included research, which can help policy makers in determining which research findings to use and which ones to ignore.

Limitations of the Review

All studies are conceived within a specific budget, timeline, and the inherent challenges that arise by conducting the study within those conditions. Therefore, it is important to acknowledge those constraints when analyzing results. In this section, we discuss the limitations of the systematic

review and the implications of such limitations when interpreting the results. We also discuss the strategies we implemented to minimize the potential effects of such limitations.

Number of Reviewers

The number of reviewers on the systematic review team increased the possibility that reviewers' subjectivities influenced rating techniques and reduced consistency. To limit subjectivity in the review of quantitative intervention research, each article was reviewed by a minimum of two quantitative research experts. We then reached consensus about the risk of bias through discussion. To limit subjectivity among the reviewers of the other research types, reviewers initially co-reviewed articles to cross-check for comparable understandings.

Large-Scale Research Question

This review is both limited and strengthened by the broad scale of the research question that guides the study. The review aims to capture every piece of research in the LAC region on early grade reading. Although the large scale of this research question made it difficult to search for and summarize all of the existing literature, it also enabled us to investigate larger questions within early grade reading. The large scope increases the relevance of the review for policy makers.

Furthermore, we mitigated the concern about the scope of the review by relying on an overarching conceptual framework that covers each of the research questions in our review. Interpreting our findings in the light of this framework enabled us to reliably assess each research question.

Inclusion of Articles

In contrast to a traditional systematic review that includes only experimental and quasi-experimental quantitative research, we included all types of quantitative research as well as qualitative studies. As such, it is important to note that the included quantitative nonintervention and qualitative studies do not present causal evidence on what works to improve early grade reading outcomes. However, these studies present some interesting hypotheses on how programs may need to be implemented to improve early grade reading.

We also created our own protocols for assessing the quality of qualitative and quantitative nonintervention articles based off of multiple examples. Finally, this review includes only published articles or grey literature that is searchable through official search databases online; we did not include research that is not available through the databases we list above.

Lack of Specific Information for Early Grade Readers

Many of the studies do not differentiate between programs that had an effect on early grade reading outcomes versus programs that had an effect on reading outcomes for other grades. As a result, we are not always able to make this distinction. Thus, we have to assume that the effects are homogeneous when interpreting our findings. We mitigated this concern by contacting authors and requesting information about the effects of programs on reading outcomes for early grade students. However, some authors did not respond and others were not able to provide differential effects by age or grade.

Recommendations

The primary end goal of all activities within the LAC Reads Capacity Program is to enhance the capacity of key stakeholders (e.g., the Ministry of Education and the government, international funders and intergovernmental entities, international NGOs, academics, and researchers and practitioners) to use evidence to choose, develop, implement, and evaluate early grade reading (EGR) strategies, programs, practices, and interventions. We developed several recommendations on the basis of the review of the EGR evidence from the LAC region and our analysis of the evidence-gaps. Some recommendations stem directly from the evidence on a particular topic, while other recommendations stem directly from gaps in the evidence and identify research gaps in need of addressing.

We also took into consideration the strength of the evidence in our recommendations by distinguishing between internal and external validity. Internal validity refers to the validity of inferences about whether the correlation between access to a certain program and the outcome variable can be considered a causal relationship (Shadish, Cook, & Campbell, 2002). To identify the internal validity of the findings, we rely on our risk of bias assessments. External validity refers to the generalizability of the evaluation's findings to different contexts (Shadish et al., 2002). To examine external validity, we assess whether the findings are specific to geographic contexts with certain contextual characteristics. We classify the recommendations on the basis of the internal and the external validity of the findings by assessing whether the findings identify a credible causal relationship and by examining the contextual characteristics to which the results can credibly be extrapolated. Recommendations about research gaps are in some cases also directly related to the relatively low internal validity of the research findings. Below we present our recommendations including a classification of the recommendations on the basis of the strength of the evidence.

Recommendations for the MOE/Government, International Funders, Intergovernmental Entities and International NGOs:

- Focus more resources on enhancing preschool quality specifically through training high quality teachers in higher middle-income and high-income countries. We did not encounter studies that credibly assess the impact of enhancing preschool quality in lower middle-income or low-income countries.
- Invest in nutrition programs in contexts with high rates of early childhood stunting and wasting to improve early grade reading outcomes. The evidence regarding the effects of nutrition programs on early grade reading outcomes is less clear in contexts with low rates of early childhood stunting and wasting.

Recommendations for Practitioners:

- Focus pedagogical approaches on the various predictors of reading skills, such as phonemic awareness, the alphabetic principle, decoding (learning the sound-symbol correspondences), vocabulary, and comprehension, which will likely contribute to reading improvements.[15]

[15] This recommendation is based on correlational and not causal research.

- Make reading activities more interesting and contextually relevant by incorporating students' ideas about potential activities and reading materials into lessons.

Recommendations Based on the Evidence Gaps:

- Ensure that language assessments include multiple reading constructs and differentiate between those constructs so it is easier to identify the effects of interventions on individual constructs.
- Fund long-term mixed-methods experimental or quasi-experimental research on the effects preschool and early childhood education on early grade reading outcomes.
- Include several early grade reading constructs in administrative data to enable researchers to conduct high-quality research on the mechanisms underlying early grade reading using large sample sizes.
- Document ongoing research to minimize publication bias so that unpublished research is known to policy makers as well and to ensure that hypotheses are pre-specified.
- Register ongoing research on early grade reading in a central, publicly available location so that everyone can see what is being done and seek to complement and add to the research base.
- Develop more interdisciplinary mixed-methods research on early grade reading that includes more than one reading construct and large sample sizes.
- Fund rigorous research that allows for an examination of the causal effects of development programs on early grade reading outcomes. These studies include both experimental and quasi-experimental studies with a sufficient sample size. These studies also need to be supplemented with qualitative research.
- Pursue more research on EGR strategies for students with disabilities.
- Pursue more research on reading in indigenous languages.
- Conduct more research on the linkages between the development of prewriting and writing skills and early grade reading outcomes.

In addition to reviewing the EGR evidence from the LAC region, the LAC Reads Capacity Program also collects and catalogues EGR pedagogical resources (e.g., supplementary reading materials, assessments, instructional materials, videos), and other EGR documents that are neither research-based evidence nor resources (e.g., policy documents, project reports, best practices documents) from the LAC region. These resources can serve as additional support for stakeholders to improve their practice. We plan to use the results of this systematic review to help determine the effectiveness and relevance of EGR resources available in each country with the goal of improving and sustaining early grade literacy performance.

We are also conducting a stakeholder analysis in the region with our local partner organizations to identify the key EGR stakeholders, determine their interests and needs, and how the evidence from this review and the resources collected can best be used to support EGR capacity and achievements

in the region. The program website (www.lacreads.org) will feature an EGR evidence and resource database, as well as publications and key information related to EGR content as well as current and future research carried out by the LRCP.

Over the remaining years of the project, the systematic review team will continue to search, evaluate and include new evidence on early grade reading in updated versions of the systematic review report. For example, USAID is currently funding the LAC Reads/Evaluation Program, which is a five-year project that rigorously evaluates and costs USAID investments in early literacy and access to education in conflict settings in the LAC region. As part of this project, Mathematica Policy Research is currently leading multi-year studies in four countries (Guatemala, Peru, Honduras, and Nicaragua) using randomized controlled trial designs supplemented with qualitative studies.

The first study is an evaluation of an early literacy intervention for bilingual populations in Guatemala and Peru, *Leer Juntos, Aprender Juntos*, implemented by Save the Children. The second study is an evaluation of the use of end-of-grade formative assessments and their impact on student learning in Honduras, implemented by the American Institutes for Research. The third study is an evaluation of a community partnership reading intervention in Nicaragua, *Espacios para Crecer* (EpC, or "Spaces to Grow"), implemented by the Community Action for Reading and Security (CARS) Project team. The fourth study is an evaluation of targeted teacher training and supports for pedagogical approaches to teaching early-grade reading, *Amazonia Lee* ("Amazon Reads"), in two Amazonian provinces of Peru. We anticipate that evidence will be emerging from each of these and other programs in the coming months and years which will continue to build the evidence base on EGR in the LAC region, and we will incorporate and share this evidence as it becomes available

Appendices

Appendix A. Citations

Abadzi, H., Crouch, L., Echegaray, M., Pasco, C., & Sampe, J. (2005). Monitoring basic skills acquisition through rapid learning assessments: A case study from Peru. *Prospects*, *35*(2), 137–156.

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

Adrogue, C., & Orlicki, M. E. (2013). Do in-school feeding programs have an impact on academic performance and dropouts? The case of public schools in Argentina. *Education Policy Analysis Archives*, *21*, 50.

Agosto, G., Citarroni, C., Briasco, I., & Garcette, N. (2012). From impact evaluations to paradigm shift. A case study of the Buenos Aires Ciudadanía Porteña conditional cash transfer programme (Working Paper 17). New Delhi: International Initiative for Impact Evaluation (3ie).

Are Latin American children's reading skills improving? Highlights of the Second and Third Regional Comparative and Explanatory Studies (SERCE & TERCE). (2015). Washington, DC: American Institutes for Research.

Athayde, M. D. L., Giacomoni, C. H., Zanon, C., & Stein, L. M. (2014). Evidências de validade do subteste de leitura do teste de desempenho escolar. *Psicologia: Teoria e prática*, *16*(2), 131–140.

August, D., Carlo, M., Dressler, C., & Snow, C. (2005). The critical role of vocabulary development for English language learners. *Learning Disabilities Research & Practice*, *20*(1), 50–57.

August, D., & Shanahan, T. (Eds.). (2006). *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bandini, H. H. M., Oliveira, C. L. A. C., & Souza, E. C. (2006). Habilidades de leitura de pré-escolares deficientes auditivos: Letramento Emergente. *Paidéia*, *16*(33), 51–58.

Bando, R. (2010). *The effect of school based management on parent behavior and the quality of education in Mexico* (Doctor of Philosophy dissertation). Available from Bancroft Digital Collections, the Bancroft Library, University of California, Berkeley. Retrieved from http://digitalassets.lib.berkeley.edu/etd/ucb/text/Bando_berkeley_0028E_10483.pdf

Barrera, S. D., & Maluf, M. R. (2003). Consciência metalingüística e alfabetização: um estudo com crianças da primeira série do ensino fundamental. *Psicologia: reflexão e crítica*, *16*(3), 491–502.

Barrera-Osorio, F., & Linden, L. L. (2009). *The use and misuse of computers in education: Evidence from a randomized experiment in Colombia* (Policy Research Working Paper 4836). The World Bank Human Development Network Education Team.

Belintane, C. (2010). Orality, literacy and reading: Dealing with differences and complexities in the public school. *Educação e Pesquisa*, *36*(3), 685–703.

Benítez, Y. G., Vargas, G. G., Hernández, A. L., Sánchez, U. D., & García, Á. H. (2007). *Habilidades lingüísticas en niños de estrato sociocultural bajo, al iniciar la primaria*. *Acta Colombiana de Psicología*, *10*(2).

Beuermann, D. W., Cristia, J., Cueto, S., Malamud, O., & Cruz-Aguayo, Y. (2015). One laptop per child at home: Short-term impacts from a randomized experiment in Peru. *American Economic Journal: Applied Economics*, *7*(2), 53–80.

Bhattacharyya, S. (2004). If you build it they will come: Creating a field for project based approach in learning science through qualitative inquiry and technology integration. Proceedings of QualIT2004: The Way Forward, Brisbane, Australia, November 24–26, 2004, Griffith University.

Bizama, M., Gutiérrez, B. A., & Sáez, K. (2011). Evaluación de la conciencia fonológica en párvulos de nivel transición 2 y escolares de primer año básico, pertenecientes a escuelas de sectores vulnerables de la provincia de Concepción, Chile. *Onomázein: Revista de lingüística, filología y traducción de la Pontificia Universidad Católica de Chile*, *23*, 81–103.

Borenstein, M. H., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Publication bias. In: *Introduction to meta-analysis* (pp. 277–294). Chichester, England: Wiley. doi:10, 9780470743386.

Bravo, L., Villalón, M., & Orellana, E. (2002). La conciencia fonológica y la lectura inicial en niños que ingresan a primer año básico. *Psykhé*, *11*(1).

Brody, C., De Hoop, T., Vojtkova, M., Warnock, R., Dunbar, M., Murthy, P., & Dworkin, S. L. (2015). Economic self-help group programs for improving women's empowerment: A systematic review. *The Campbell Collaboration*, 1–184.

Brody, C., De Hoop, T., Vojtkova, M., Warnock, R., Dunbar, M., Murthy, P., & Dworkin, S. L. (2016). Can self-help group programs improve women's empowerment? A systematic review. *Journal of Development Effectiveness*, 1-26.

Bryman, A. (2007). Effective leadership in higher education: A literature review. *Studies in Higher Education*, *32*(6), 693–710.

Caldeira, M. M., & Ward, J. M. (2003). Using resource-based theory to interpret the successful adoption and use of information systems and technology in manufacturing small and medium-sized enterprises. *European Journal of Information Systems*, *12*(2), 127–141.

Caldera de Briceño, R., Escalante de Urrecheaga, D., & Terán de Serrentino, M. (2010). Práctica pedagógica de la lectura y formación docente. *Revista de Pedagogía*, *31*(88).

Cameron, D. B., Mishra, A., & Brown, A. N. (2015). The growth of impact evaluation for international development: how much have we learned? *Journal of Development Effectiveness*, *8*(1), 1–21.

Campos, M. M., Bhering, E. B., Esposito, Y., Gimenes, N., Abuchaim, B., Valle, R., & Unbehaum, S. (2011). The contribution of quality early childhood education and its impacts on the beginning of fundamental education. *Educação e Pesquisa*, *37*(1), 15–33.

Capovilla, A. G. S., Capovilla, F. C., & Suiter, I. (2004). Processamento cognitivo em crianças com e sem dificuldades de leitura. *Psicologia em estudo*, *9*(3), 449–458.

Capovilla, A. G. S., Gutschow, C. R. D., & Capovilla, F. C. (2004). Habilidades cognitivas que predizem competência de leitura e escrita. *Psicologia: Teoria e prática*, *6*(2), 13–26.

Cardoso-Martins, C., & Da Silva, J. R. (2010). Cognitive and language correlates of hyperlexia: Evidence from children with autism spectrum disorders. *Reading and Writing*, *23*(2), 129–145.

Cardoso-Martins, C., & Fulanete Correa, M. (2008). O desenvolvimento da escrita nos anos pré-escolares: Questões acerca do estágio silábico. *Psicologia: Teoria e pesquisa*, *24*(3), 279–286.

Cardoso-Martins, C., Mesquita, T. C. L., & Ehri, L. (2011). Letter names and phonological awareness help children to learn letter–sound relations. *Journal of Experimental Child Psychology*, *109*(1), 25–38.

Castanheira, M. L., Neves, V. F. A., & Gouvêa, M. C. S. D. (2013). Eventos interacionais e eventos de letramento: um exame das condições sociais e semióticas da escrita em uma turma de educação infantil. *Cadernos CEDES*, *33*(89).

Castro, D. C., Lubker, B. B., Bryant, D. M., & Skinner, M. (2002). Oral language and reading abilities of first-grade Peruvian children: Associations with child and family factors. *International Journal of Behavioral Development*, *26*(4), 334–344.

Cervini, R. A. (2015). Trabajo infantil y logro escolar en América Latina-los datos del SERCE. *Revista electrónica de investigación educativa*, *17*(2), 130–146.

Cipielewski, J., & Stanovich, K. E. (1992). Predicting growth in reading ability from children's exposure to print. *Journal of Experimental Child Psychology*, *54*(1), 74–89.

Compton-Lilly, C. (2007). *Re-reading families: The literate lives of urban children, four years later* (Practitioner Inquiry series). Teachers College Press.

Corral Verdugo, V., Bazán Ramirez, A., & Sánchez Hernandez, B. (2000). Validez de constructos funcionales y morfológicos en tareas de lecto-escritura: Un estudio con niños de educación básica. *Acta comportamentalia*, *8*(2), 226–252.

Correa, J., & Dockrell, J. E. (2007). Unconventional word segmentation in Brazilian children's early text production. *Reading and Writing*, *20*(8), 815–831.

Cristia, J., Ibarrarán, P., Cueto, S., Santiago, A., & Severín, E. (2012). *Technology and child development: Evidence from the one laptop per child program* (Discussion Paper No. 6401). Bonn, Germany: Institute for the Study of Labor.

Cueto, S., & Díaz, J. J. (2013). Impacto de la educacion inicial en el rendimiento en primer grado de primaria en escuelas publicas urbanas de Lima. *Revista de Psicología*, *17*(1), 73–91.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, *49*(2), 222–251.

Cunningham, A. E., & Stanovich, K. E. (1991). Tracking the unique effects of print exposure in children: Associations with vocabulary, general knowledge, and spelling. *Journal of Educational Psychology*, *83*(2), 264.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, *49*(2), 222–251.

Denzin, N. K. (1978). *The research act: A theoretical introduction to research methods* (2nd ed.). New York: McGraw-Hill.

De Abreu, M. D., & Cardoso-Martins, C. (1998). Alphabetic access route in beginning reading acquisition in Portuguese: The role of letter-name knowledge. *Reading and Writing*, *10*(2), 85–104.

Diaz, J. J., & Handa, S. (2006). An assessment of propensity score matching as a nonexperimental impact estimator: Evidence from Mexico's PROGRESA program. *The Journal of Human Resources, XLI* (2), 319–345.

Dias, T. L., Enumo, S. R. F., & Turini, F. A. (2006). Avaliação do desempenho acadêmico de alunos do ensino fundamental em Vitória, Espírito Santo. *Estudos de Psicologia (Campinas)*, *23*(4), 381–390.

Diuk, B. (2007). El aprendizaje inicial de la lectura y la escritura de palabras en español: un estudio de caso. *Psykhe (Santiago)*, *16*(1), 27–39.

Duvendack, M., Palmer-Jones, R., Copestake, J. G., Hooper, L., Loke, Y., & Rao, N. (2011). *What is the evidence of the impact of microfinance on the well-being of poor people*? London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Egozi, O., Markovitch, S., & Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)*, *29*(2), 8.

Evans, D., & Popova, A. (2015). What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews (Policy Research Working Paper 7203). World Bank Group.

Favila, A., Yánez, G., Bernal, J., Silva, J., Marosi, E., Rodríguez, M., & Fernández, T. (1999). La conciencia y la memoria fonológicas son factores predictores del nivel de lectura y escritura alcanzado en niños de primer grado de primaria. *Revista Mexicana de Psicologia*, 16(1), 57–63.

Felício, F. D., Terra, R., & Zoghbi, A. C. (2012). The effects of early childhood education on literacy scores using data from a new Brazilian assessment tool. *Estudos Econômicos (São Paulo)*, 42(1), 97–128.

Ferrando, M., Machado, A., Perazzo, I., & Vernengo, A. (2011). Aprendiendo con las XO: El impacto del Plan Ceibal en el aprendizaje. Serie Documentos de Trabajo/FCEA-IE; DT03/11.

Ford, E. D. (2009). The importance of a research data statement and how to develop one. *Annales Zoologici Fennici*, *46*(2), 82–92.

Foy, J. G., & Mann, V. (2003). Home literacy environment and phonological awareness in preschool children: Differential effects for rhyme and phoneme awareness. *Applied Psycholinguistics*, *24*(01), 59–88.

Francis, N. (1999). Applications of cloze procedure to reading assessment in special circumstances of literacy development. *Reading Horizons*, *40*(1), 23.

Fuller, B., Dellagnelo, L., Strath, A., Bastos, E. S. B., Maia, M. H., de Matos, K. S. L., ... & Vieira, S. L. (1999). How to raise children's early literacy? The influence of family, teacher, and classroom in northeast Brazil. *Comparative Education Review*, 1–35.

Gabrilovich, E., & Markovitch, S. (2006, July). *Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge* (pp. 1301–1306). AAAI'06 proceedings of the 21st national conference on artificial intelligence, Volume 2. Menlo Park, CA: AAAI Press.

Giacomoni, C. H., Athayde, M. D. L., Zanon, C., & Stein, L. M. (2015). Teste do Desempenho Escolar: evidências de validade do subteste de escrita. *Psico USF*, *20*(1), 133–140.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–8.

Goldenberg, C., Tolar, T. D., Reese, L., Francis, D. J., Bazán, A. R., & Mejía-Arauz, R. (2014). How important is teaching phonemic awareness to children learning to read in Spanish? *American Educational Research Journal*, *51*(3), 604–633.

Gomez Franco, L. E. (2014). Exploring teachers' read-aloud practices as predictors of children's language skills: The case of low-income Chilean preschool classrooms (Doctoral dissertation, Boston College. Lynch School of Education).

Gómez Nashiki, A. (2008). La práctica docente y el fomento de la lectura en Colima: Estrategias y recomendaciones de los docentes de educación básica. *Revista Mexicana de Investigación Educativa*, *13*(39), 1017–1053.

Gómez-Pérez, M. A., Sierra, M. D. L. D. V., Jiménez, J., & Méndez, A. M. (2011). Evaluación de los procesos cognitivos de la lectura a través del SICOLE-R-Primaria en niños que cursan la educación primaria: un estudio transversal (Evaluate the cognitive processes of reading through a SICOLE-R-Primary in children attending primary school. A transversal study). *Revista de Psicologéa de la Educación*, 6.

Gonçalves Medeiros, J., Antunes, L., Pokrevieschi, J. E. J., Bottenberg, D. G., de Amorim Ferreira, C., & Eickhoff Cavalhieri, K. (2011). Emergência de leitura de frases a partir do ensino de suas unidades constituintes. *Acta Comportamentalia: Revista Latina de Análisis del Comportamiento*, *19*(3).

Gonzalez, Y. Y. Y., Saenz, L. F., Bermeo, J. A., & Chaves, A. F. C. (2013). El papel del trabajo colaborativo en el desarrollo de las habilidades de escritura de estudiantes de primaria. *Revista PROFILE*, *15*(1), 11–26.

Grimshaw, J., McAuley, L. M., Bero, L. A., Grilli, R., Oxman, A. D., Ramsay, C., ... & Zwarenstein, M. (2003). Systematic reviews of the effectiveness of quality improvement strategies and programmes. *Quality and Safety in Health Care*, *12*(4), 298–303.

Guardia, P. (2003). Reladones entre habilidades de alfabetizacion emergente y la lectura, desde nivel transición mayor a primero básico. *Psykhe*, *12*(2), 63–79.

Guevara, B., López, H., García, V., Delgado, S., Hermosillo, G., & Rugerio, J. P. (2008). Habilidades de lectura en primer grado en alumnos de estrato sociocultural bajo. *Revista Mexicana de Investigación Educativa*, *13*(37), 573–597.

Guevara, D. C., & Ordoñez, C. L. (2012). Teaching English to very young learners through authentic communicative performances. *Colombian Applied Linguistics Journal*, *14*(2), 9–27.

Hoddinott, J., Behrman, J. R., Maluccio, J. A., Melgar, P., Quisumbing, A. R., Ramirez-Zea, M., & Martorell, R. (2013). Adult consequences of growth failure in early childhood. *The American Journal of Clinical Nutrition*, ajcn-064584.

Hombrados, J. G., &Waddington, H. (2012). Internal validity in social experiments and quasiexperiments: An assessment tool for reviewers. Mimeo: 3ie.

Iparraguirre, M. S. (2014). Elementary school students as authors of a description: Stages in the learning of writing and linguistic-discursive styles/Alumnos de tercer y séptimo grado de nivel primario como autores de una descripción: etapas en el aprendizaje de la escritura y estilos lingüístico-discursivos. *Infancia y Aprendizaje, 37*(4), 740–784.

Ismail, S. J., Jarvis, E. A., & Borja-Vega, C. (2014). 11 Guyana's hinterland community-based school feeding program (SFP). *Improving Diets and Nutrition*, 124.

Jaichenco, V., & Wilson, M. (2013). El rol de la morfología en el proceso de aprendizaje de la lectura en español. *Interdisciplinaria, 30*(1), 85–99.

Janus, M. (2011). Impact of impairment on children with special needs at school entry: Comparison of school readiness outcomes in Canada, Australia, and Mexico. *Exceptionality Education International, 21*(2), 29–44.

Jiménez, V., Puente, A., Alvarado, J. M., & Arrebillaga, L. (2009). Measuring metacognitive strategies using the reading awareness scale ESCOLA. *Electronic Journal of Research in Educational Psychology*, *7*(2), 779–804.

Jiménez, R. T., Smith, P. H., & Martínez-León, N. (2003). Freedom and form: The language and literacy practices of two Mexican schools. *Reading Research Quarterly*, *38*(4), 488–508.

Kim, Y. S., & Pallante, D. (2012). Predictors of reading skills for kindergartners and first grade students in Spanish: A longitudinal study. *Reading and Writing, 25*(1), 1–22.

Kessler, B., Pollo, T. C., Treiman, R., & Cardoso-Martins, C. (2013). Frequency analyses of prephonological spellings as predictors of success in conventional spelling. *Journal of Learning Disabilities, 46*(3), 252–259.

Kim, Y. S., & Pallante, D. (2012). Predictors of reading skills for kindergartners and first grade students in Spanish: A longitudinal study. *Reading and Writing*, *25*(1), 1–22.

King, E., Samii, C., & Snilstveit, B. (2010). Interventions to promote social cohesion in sub-saharan Africa. *Journal of Development Effectiveness, 2*(3), 336–370.

Kinkhead-Clark, Z. (2014). Family, culture, literacy and the kindergarten classroom: Perspectives of an immigrant teacher. *The International Journal of Early Childhood Learning*, *21*, 1–14.

Koda, K. (2008). Impacts of prior literacy experience on second language learning to read. In K. Koda & A. M. Zehler (Eds.), *Learning to read across languages: Cross-linguistic relationships in first- and second language literacy development* (pp. 68–96). New York, NY: Routledge.

Koda, K., & Reddy, P. (2008). Cross-linguistic transfer in second language reading. *Language Teaching*, *41*(04), 497–508.

Kudo, I., & Bazan, J. (2009). *Measuring beginner reading skills: An empirical evaluation of alternative instruments and their potential use for policymaking and accountability in Peru* (Policy Research Working Paper 4812).

Langou, G. D., & Forteza, P. (2012). *Validating one of the world's largest conditional cash transfer programmes: A case study on how an impact evaluation of Brazil's Bolsa Família Programme helped silence its critics and improve policy* (Working paper 16). New Delhi: International Initiative for Impact Evaluation (3ie).

Larraín, A., Strasser, K., & Lissi, M. R. (2012). Lectura compartida de cuentos y aprendizaje de vocabulario en edad preescolar: un estudio de eficacia. *Estudios de Psicología, 33*(3), 379–383.

Leal Carretero, F., & Suro Sánchez, J. (2012). Las tareas de conciencia fonológica en preescolar: una revisión de las pruebas empleadas en población hispanohablante. *Revista Mexicana de Investigación Educativa, 17*(54), 729–757.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry* (Vol. 75). Sage Publications.

Lincoln, Y. S., and E. G. Guba. (2005). Techniques for collecting and analyzing data (pp. 27–40). In A. Carol Rusaw (Ed.), *Transforming the character of public organizations: Techniques for change agents*. Westport, CT: Quorum Books.

Lipsey, M. W. (1999). Can intervention rehabilitate serious delinquents? *The Annals of the American Academy of Political and Social Science, 564*(1), 142–166.

Lipsey, M. W. (2009). The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims and Offenders, 4*(2), 124–147.

Lockheed, M., Harris, A., & Jayasundera, T. (2010). School improvement plans and student learning in Jamaica. *International Journal of Educational Development, 30*(1), 54–66.

Mahurt, S. F. (1993). *Teacher in transition: A case study of the change process from skills-based to whole language teaching. Research in progress*. Paper presented at the 43rd Annual Meeting of the National Reading Conference, December 1–4, 1993, Charleston, SC.

Maluccio, J. A., Hoddinott, J., Behrman, J. R., Martorell, R., Quisumbing, A. R., & Stein, A. D. (2009). The impact of improving nutrition during early childhood on education among Guatemalan adults. *The Economic Journal, 119*(537), 734–763.

Manrique, M. S., & Borzone, A. M. (2010). La comprensión de cuentos como resolución de problemas en niños de 5 años de sectores urbano-marginales. *Interdisciplinaria, 27*(2), 209–228.

Manrique, A. M. B., & Signorini, A. (1994). Phonological awareness, spelling and reading abilities in Spanish-speaking children. *British Journal of Educational Psychology, 64*(3), 429–439.

Massone, M. I., & Baez, M. (2009) Deaf children's construction of writing. *Sign Language Studies*, *9*(4), 457–479.

Matute, E., Montiel, T., Pinto, N., Rosselli, M., Ardila, A., & Zarabozo, D. (2012). Comparing cognitive performance in illiterate and literate children. *International Review of Education*, *58*(1), 109–127.

McMillan, J. H., & Schumacher, S. (2001). *Research in education: A conceptual introduction*. Little, Brown.

Medeiros, J. G., Antunes, L., Pokreviescki, J. E. J., Bottenberg, D. G., Ferreira, C. D. A., & Cavalhieri, K. E. (2011). Emergência de leitura de frases a partir do ensino de suas unidades constituintes. *Acta Comportamentalia, 19*(3), 317–342.

Medina, C. L., & Costa, M. D. R. (2013). Latino media and critical literacy pedagogies: Children's scripting of telenovelas discourses. *Journal of Language and Literacy Education*, *9*(1), 161–184.

Melchiori, L. E., de Souza, D. G., & de Rose, J. C. (2012). Aprendizagem de leitura por meio de um procedimento de discriminação sem erros (exclusão): uma replicação com pré-escolares. *Psicologia: Teoria e Pesquisa, 8*(1), 101–111.

Mendive, S., Weiland, C., Yoshikawa, H., & Snow, C. (2016). Opening the black box: Intervention fidelity in a randomized trial of a preschool teacher professional development program. *Journal of Educational Psychology*, *108*(1), 130.

Miguel, E., & Kremer, M. (2004). Worms: Identifying impacts on education and health in the present of treatment externalities. *Econometrica, 72*(1), 159–217.

Moneda, I. X. G., Velasco, A. S., Figueroa, S. P., & Flores, T. P. (2009). Habilidades psicolingüísticas al ingreso y egreso del jardín de niños. *Revista Intercontinental de Psicología y Educación*, *11*(2), 13–36.

Morales, A. M. F., Van de Vijver, F. J., & Poortinga, Y. H. (2013). Differential item functioning and educational risk factors in Guatemalan reading assessment. *Revista Interamericana de Psicología, 47*, 3.

Muñoz, C. (2002). Aprendizaje de la lectura y conciencia fonológica: Un enfoque psicolingüístico del proceso de alfacetización inicial. *Psykhe*, *11*(1), 29–42.

Murad, C. R., & Topping, K. J. (2000). Parents as reading tutors for first graders in Brazil. *School Psychology International*, *21*(2), 152–171.

Nakamura, P., & de Hoop, T. (2014). *Facilitating reading acquisition in multilingual environments in India (FRAME-India). Final report*. American Institutes for Research.

Neugebauer, S. R., & Currie-Rubin, R. (2009). Read-alouds in Calca, Peru: A bilingual indigenous context. *The Reading Teacher*, *62*(5), 396–405.

Oliveira, Q. L. D. (1996). Três instrumentos de avaliação de habilidades para aprendizagem da leitura e escrita. *Psicologia: Teoria e Pesquisa, 12*(1), 83–96.

Páez, M. M., Tabors, P. O., & López, L. M. (2007). Dual language and literacy development of Spanish-speaking preschool children. *Journal of Applied Developmental Psychology, 28*(2), 85–102.

Pallante, D. H., & Kim, Y. S. (2013). The effect of a multicomponent literacy instruction model on literacy growth for kindergartners and first-grade students in Chile. *International Journal of Psychology, 48*(5), 747–761.

Patashnick, J. and M. Rich (2004). Researching human experience: Video intervention/prevention assessment (VIA). *Australian Journal of Information Systems 12*(2), 103–111.

Patton, M. Q. (2015a). *Qualitative Evaluation and Research Methods*. SAGE Publications, Inc.

Patton, M. Q. (2015b). In D. M. Fetterman, S. J. Kaftarian, and A. Wandersman (Eds.), Empowerment evaluation: Knowledge and tools for self-assessment, evaluation, capacity building, and accountability (2nd ed.). Sage Publications. *Evaluation and Program Planning, 52,* 15–18.

Pino, M., & Bravo, L. (2005). La memoria visual como predictor del aprendizaje de la lectura. *Psykhe (Santiago), 14*(1), 47–53.

Plana, M. D., & Fumagalli, J. (2013). Habilidades y conocimientos constitutivos de la alfabetización temprana: Semejanzas y diferencias según el entorno social y las oportunidades educativas. *Interdisciplinaria, 30*(1), 5–24.

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., ... & Duffy, S. (2006). Guidance on the conduct of narrative synthesis in systematic reviews. *ESRC Methods Programme, 15*(1), 47–71.

Powell, C. A., Walker, S. P., Chang, S. M., & Grantham-McGregor, S. M. (1998). Nutrition and education: A randomized trial of the effects of breakfast in rural primary school children. *The American Journal of Clinical Nutrition, 68*(4), 873–879.

Porras González, N. I. (2010). Teaching English through stories: A meaningful and fun way for children to learn the language. *Profile Issues in Teachers' Professional Development, 12*(1), 95–106.

Pritchett, L., & Sandefur, J. (2013). Context matters for size: Why external validity claims and development practice do not mix. *Journal of Globalization and Development, 4*(2), 161–197.

Querejeta, M., Piacente, T., Guerrero Ortiz-Hernán, B., & Alva Canto, E. A. (2013). La separación entre palabras en la escritura de niños que inician la escolaridad primaria. *Interdisciplinaria, 30*(1), 45–64.

Ramírez, A. B., Verdugo, V. C., & Sánchez, B. (2000). Predictores del desempeño en lectura y escritura de niños de primer grado. *Revista de Psicología*, *18*(2), 295–314.

Rego, L. L. B. (1997). The connection between syntactic awareness and reading: Evidence from portuguese-speaking children taught by a phonic method. *International Journal of Behavioral Development*, *20*(2), 349–365.

Reigosa-Crespo, V., González-Alemañy, E., León, T., Torres, R., Mosquera, R., & Valdés-Sosa, M. (2013). Numerical capacities as domain-specific predictors beyond early mathematics learning: A longitudinal study. *PLoS One*, *8*(11), e79711.

Reynoso-Alcántara, V., Bernal, J., Silva-Pereyra, J., Rodríguez, M., Yáñez, G., Fernández, T., & del Río, Y. (2010). Procesamiento fonológico y léxico en niños normolectores de educación primaria. *Infancia y Aprendizaje*, *33*(3), 413–425.

Ribeiro, N., & Souza, L. A. D. P. (2012). Efeitos do (s) letramento (s) na constituição social do sujeito: considerações fonoaudiológicas. *Revista CEFAC*, *14*(5), 808–15.

Rindermann, H., Baumeister, A. E. E., & Gröper, A. (2014). Cognitive ability of preschool, primary and secondary school children in Costa Rica. *Journal of Biosocial Science*, *46*, 199–213.

Roofe, C. G. (2014). One size fits all: Perceptions of the revised primary curriculum at grades one to three in Jamaica. *Research in Comparative and International Education*, *9*(1), 4–15.

Rosado, C. T. D. C. L., & Campelo, M. E. C. H. (2011). Elementary education: The time and voice of children. *Ensaio: Avaliação e Políticas Públicas em Educação*, *19*(71), 401–424.

Rosas, R., Ceric, F., Aparicio, A., Arango, P., Arroyo, R., Benavente, C., ... & Tenorio, M. (2015). Traditional assessment or invisible assessment using games? New frontiers in cognitive assessment. *Psykhe*, *24*(1).

Salazar, C. E., Amon, E., & Ortiz de Urdiales, J. (1996). Pruebas que se usan para predecir adquisición de lectura en la ciudad de Guatemala: Validez predictiva y reanálisis del ABC. *Revista Latinoamericana de Psicología*, *28*(2), 273–292.

Salazar-Reyes, L., & Vega-Pérez, L. O. (2013). Relaciones diferenciales entre experiencias de alfabetización y habilidades de alfabetización emergente. *Educación y Educadores*, *16*(2), 311–325.

Salles, J. F. D., & Parente, M. A. D. M. P. (2002). Processos cognitivos na leitura de palavras em crianças: relações com compreensão e tempo de leitura. *Psicologia: Reflexão e Crítica. Porto Alegre*, *15*(2), 321–331.

Schuelke-Leech, B. A., Barry, B., Muratori, M., & Yurkovich, B. J. (2015). Big Data issues and opportunities for electric utilities. *Renewable and Sustainable Energy Reviews*, *52*, 937–947.

Sénéchal, M., & LeFevre, J. A. (2002). Parental involvement in the development of children's reading skill: A five-year longitudinal study. *Child Development, 73*(2), 445–460.

Shadish, W. R. (2002). Revisiting field experimentation: field notes for the future. *Psychological Methods, 7*(1), 3.

Silva, L. S. L., Aristizabal, C. P. D., De Luque, G. L. C., Muñoz, E. D. C. A., Cantillo, M. Á., & Kemp, S. (2013). Habilidades Prelectoras de estudiantes de preescolar en la region caribe colombiana. *Zona Próxima, 19*.

Silva, M., Strasser, K., & Cain, K. (2014). Early narrative skills in Chilean preschool: Questions scaffold the production of coherent narratives. *Early Childhood Research Quarterly, 29*(2), 205–213.

Simeon, D. T., Grantham-McGregor, S. M., & Wong, M. S. (1995). Treatment of Trichuris trichiura infections improves growth, spelling scores and school attendance in some children. *Community and International Nutrition, 125*, 1875–1883.

Snilstveit, B., Oliver, S., & Vojtkova, M. (2012). Narrative approaches to systematic review and synthesis of evidence for international development policy and practice. *Journal of Development Effectiveness, 4*(3), 409–429.

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

Spier, E., Britto, P., Pigott, T., Kidron, Y., Lane, J., Roehlkepartain, E., Scales, P., Wagner, D., McCarthy, M., Song, M., & Glover, J. (2016). *Parental, familial, and community support interventions to improve children's literacy in developing countries: A systematic review*. The Campbell Collaboration Library of Systematic Reviews.

Stein, A., & Rosemberg, C. R. (2012). Redes de colaboración en situaciones de alfabetización con niños pequeños: Un estudio en poblaciones urbano marginales de Argentina. *Interdisciplinaria, 29*(1), 95–108.

Tapia, J. P. R., & Benítez, Y. G. (2013). Desarrollo de habilidades conductuales maternas para promover la alfabetización inicial en niños preescolares. *Acta Colombiana de Psicología, 16*(1), 81–90.

The Joanna Briggs Institute. (2014). *Joanna Briggs Institute reviewers' manual 2014 edition*. South Australia: University of Adelaide.

Torrecilla, F. J. M., & Carrasco, M. R. (2014). Consecuencias del trabajo infantil en el desempeño escolar: Estudiantes latinoamericanos de educación primaria. *Latin American Research Review, 49*(2), 84–106.

Treiman, R., Kessler, B., & Pollo, T. C. (2006). Learning about the letter name subset of the vocabulary: Evidence from U.S. and Brazilian preschoolers. *Applied Psycholinguistics, 27*(2), 211.

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2014). *Regional overview: Latin America and the Caribbean: Education for all global monitoring report 2015*. UNESCO.

Verhoeven, L. (1994). Transfer in bilingual development: The linguistic interdependence hypothesis revisited. *Language Learning, 44*, 381–415.

Villalon, M., & San Francisco, A. (2001). *Assessment and evaluation of phonological awareness, concepts of print, and early reading and writing in young Chilean children: A comparison with international results*. ERIC.

Vivas, E. (1996). Effects of story reading on language. *Language Learning, 46*(2), 189–216.

Volk, D., & de Acosta, M. (2001). "Many differing ladders, many ways to climb...": Literacy events in the bilingual classroom, homes, and community of three Puerto Rican kindergartners. *Journal of Early Childhood Literacy, 1*(2), 193–224.

Volk, D., & de Acosta, M. (2003). Reinventing texts and contexts: Syncretic literacy events in young Puerto Rican children's homes. *Research in the Teaching of English*, 8–48.

Waddington, H., White, H., Snilstveit, B., Hombrados, J. G., Vojtkova, M., Davies, P., ... & Valentine, J. C. (2012). How to do a good systematic review of effects in international development: A tool kit. *Journal of Development Effectiveness*, *4*(3), 359–387.

Warrican, S. J., Down, L., & Spencer-Ernandez, J. (2008). Exemplary teaching in the Caribbean: Experiences from early literacy classrooms. *Journal of Eastern Caribbean Studies*, *33*(1).

Webster, P. S. (2009). Exploring the literature of fact. *The Reading Teacher, 62*(8), 662–671.

White, H., & Phillips, D. (2012). *Addressing attribution of cause and effect in small n impact evaluations: Towards an integrated framework* (Working paper 15). New Delhi: International Initiative for Impact Evaluation (3ie).

World Conference on Education for All. (1990). *Meeting back learning needs: A vision for the 1990s* (Background document, World Conference on Education for All, March 5-8, 1990, Jomtien, Thailand). New York: The Inter-Agency Commission (UNDO, UNESCO, UNICEF, World Bank) for the World Conference on Education for All and UNICEF House.

Yoshikawa, H., Leyva, D., Snow, C. E., Treviño, E., Barata, M., Weiland, C., Gomez, C., Moreno, L., Rolla, A., D'Sa, N. & Arbour, M. C. (2015). Experimental impacts of a teacher professional development program in Chile on preschool classroom quality and child outcomes. *Developmental Psychology, 51*(3), 309.

Appendix B. Search String Modification Process

Science Direct - 3748 search results

"Read*" OR Literacy AND "primary school*" OR "primary grade*" OR {grades 1 through 3} OR {grades 1 to 3} OR {grades 1-3} OR {first through third} OR {Grade 1} OR {first grade*} OR {grade 2} OR {second grade*} OR {grade 3} OR "third grade*" OR "early grade*" OR elementary OR "kindergarten*" OR "pre-school*" OR "preschool*" OR "pre-kindergarten*" OR "prekindergarten*" OR preK OR "pre-K" OR {early childhood} AND "Latin America*" OR Caribbean OR "South America*" OR {Antigua and Barbuda} OR Argentina OR Aruba OR Bahamas OR Barbados OR Belize OR Bermuda OR "Bolivia*" OR "Brazil*" OR "British Virgin Islands" OR "Cayman Islands" OR "Chile*" OR "Colombia*" OR "Costa Rica*" OR "Cuba*" OR Curacao OR "Dominica*" OR "Dominican Republic" OR "Ecuador*" OR "El Salvador*" OR "French Guiana*" OR "Grenada*" OR Guadeloupe OR "Guatemala*" OR "Guyana*" OR "Haiti*" OR Honduras OR "Jamaica*" OR Martinique OR Mexico OR Mont Serrat OR "Netherlands Antilles" OR "Nicaragua*" OR "Panama*" OR "Paraguay*" OR "Peru*" OR "Puerto Rico" OR "Saint Barthelemy" OR "Saint Kitts and Nevis" OR "Saint Lucia*" OR "Saint-Martin" OR {Saint Vincent and the Grenadines} OR "Sint Maarten" OR Suriname OR {Trinidad and Tobago} OR {Turks and Caicos} OR Uruguay OR {Virgin Islands} OR Venezuela

According to the rules of Science Direct, a phrase must be enclosed in { } to ensure that the phrase is exact, and includes stop words. I enclosed only those phrases with stop words. Date range: 1990 to 2016

2. Removed all asterisks and added parentheses between the three components to ensure proper order of operations.

("Read" OR Literacy) AND ("primary school" OR "primary grade" OR {grades 1 through 3} OR {grades 1 to 3} OR {grades 1-3} OR {first through third} OR {Grade 1} OR {first grade} OR {grade 2} OR {second grade} OR {grade 3} OR "third grade" OR "early grade" OR elementary OR "kindergarten" OR "pre-school" OR "preschool" OR "pre-kindergarten" OR "prekindergarten" OR preK OR "pre-K" OR {early childhood}) AND ("Latin America" OR Caribbean OR "South America" OR {Antigua and Barbuda} OR Argentina OR Aruba OR Bahamas OR Barbados OR Belize OR Bermuda OR "Bolivia" OR "Brazil" OR "British Virgin Islands" OR "Cayman Islands" OR "Chile" OR "Colombia" OR "Costa Rica" OR "Cuba" OR Curacao OR "Dominica" OR "Dominican Republic" OR "Ecuador" OR "El Salvador" OR "French Guiana" OR "Grenada" OR Guadeloupe OR "Guatemala" OR "Guyana" OR "Haiti" OR Honduras OR "Jamaica" OR Martinique OR Mexico OR Mont Serrat OR "Netherlands Antilles" OR "Nicaragua" OR "Panama" OR "Paraguay" OR "Peru" OR "Puerto Rico" OR "Saint Barthelemy" OR {Saint Kitts and Nevis} OR "Saint Lucia" OR "Saint-Martin" OR {Saint Vincent and the Grenadines} OR "Sint Maarten" OR Suriname OR {Trinidad and Tobago} OR {Turks and Caicos} OR Uruguay OR {Virgin Islands} OR Venezuela)

3. It wasn't immediately clear that the relevant articles yielded on the first search were also present in the second search, so I selected a few relevant articles from the first set of results to look for within the second set of results. I found these same articles within the second set of results, so the smaller number of results also include the relevant articles yielded from the first entry.

Results: 2,053

SAGE

"early grade" AND literacy (all fields)

OR "early grade" AND reading (all fields)

OR childhood AND reading (all fields)

OR childhood AND literacy (all fields)

AND South America OR Latin America (all fields)

OR Caribbean OR Central America (all fields)

From Jan 1990 through Jan 2016

Method 1: Manually selected disciplines (4680 results)

Education

Ethnic Studies

Family Studies

Gender Studies

Group Studies

Language & Linguistics

Regional Studies

Research Methods & Evaluation

Special Education

Method 2: Manually selected Sage journals included (964 results)

American Educational Research Journal

Australian Journal of Education

Child Language Teaching and Therapy

Childhood: A journal of global child research

Contemporary Education Dialogue

Contemporary Issues in Early Childhood

Education and Urban Society

Education, Citizenship, and Social Justice

Educational Administration Quarterly: The Journal of Leadership for Effective & Equitable Organizations

Educational Evaluations and Policy Analysis

Educational Horizons

Educational Policy: An Interdisciplinary Journal of Policy and Practice

Educational Researcher

European Educational Research Journal

Exceptional Children

Gifted Children Quarterly

Gifted Child Today

Gifted Education International

Global Studies of Childhood

International Journal of Christianity & Education

Journal for the Education of the Gifted

Journal of Early Childhood Literacy

Journal of Early Childhood Research

Journal of Education for Sustainable Development

Journal of Educational and Behavioral Statistics

Journal of Experiential Education

Journal of Literacy Research

Journal of Planning Education and Research

Journal of Research in International Education

The Journal of Special Education

Journal of Studies in International Education

Journal of Transformative Education

Language and Linguistics

Language and Literature

Language and Speech

Language Teaching Research

Management in Education

Power and Education

Remedial and Special Education

Research in Comparative and International Education

Review of Educational Research

Review of Research in Education

Sociology of Education

Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children

TEACHING Exceptional Children

Theory and Research in Education

Topics in Early Childhood Special Education

Urban Education

Young

Young Exceptional Children

Youth & Society

Taylor & Francis 3442 results

("early childhood" OR "early grade" AND Read OR Literacy) AND ("Latin America" OR Caribbean OR "South America" OR "Central America")

From Jan 1990 through Jan 2016

Subject Areas

Education

Language and Literature

Note: I purposely excluded other subject areas such as “Latin American Studies” because they yielded irrelevant results.

Updated search terms as follows:

("early childhood" OR "early grade") AND (Read* OR Literacy) AND ("Latin America" OR Caribbean OR "South America" OR "Central America")

Added date parameters, but did not need to add additional subject area limitations

1258 results

JSTOR

Original search string was too long to accept. The number of characters is limited across 7 fields.

By entering the search terms manually, as follows, I got over a million results:

read* OR literacy

AND “early grade”

OR “early child”

AND “Latin America*”

OR Caribbean

OR “South America*”

Within the search engine, the logic structure was depicted as follows:

((((((Read* OR Literacy) AND ("early grade*")) OR ("early child*")) AND (Latin America*)) OR (Caribbean)) OR (South America*))

I changed this structure according to the Boolean logic structure, which is as follows:

(Read* OR Literacy) AND (("early grade*") OR ("early child*")) AND ((Latin America*) OR (Caribbean) OR (South America*)) **Results: 2,801**

I removed the parentheses and quotation marks to determine relevancy and number of articles and found that quotation marks are necessary to keep phrases together.

(Read* OR Literacy) AND (("early grade*") OR ("early child*")) AND (("Latin America*") OR (Caribbean) OR ("South America*")) **Results: 258**

(Read* OR Literacy) AND ("early grade*" OR "early child*") AND ("Latin America*" OR Caribbean OR "South America*") **Results: 258**

(Read* OR Literacy) AND ((early grade*) OR (early child*)) AND ((Latin America*) OR (Caribbean) OR (South America*)) **Results: 588,645**

Parentheses are good for ordering, while quotation marks are good for phrases, even with the * included for variation.

I added the term “primary grade” to include more relevant results

(Read* OR Literacy) AND ("early grade*" OR "early child*" OR "primary grade") AND (Latin America* OR Caribbean OR South America*) **Results: 3,652**

I removed all asterisks from phrases:

(Read* OR Literacy) AND ("early grade" OR "early child" OR "primary grade") AND ("Latin America" OR Caribbean OR "South America")

Some relevant results, but mostly not – eliminated some relevant results in previous search

Results: 336

EBSCO

Databases searched:

ERIC

Academic Search Premier

Education Source

PsycINFO

CINAHL

Psychology and Behavioral Sciences Collection

SocINDEX with Full Text

EconLit

The original search string was used in the format provided.

Read* OR Literacy

AND

"primary school*" OR "primary grade*" OR “grades 1 through 3” OR “grades 1 to 3” OR “grades 1-3” OR "first through third" OR “Grade 1” OR “first grade*” OR “grade 2” OR “second grade*” OR “grade 3” OR “third grade*” OR “early grade*” OR elementary OR kindergarten* OR pre-school* OR preschool* OR pre-kindergarten* OR prekindergarten* OR preK OR pre-K OR “early childhood”

AND

“Latin America*” OR Caribbean OR “South America*” OR Antigua* and Barbuda OR Argentina OR Aruba OR Bahamas OR Barbados OR Belize OR Bermuda OR Bolivia* OR Brazil* OR “British Virgin Islands” OR “Cayman Islands” OR Chile* OR Colombia* OR “Costa Rica*” OR Cuba* OR Curacao OR Dominica* OR “Dominican Republic” OR Ecuador* OR “El Salvador*” OR “French Guiana*” OR Grenada* OR Guadeloupe OR Guatemala* OR Guyana* OR Haiti* OR Honduras OR Jamaica* OR Martinique OR Mexico OR Mont Serrat OR “Netherlands Antilles” OR Nicaragua* OR Panama* OR Paraguay* OR Peru* OR “Puerto Rico” OR “Saint Barthelemy” OR “Saint Kitts and Nevis” OR “Saint Lucia*” OR “Saint-Martin” OR “Saint Vincent and the Grenadines” OR “Sint Maarten” OR Suriname OR Trinidad and Tobago OR “Turks and Caicos” OR Uruguay OR “Virgin Islands” OR Venezuela

From 1990 to 2015

Results: 2,779

Modified research results have removed asterisks that are within quotes, and is written as follows:

Read* OR Literacy

AND

"primary school" OR "primary grade" OR “grades 1 through 3” OR “grades 1 to 3” OR “grades 1-3” OR "first through third" OR “Grade 1” OR “first grade” OR “grade 2” OR “second grade” OR “grade 3” OR “third grade” OR “early grade” OR elementary OR kindergarten* OR pre-school* OR preschool* OR pre-kindergarten* OR prekindergarten* OR preK OR pre-K OR “early childhood”

AND

“Latin America” OR Caribbean OR “South America” OR Antigua* and Barbuda OR Argentina OR Aruba OR Bahamas OR Barbados OR Belize OR Bermuda OR Bolivia* OR Brazil* OR “British Virgin Islands” OR “Cayman Islands” OR Chile* OR Colombia* OR “Costa Rica” OR Cuba* OR Curacao OR Dominica* OR “Dominican Republic” OR Ecuador* OR “El Salvador” OR “French Guiana” OR Grenada* OR Guadeloupe OR Guatemala* OR Guyana* OR Haiti* OR Honduras OR Jamaica* OR Martinique OR Mexico OR Mont Serrat OR “Netherlands Antilles” OR Nicaragua* OR Panama* OR Paraguay* OR Peru* OR “Puerto Rico” OR “Saint Barthelemy” OR “Saint Kitts and Nevis” OR “Saint Lucia” OR “Saint-Martin” OR “Saint Vincent and the Grenadines” OR “Sint Maarten” OR Suriname OR Trinidad and Tobago OR “Turks and Caicos” OR Uruguay OR “Virgin Islands” OR Venezuela

Results: 2,612

Cochrane

This is a medical database that is part of Wiley Online journal. See Wiley for explanation.

Wiley

Tried entering the original string, response said, “search terms should be more than 1 characters long”

Tried entering into the smaller search engine, but the database could not handle computing the command

Entered:

("Read*" OR Literacy) AND ("primary school*" OR "primary grade*" OR "grades 1 through 3" OR "grades 1 to 3" OR "grades 1-3" OR "first through third" OR "Grade 1" OR "first grade*" OR "grade 2" OR "second grade*" OR "grade 3" OR "third grade*" OR "early grade*" OR elementary OR "kindergarten*" OR "preschool*" OR "prekindergarten*" OR preK OR "early childhood") AND ("Latin America*" OR Caribbean OR "South America*" OR "Central America*") **Results: 2580083**

A mix of relevant and irrelevant results.

I tried again by entering the same string, but selected “full text” for the fields. Excessive and irrelevant results. I tried “abstract” with excessive and irrelevant results.

I tried a new string:

("Read*" OR Literacy) AND ("primary school*" OR "primary grade*" OR "early grade*" OR elementary OR "kindergarten*" OR "preschool*" OR "prekindergarten*" OR preK OR "early childhood") AND ("Latin America*" OR Caribbean OR "South America*" OR "Central America*") **Results: 12962**

Irrelevant results.

To weed out irrelevant results, I tried adding NOT psychology* NOT disease*

("Read*" OR Literacy) AND ("primary school*" OR "primary grade*" OR "early grade*" OR elementary OR "kindergarten*" OR "preschool*" OR "prekindergarten*" OR preK OR "early childhood") AND ("Latin America*" OR Caribbean OR "South America*" OR "Central America*") NOT psycholog* NOT disease* **Results: 3,540**

These articles seem relevant.

I removed quotation marks on one-word entries, and asterisks from phrases. I also added an asterisk **before** kindergarten to account for prekindergarten.

(Read* OR Literacy) AND ("primary school" OR "primary grade" OR "early grade" OR elementary OR *kindergarten* OR preschool* OR preK OR "early childhood") AND ("Latin America" OR Caribbean OR "South America" OR "Central America") NOT psycholog* NOT disease* **Results: 2390**

Checked to see if the same relevant articles that appeared in entry from step #6 appeared for the entry from step #7, confirmed availability.

ProQuest

Signed up for a free trial and was limited to 6 journals, selected the following journals:

Australian Education Index

CBCA Education

ERIC

Linguistics and Language Behavior Abstracts

Proquest Learning: Literature

Proquest Education Journals

Got confirmation and notification that they will email me more information about this free trial in a few days

No information as of 7/20

The Campbell Library

Entered the string as follows:

(Read* OR Literacy)

AND

("primary school*" OR "primary grade*" OR “grades 1 through 3” OR “grades 1 to 3” OR “grades 1-3” OR "first through third" OR “Grade 1” OR first grade* OR “grade 2” OR second grade* OR “grade 3” OR third grade* OR early grade* OR elementary OR kindergarten* OR pre-school* OR preschool* OR pre-kindergarten* OR prekindergarten* OR preK OR pre-K OR “early childhood”)

AND

(Latin America* OR Caribbean OR South America* OR Antigua* and Barbuda OR Argentina OR Aruba OR Bahamas OR Barbados OR Belize OR Bermuda OR Bolivia* OR Brazil* OR “British Virgin Islands” OR “Cayman Islands” OR Chile* OR Colombia* OR Costa Rica* OR Cuba* OR Curacao OR Dominica* OR “Dominican Republic” OR Ecuador* OR El Salvador* OR French Guiana* OR Grenada* OR Guadeloupe OR Guatemala* OR Guyana* OR Haiti* OR Honduras OR Jamaica* OR Martinique OR Mexico OR Mont Serrat OR “Netherlands Antilles” OR Nicaragua* OR Panama* OR Paraguay* OR Peru* OR “Puerto Rico” OR “Saint Barthelemy” OR “Saint Kitts and Nevis” OR Saint Lucia* OR “Saint-Martin” OR “Saint Vincent and the Grenadines” OR “Sint Maarten” OR Suriname OR Trinidad and Tobago OR “Turks and Caicos” OR Uruguay OR “Virgin Islands” OR Venezuela)

I modified the string to enclose all the countries with “and” in their names

(Read* OR Literacy)

AND

("primary school*" OR "primary grade*" OR "grades 1 through 3" OR "grades 1 to 3" OR "grades 1-3" OR "first through third" OR "Grade 1" OR first grade* OR "grade 2" OR second grade* OR "grade 3" OR third grade* OR early grade* OR elementary OR kindergarten* OR pre-school* OR preschool* OR pre-kindergarten* OR prekindergarten* OR preK OR pre-K OR "early childhood")

AND

(Latin America* OR Caribbean OR South America* OR Antigua* and Barbuda OR Argentina OR Aruba OR Bahamas OR Barbados OR Belize OR Bermuda OR Bolivia* OR Brazil* OR “British Virgin Islands” OR “Cayman Islands” OR Chile* OR Colombia* OR “Costa Rica*” OR Cuba* OR Curacao OR Dominica* OR “Dominican Republic” OR Ecuador* OR “El Salvador*” OR “French Guiana*” OR Grenada* OR Guadeloupe OR Guatemala* OR Guyana* OR Haiti* OR Honduras OR Jamaica* OR Martinique OR Mexico OR Mont Serrat OR “Netherlands Antilles” OR Nicaragua* OR Panama* OR Paraguay* OR Peru* OR “Puerto Rico” OR “Saint Barthelemy” OR “Saint Kitts and Nevis” OR “Saint Lucia*” OR “Saint-Martin” OR “Saint Vincent and the Grenadines” OR “Sint Maarten” OR Suriname OR “Trinidad and Tobago” OR “Turks and Caicos” OR Uruguay OR “Virgin Islands” OR Venezuela)

Since the earliest publication goes back to 2003, I did not need to make additional date adjustments. Results look relevant. R**esults: 217**

Updated entries to remove all asterisks within quotes

(Read* OR Literacy)

AND

("primary school" OR "primary grade" OR "grades 1 through 3" OR "grades 1 to 3" OR "grades 1-3" OR "first through third" OR "Grade 1" OR first grade* OR "grade 2" OR second grade* OR "grade 3" OR third

grade* OR early grade* OR elementary OR kindergarten* OR pre-school* OR preschool* OR pre-kindergarten* OR prekindergarten* OR preK OR pre-K OR "early childhood")

AND

(“Latin America” OR Caribbean OR “South America” OR Antigua* and Barbuda OR Argentina OR Aruba OR Bahamas OR Barbados OR Belize OR Bermuda OR Bolivia* OR Brazil* OR “British Virgin Islands” OR “Cayman Islands” OR Chile* OR Colombia* OR “Costa Rica” OR Cuba* OR Curacao OR Dominica* OR “Dominican Republic” OR Ecuador* OR “El Salvador” OR “French Guiana” OR Grenada* OR Guadeloupe OR Guatemala* OR Guyana* OR Haiti* OR Honduras OR Jamaica* OR Martinique OR Mexico OR Mont Serrat OR “Netherlands Antilles” OR Nicaragua* OR Panama* OR Paraguay* OR Peru* OR “Puerto Rico” OR “Saint Barthelemy” OR “Saint Kitts and Nevis” OR “Saint Lucia” OR “Saint-Martin” OR “Saint Vincent and the Grenadines” OR “Sint Maarten” OR Suriname OR “Trinidad and Tobago” OR “Turks and Caicos” OR Uruguay OR “Virgin Islands” OR Venezuela)

No hits! I backtracked, and step 2 also yielded no hits! I updated the quotation marks to reflect Unicode, and yielded 8 hits. Then I again removed all asterisks within quotation marks, as follows. This also yielded 8 hits.

(Read* OR Literacy)

AND

("primary school" OR "primary grade" OR "grades 1 through 3" OR "grades 1 to 3" OR "grades 1-3" OR "first through third" OR "Grade 1" OR first grade* OR "grade 2" OR second grade* OR "grade 3" OR third grade* OR early grade* OR elementary OR kindergarten* OR pre-school* OR preschool* OR pre-kindergarten* OR prekindergarten* OR preK OR pre-K OR "early childhood")

AND

("Latin America" OR Caribbean OR "South America" OR Antigua* and Barbuda OR Argentina OR Aruba OR Bahamas OR Barbados OR Belize OR Bermuda OR Bolivia* OR Brazil* OR "British Virgin Islands" OR "Cayman Islands" OR Chile* OR Colombia* OR "Costa Rica" OR Cuba* OR Curacao OR Dominica* OR "Dominican Republic" OR Ecuador* OR "El Salvador" OR "French Guiana" OR Grenada* OR Guadeloupe OR Guatemala* OR Guyana* OR Haiti* OR Honduras OR Jamaica* OR Martinique OR Mexico OR Mont Serrat OR "Netherlands Antilles" OR Nicaragua* OR Panama* OR Paraguay* OR Peru* OR "Puerto Rico" OR "Saint Barthelemy" OR "Saint Kitts and Nevis" OR "Saint Lucia" OR "Saint-Martin" OR "Saint Vincent and the Grenadines" OR "Sint Maarten" OR Suriname OR "Trinidad and Tobago" OR "Turks and Caicos" OR Uruguay OR "Virgin Islands" OR Venezuela)

Search “help” only says, to use an asterisk to search for multiple characters after a search strings, so I removed all quotes, which yielded no hits. Then I entered it as follows (enclosing phrases in quotes, except those with asterisks):

Read* OR Literacy

AND

"primary school" OR "primary grade" OR "grades 1 through 3" OR "grades 1 to 3" OR "grades 1-3" OR "first through third" OR "Grade 1" OR first grade* OR "grade 2" OR second grade* OR "grade 3" OR third

grade* OR early grade* OR elementary OR kindergarten* OR pre-school* OR preschool* OR *kindergarten* OR preK OR pre-K OR "early childhood"

AND

Latin America* OR Caribbean OR South America* OR "Antigua and Barbuda" OR Argentina OR Aruba OR Bahamas OR Barbados OR Belize OR Bermuda OR Bolivia* OR Brazil* OR "British Virgin Islands" OR "Cayman Islands" OR Chile* OR Colombia* OR "Costa Rica" OR Cuba* OR Curacao OR Dominica* OR "Dominican Republic" OR Ecuador* OR "El Salvador" OR "French Guiana" OR Grenada* OR Guadeloupe OR Guatemala* OR Guyana* OR Haiti* OR Honduras OR Jamaica* OR Martinique OR Mexico OR "Mont Serrat" OR "Netherlands Antilles" OR Nicaragua* OR Panama* OR Paraguay* OR Peru* OR "Puerto Rico" OR "Saint Barthelemy" OR "Saint Kitts and Nevis" OR "Saint Lucia" OR "Saint-Martin" OR "Saint Vincent and the Grenadines" OR "Sint Maarten" OR Suriname OR "Trinidad and Tobago" OR "Turks and Caicos" OR Uruguay OR "Virgin Islands" OR Venezuela

Results: 189

Dissertation Abstracts

This is part of Proquest. Due to limited access to Proquest via free trial subscriptions, I cannot access this.

Directory of Open Access Journals (DOAJ)

Entered original search string, yielded 0 results

Eliminated the countries, yielded 0 results

Entered ("read*" OR literacy) AND ("early grade" OR childhood) AND ("South America" OR "Latin America" OR "Central America" OR Caribbean), yielded 93 results, mixed results

Tried filtering, but it eliminated some relevant results

Does not allow for date restrictions, but all the articles are recent

Results: 94

Modified the entry to be as follows (removed quotes from “read”):

(read* OR literacy) AND ("early grade" OR childhood) AND ("South America" OR "Latin America" OR "Central America" OR Caribbean)

Did not make a difference in search results. Both entries (#3 and #6) work, and yield the same results.

Directory of Open Access Books (DOAB)

Entered original search string into “simple search”, yielded 104 irrelevant results

Entered the same search string into “advanced search”, yielded 10 irrelevant results

Entered modified search string into “advanced search” as follows:

(Read* OR all:Literacy) AND ("primary all:school*" OR all:"primary OR all:"grades OR all:"grades OR all:"grades OR all:"first OR all:"Grade OR all:first all:grade* OR all:"grade OR all:second all:grade* OR all:"grade OR all:third all:grade* OR all:early all:grade* OR all:elementary OR all:kindergarten* OR all:pre-school* OR all:preschool* OR all:pre-kindergarten* OR all:prekindergarten* OR all:preK OR all:pre-K OR all:"early childhood" all:)

AND

(all:("South all:America" OR all:"Latin OR all:"Central OR all:Caribbean))

Yielded 17 results that were irrelevant. It appears this database does not have any useful results

3ie

Cannot fit original search string into search engine

Modified search string and got 1 irrelevant result:

("Read*" OR Literacy) AND ("primary school*" OR "primary grade*" OR "early grade") AND ("South America*" OR "Latin America*" OR "Central America*" OR Caribbean)

Modified search string and got 3 irrelevant results:

(Read OR Literacy) AND (primary school OR primary grade OR early grade) AND (South America OR Latin America OR Central America OR Caribbean)

Modified search string and got 6 results, only 1 of which was relevant:

(Read OR Literacy) AND (primary school OR primary grade OR early grade OR childhood) AND (South America OR Latin America OR Central America OR Caribbean)

Link to relevant article:

http://www.scielo.br/scielo.php?pid=S0101-41612012000100004&script=sci_arttext

To further ensure the accuracy of these results, I experimented and tried entering string #2, but with the quotes removed from “read”. I also got irrelevant results.

International Bibliography of the Social Sciences

See Proquest

British Library for Development Studies

Entered the results as follows, with no results

(Read* OR Literacy)

AND

("primary school*" OR "primary grade*" OR "grades 1 through 3" OR "grades 1 to 3" OR "grades 1-3" OR "first through third" OR "Grade 1" OR first grade* OR "grade 2" OR second grade* OR "grade 3" OR third grade* OR early grade* OR elementary OR kindergarten* OR pre-school* OR preschool* OR pre-kindergarten* OR prekindergarten* OR preK OR pre-K OR "early childhood")

AND

("Latin America*" OR Caribbean OR "South America*" OR Antigua* and Barbuda OR Argentina OR Aruba OR Bahamas OR Barbados OR Belize OR Bermuda OR Bolivia* OR Brazil* OR "British Virgin Islands" OR "Cayman Islands" OR Chile* OR Colombia* OR "Costa Rica*" OR Cuba* OR Curacao OR Dominica* OR "Dominican Republic" OR Ecuador* OR "El Salvador*" OR "French Guiana*" OR Grenada* OR Guadeloupe OR Guatemala* OR Guyana* OR Haiti* OR Honduras OR Jamaica* OR Martinique OR Mexico OR Mont Serrat OR "Netherlands Antilles" OR Nicaragua* OR Panama* OR Paraguay* OR Peru* OR "Puerto Rico" OR "Saint Barthelemy" OR "Saint Kitts and Nevis" OR "Saint Lucia*" OR "Saint-Martin" OR "Saint Vincent and the Grenadines" OR "Sint Maarten" OR Suriname OR "Trinidad and Tobago" OR "Turks and Caicos" OR Uruguay OR "Virgin Islands" OR Venezuela)

Cut down thread, and entered the following with no results:

("Read*" OR Literacy) AND ("primary school*" OR "primary grade*" OR "early grade") AND ("South America*" OR "Latin America*" OR "Central America*" OR Caribbean)

Entered "early childhood reading" with no results

Entered "early grade reading" with no results

Entered "child literacy" with 39 irrelevant results

This journal is useless. Zero relevant results.

Based on not finding results for #3-5, I will not try removing asterisks

Education International

This search engine allows you to choose a region, so I chose Latin America, which yielded 189 results

I unchecked the following for types of resources, which reduced the results to 48:

News

Events

Urgent Action Appeals

I unchecked the following for subject matter, got 18 results that were not relevant:

About EI

Trade & Education

Higher Education & Research

HIV/AIDS

Human & Trade Union Rights

Professional Ethics

Sexual Orientation

Health and Safety in Schools

Solidarity Fund

Migrant Rights

Racism and Xenophobia

Economic Crisis

Congress 7

I selected all options again, and tried:

Entering “reading” and “literacy” but with no results

Entering “early” with 2 irrelevant results

I chose another region, North America-Caribbean, 375 irrelevant results

Entered “reading” and “literacy,” the latter yielding 5 irrelevant results

Entered “early” with 8 irrelevant results

I didn’t find anything useful here. Zero relevant results.

Google Scholar

Couldn’t enter original search string due to character limit

Entered with 311,000 results:

("Read*" OR Literacy) AND ("primary school*" OR "primary grade*" OR "early grade") AND ("South America*" OR "Latin America*" OR "Central America*" OR Caribbean)

Added **date parameters** and got 17,800 results

Removed quotation marks on all except “early grade” and got 18,800 results, mixed relevance

(Read* OR Literacy) AND (primary school* OR primary grade* OR "early grade") AND (South America* OR Latin America* OR Central America* OR Caribbean)

Modified results as follows, for **20,800** results:

("early grade literacy" OR "early grade reading" OR ("early childhood" AND (reading OR literacy)) AND (Latin America OR South America OR Central America OR Caribbean)

The results seem relevant, even after skipping several pages of results.

Tried adding “NOT” to make results more relevant

("early grade literacy" OR "early grade reading" OR ("early childhood" AND (reading OR literacy)) AND (Latin America OR South America OR Central America OR Caribbean) NOT mathematics

I am not making any further edits to this search engine because we decided not to use this search engine.

HAPI

Required subscription that I do not have.

LANIC

Original search string did not yield any results.

Removed all numbered grade references, did not yield any results either.

Removed all country references, since this is a database on Latin America. No results.

Entered results as follows:

(Read OR Literacy) AND ("primary school" OR "primary grade" OR "early grade" OR elementary OR kindergarten OR preschool OR "early childhood")

Mixed results, **results: 208**

Added asterisks as follows:

(Read* OR Literacy) AND ("primary school" OR "primary grade" OR "early grade" OR elementary OR kindergarten* OR preschool* OR "early childhood")

No results. If I leave asterisk only on "read", it yields **187 results**. These results are mixed. It seems that the articles are among the results because a teacher provides a narrative of what they do: "I teach high school students who read English on a primary-grade level" for an article entitled "Animals of Ecuador and Virginia."

(Read* OR Literacy) AND ("primary school" OR "primary grade" OR "early grade" OR elementary OR kindergarten OR preschool OR "early childhood")

Removed quotation marks, got 162 results, but less relevant.

I think that this database is not useful for our purposes.

DEC

This is a Google powered engine, and just as I couldn't enter the original search string in Google, I can't do so here either. So I borrowed from my first string in Google, but modified to remove asterisks from phrases, and quotation marks from single words.

(Read* OR Literacy) AND ("primary school" OR "primary grade" OR "early grade") AND ("South America" OR "Latin America" OR "Central America" OR Caribbean)

Results: 3910

Within the date parameters, there are **2231 results**. However, the results must be filtered in categories. For dates, the results are filtered by decade: 1990-1999 (1028 results), 2000-2009 (860 results), and 2010 or later (343 results)

WorldCat

Entered original search string, with date parameters of 1990-2016, into one field:

("Read" OR Literacy) AND ("primary school" OR "primary grade" OR "grades 1 through 3" OR "grades 1 to 3" OR "grades 1-3" OR "first through third" OR "Grade 1" OR "first grade" OR "grade 2" OR "second grade" OR "grade 3" OR "third grade" OR "early grade" OR elementary OR "kindergarten" OR "pre-school" OR "preschool" OR "pre-kindergarten" OR "prekindergarten" OR preK OR "pre-K" OR "early childhood") AND ("Latin America" OR Caribbean OR "South America" OR "Antigua and Barbuda" OR Argentina OR Aruba OR Bahamas OR Barbados OR Belize OR Bermuda OR "Bolivia" OR "Brazil" OR "British Virgin Islands" OR "Cayman Islands" OR "Chile" OR "Colombia" OR "Costa Rica" OR "Cuba" OR Curacao OR "Dominica" OR "Dominican Republic" OR "Ecuador" OR "El Salvador" OR "French Guiana" OR "Grenada" OR Guadeloupe OR "Guatemala" OR "Guyana" OR "Haiti" OR Honduras OR "Jamaica" OR Martinique OR Mexico OR Mont Serrat OR "Netherlands Antilles" OR "Nicaragua" OR "Panama" OR "Paraguay" OR "Peru" OR "Puerto Rico" OR "Saint Barthelemy" OR "Saint Kitts and Nevis" OR "Saint Lucia" OR "Saint-Martin" OR "Saint Vincent and the Grenadines" OR "Sint Maarten" OR Suriname OR "Trinidad and Tobago" OR "Turks and Caicos" OR Uruguay OR "Virgin Islands" OR Venezuela)

Yielded system error, so I divided the string by three:

("Read" OR Literacy)

AND

("primary school" OR "primary grade" OR "grades 1 through 3" OR "grades 1 to 3" OR "grades 1-3" OR "first through third" OR "Grade 1" OR "first grade" OR "grade 2" OR "second grade" OR "grade 3" OR "third grade" OR "early grade" OR elementary OR "kindergarten" OR "pre-school" OR "preschool" OR "pre-kindergarten" OR "prekindergarten" OR preK OR "pre-K" OR "early childhood")

AND

("Latin America" OR Caribbean OR "South America" OR "Antigua and Barbuda" OR Argentina OR Aruba OR Bahamas OR Barbados OR Belize OR Bermuda OR "Bolivia" OR "Brazil" OR "British Virgin Islands" OR "Cayman Islands" OR "Chile" OR "Colombia" OR "Costa Rica" OR "Cuba" OR Curacao OR "Dominica" OR "Dominican Republic" OR "Ecuador" OR "El Salvador" OR "French Guiana" OR "Grenada" OR Guadeloupe OR "Guatemala" OR "Guyana" OR "Haiti" OR Honduras OR "Jamaica" OR Martinique OR Mexico OR Mont Serrat OR "Netherlands Antilles" OR "Nicaragua" OR "Panama" OR "Paraguay" OR "Peru" OR "Puerto Rico" OR "Saint Barthelemy" OR "Saint Kitts and Nevis" OR "Saint Lucia" OR "Saint-Martin" OR "Saint Vincent and the Grenadines" OR "Sint Maarten" OR Suriname OR "Trinidad and Tobago" OR "Turks and Caicos" OR Uruguay OR "Virgin Islands" OR Venezuela)

System error. I removed the parentheses, got 64,126 hits. Even though the "help" section talks about using parentheses to create more precise searches, I get error responses (http://www.oclc.org/support/help/navpatron/ApplicationHelp.htm).

Added date parameters for 1990-2016, got 38,300 hits. After looking through the results, I did not find relevant results.

Modified search string to remove all the "grade 1" "grade 2" and "grade 3" references, got the same number of results

Removed all references to grade (i.e., kindergarten, preschool), still got the same number of results.

Removed all country references, to focus on regional, and got 11,213 results:

Read* OR Literacy

AND

"primary school" OR "primary grade" OR "early grade" OR "early childhood"

AND

"Latin America" OR Caribbean OR "South America" OR "Central America"

Tried further filtering topics by clicking on the topic of "Education" and got 875 results, but the results included titles clearly irrelevant to the subject, i.e., "Examining the impacts of dynamic downscaling method and vegetation biophysical processes on the South American regional climate simulation" and some that were tangentially relevant, "The experiences of African, Caribbean and south Asian women in initial teacher education"

Tried entering data into the fields, one by one, got 15,559 irrelevant results:

kw:Read* OR kw:Literacy AND kw:primary school OR kw:primary grade OR kw:early grade OR kw:elementary OR kw:*kindergarten* OR kw:preschool* OR kw:prek OR kw:early childhood AND kw:Latin America* OR kw:Central America* OR kw:South America* OR kw:Caribbean AND yr:1990..2016

Filtered by education, got 1,408 results with irrelevant results, for example: "The determinants of remittances: Latin America and the Caribbean, 1982-2001" and "Taxonomy of larval blennioidei of Belize, Central America". If I go through the results, I find a few potentially relevant results (although it isn't immediately clear upon reading the title).

Backtracked to step 2 (without using parentheses) and removed quotations from single words such as "Venezuela" and "Brazil," and still got irrelevant results (3,542 results).

Filtered out by "Education," irrelevant results (194 results). Removed the following fields: "Individual Institutions," "Higher Education," and "Individual Institutions – America – Except U.S." got 37 results.

Backtracked again to step 2 (without using parentheses), and tried the same entry WITH quotations, and repeated step 12, with mixed results.

Backtracked to step 7, removed all quotation marks, and filtered according to step 12 (98 results). Even though the results are related to Education, they are not specifically related to early grade reading.

Backtracked again to step 1, and removed parentheses and quotation marks on one-word entries. Date parameters set to 1990-2016. Entered as follows:

Read* OR Literacy

AND

"primary school" OR "primary grade" OR "grades 1 through 3" OR "grades 1 to 3" OR "grades 1-3" OR "first through third" OR "Grade 1" OR "first grade" OR "grade 2" OR "second grade" OR "grade 3" OR "third grade" OR "early grade" OR elementary OR kindergarten OR "pre-school" OR "preschool" OR "pre-kindergarten" OR "prekindergarten" OR preK OR "pre-K" OR "early childhood"

AND

"Latin America" OR Caribbean OR "South America" OR "Antigua and Barbuda" OR Argentina OR Aruba OR Bahamas OR Barbados OR Belize OR Bermuda OR Bolivia OR Brazil OR "British Virgin Islands" OR "Cayman Islands" OR Chile OR Colombia OR "Costa Rica" OR Cuba OR Curacao OR Dominica OR "Dominican Republic" OR Ecuador OR "El Salvador" OR "French Guiana" OR Grenada OR Guadeloupe OR Guatemala OR Guyana OR Haiti OR Honduras OR Jamaica OR Martinique OR Mexico OR Mont Serrat OR "Netherlands Antilles" OR Nicaragua OR Panama OR Paraguay OR Peru OR "Puerto Rico" OR "Saint Barthelemy" OR "Saint Kitts and Nevis" OR "Saint Lucia" OR "Saint-Martin" OR "Saint Vincent and the Grenadines" OR "Sint Maarten" OR Suriname OR "Trinidad and Tobago" OR "Turks and Caicos" OR Uruguay OR "Virgin Islands" OR Venezuela

Got 38,312 results, and added parentheses in the search box as follows:

(kw:Read* OR Literacy) AND (kw:"primary school" OR "primary grade" OR "grades 1 through 3" OR "grades 1 to 3" OR "grades 1-3" OR "first through third" OR "Grade 1" OR "first grade" OR "grade 2" OR "second grade" OR "grade 3" OR "third grade" OR "early grade" OR elementary OR kindergarten OR "pre-school" OR "preschool" OR "pre-kindergarten" OR "prekindergarten" OR preK OR "pre-K" OR "early childhood") AND (kw:"Latin America" OR Caribbean OR "South America" OR "Antigua and Barbuda" OR Argentina OR Aruba OR Bahamas OR Barbados OR Belize OR Bermuda OR Bolivia OR Brazil OR "British Virgin Islands" OR "Cayman Islands" OR Chile OR Colombia OR "Costa Rica" OR Cuba OR Curacao OR Dominica OR "Dominican Republic" OR Ecuador OR "El Salvador" OR "French Guiana" OR Grenada OR Guadeloupe OR Guatemala OR Guyana OR Haiti OR Honduras OR Jamaica OR Martinique OR Mexico OR Mont Serrat OR "Netherlands Antilles" OR Nicaragua OR Panama OR Paraguay OR Peru OR "Puerto Rico" OR "Saint Barthelemy" OR "Saint Kitts and Nevis" OR "Saint Lucia" OR "Saint-Martin" OR "Saint Vincent and the Grenadines" OR "Sint Maarten" OR Suriname OR "Trinidad and Tobago" OR "Turks and Caicos" OR Uruguay OR "Virgin Islands" OR Venezuela) AND yr:1990..2016

Hit search again, and got 102 results, relevant.

Appendix C. Quantitative Risk of Bias Assessment Tool and Risk of Bias Assessment for Included Quantitative Intervention Studies

Code description	Code	Comment
Study ID	Last name of author, year	Open answer
Justification of use	Study design and methodology	Open Answer
Ask these questions for all quantitative studies		
Did the outcome measure include some measure of reading or a reading sub-skill (e.g., fluency, phonological awareness, language, decoding, letter knowledge, comprehensions etc.)?	Yes No Unclear Not applicable	Comment: Open answer
If the study did not include a measurement of reading or a reading sub-skill, is literacy measured in a different manner?		
Does the study show baseline reading/literacy abilities for beneficiaries and non-beneficiaries?		
If reading/literacy scores are not available at baseline, does the study show characteristics of beneficiaries and non-beneficiaries that are not likely to be affected by the intervention?		
Are the mean values or the distributions of the covariates at baseline statistically different for beneficiaries and non-beneficiaries ($p<0.05$)		
If there are statistically significant differences between beneficiaries and non-beneficiaries are these differences controlled for using covariate analysis in the impact evaluation?		
If baseline characteristics are not available, does the study qualitatively assess why beneficiaries are likely/unlikely to be a random draw of the population at baseline?		
Confounding and selection bias (ask questions for all quantitative studies)		
Does the study use a comparison/control group of students/households without access to the program?	Yes No Unclear Not applicable	Comment: Open answer
Does the study use a comparison/control group of students/households with access to the program but that did not choose to participate in the program?		
Does the study include data at baseline and endline (before and after the intervention)?		

Code description	Code	Comment
Are the data on covariates collected at the baseline?		
Is difference in differences estimation used?		
If the study is quasi-experimental and uses difference-in-difference estimation do the authors assess the parallel trends assumption?		
If the study does not use difference in difference, does the study control for baseline values of the outcome of interest		
If the study does not use difference in difference and does not control for baseline values of the outcome variable, does the study control for other covariates at baseline		
If the study does not use difference in differences estimation, is there any assessment of likely risk of bias from time invariant characteristics driving both participation and outcome?		
If the study does not use difference in difference estimation but does assess likely risk of bias from time invariant characteristics, are these time invariant characteristics likely to bias the impact estimates		
Does the study report the table with the results of the outcome equation (including covariates)? Where full results of the outcome equation are not reported, is it clear which covariates have been used?		
Are all relevant observable covariates (confounding variables) included in the outcome equation which might explain outcomes, if estimation does not use a statistical technique to control for selection bias (RCT, PSM or covariate matching, IV or switching regression)? This might, for example, include control for ability, and/or social capital.		
Attrition (ask questions for all quantitative studies)		
For studies including baseline data, does the study report attrition (drop-out) from the study?	Yes No Unclear Not applicable	Comment: Open answer
Is the attrition rate below 10%?		
Does the study assess whether drop-outs are random draws from the sample (e.g., by examining correlation with determinants of outcomes, in both treatment comparison group)?		

Code description	Code	Comment
Spillovers and contamination (ask questions for all quantitative studies)		
Spillovers: are comparisons sufficiently isolated from the intervention (e.g., participants and non-participants are sufficiently geographically or socially separated) or are spillovers estimated by comparing non-beneficiaries with access to the intervention to non-beneficiaries without access to the intervention and/or through social network analysis?	Yes No Unclear Not applicable	Comment: Open answer
Spillovers; if spillovers are not estimated, is the study likely to bias the impact of the program?		
Contamination: does the study assess whether the control group receives the intervention?		
Contamination: if the control group receives the intervention but for a shorter amount of time does the study assess the likelihood that the control group has received equal benefits as the treatment group		
Contamination: if the control group receives the intervention have they received the intervention sufficiently long to argue that they have benefited from the intervention		
Contamination: does the study describe and control for other interventions which might explain changes in outcomes?		
Other threats to validity (ask questions for all quantitative studies)		
Does the evidence suggest analysis reporting biases are a serious concern? Analysis reporting biases include failure to report important treatment effects (possibly relating to intermediate outcomes), or justification for (uncommon) estimation methods, especially multivariate analysis for outcomes equations.	1 = Yes 2 = No 9 = Unclear 99 = Not applicable	Comment: Open answer
Are there concerns about baseline data collected retrospectively		
Are there concerns about courtesy bias from outcomes collected through self-reporting?		
Construct Validity (ask questions for all quantitative studies)		
Were reading outcomes measured in the majority of the appropriate languages?	1 = Yes 2 = No 9 = Unclear 99 = Not applicable	Comment: Open answer
Does the study describe the implementation of the program in sufficient detail?		
Was the unit of allocation and the unit of analysis the same?		
Do all students targeted by the study take the reading test/answer the survey questions?		

Code description	Code	Comment
Does the study take into consideration potential implementation failures		
Does the study use a proper theory of change, logframe and/or other proper conceptual or theoretical framework?		
Does the study analyze the outcome measures put forward in the theory of change or logframe?		
External Validity (ask questions for all quantitative studies)		
Do the authors clearly distinguish between the intention-to-treat effect and the treatment effect on the treated?	1 = Yes 2 = No 9 = Unclear 99 = Not applicable	Comment: Open answer
Do the authors highlight the intention-to-treat effect?		
Hawthorne and John Hendry Effects (ask questions for all quantitative studies)		
Do the authors argue convincingly that it is not likely that being monitored influences the behavior of the beneficiaries and non-beneficiaries in different ways?	1 = Yes 2 = No 9 = Unclear 99 = Not applicable	Comment: Open answer
Confidence Intervals (ask questions for all quantitative studies)		
Does the study account for lack of independence between observations within assignment clusters if the outcome variables are clustered?	1 = Yes 2 = No 9 = Unclear 99 = Not applicable	Comment: Open answer
Is the sample size likely to be sufficient to find significant effects of the intervention?		
Do the authors control for heteroskedasticity and/or use robust standard errors?		
Ask questions below only for studies that apply randomization		
Does the study apply randomized assignment?	1 = Yes 2 = No 9 = Unclear 99 = Not applicable	Comment: Open answer
Does the study use a unit of allocation with a sufficiently large sample size to ensure equivalence between the treatment and the control group?		

Code description	Code	Comment
Ask questions below only for studies that apply regression discontinuity designs		
Is the allocation of the program based on a predetermined continuity on a continuous variable and blinded to the beneficiaries or if not blinded, individuals cannot reasonably affect the assignment variable in response to knowledge of the participation rule?	1 = Yes 2 = No 9 = Unclear 99 = Not applicable	Comment: Open answer
Is the sample size immediately at both sides of the cut-off point sufficiently large to equate groups on average?		
Is the mean of the covariates of individuals immediately at both sides of the cut-off point statistically significantly different for beneficiaries and non-beneficiaries?		
If there are statistically significant differences between beneficiaries and non-beneficiaries are these differences controlled for using covariate analysis?		
Ask questions below only for studies that apply matching		
Quality of matching (PSM, covariate matching)		
Are beneficiaries and non-beneficiaries matched on all relevant characteristics?	1 = Yes 2 = No 9 = Unclear 99 = Not applicable	Comment: Open answer
Does the study report the results of the matching function (e.g., for PSM the logit function)?		
Does the study report the matching method?		
Does the study exclude observations outside the common support?		
Does the study use variables at follow-up that can be affected by the intervention in the matching equation?		
Are matches found for the majority of participants (>90%)?		
If >=10% of participants failed to be matched, is sensitivity analysis used to re-estimate results using different matching methods?		
For nearest-neighbor PSM, does the study report the mean or distribution of the propensity scores in the treatment and control groups after matching?		
For nearest-neighbor PSM, are propensity scores similar, based on tests for statistical differences at the means or other quantiles of the distribution)?		
Does the study report the mean or distribution for the covariates of the treatment and control groups after matching?		

Code description	Code	Comment
Are these characteristics similar, based on tests for statistically significant differences (p>0.5)?		
Sensitivity analysis (only for studies that apply PSM)		
For PSM, where propensity score distributions and/or covariates of the treatment and control groups are not reported, or they are reported but there are differences in means or distributions of the covariates or propensity scores (usually only applicable to methods which do not exclude treatment observations such as nearest neighbor), is robustness assessed using an additional matching technique?	1 = Yes 2 = No 9 = Unclear 99 = Not applicable	Comment: Open answer
Is sensitivity to hidden bias assessed statistically, e.g., using the Rosenbaum bounds test?		
Ask questions below only for studies that apply instrumental variable estimation		
Quality of IV, two-steps endogenous switching regression approach		
Does the study describe clearly the instrumental variable(s)/identifier used?	1 = Yes 2 = No 9 = Unclear 99 = Not applicable	Comment: Open answer
Are the results of the participation equation reported?		
Are the instruments jointly significant at the level of F ≥ 10? If an F test is not reported, does the author report and assess whether the R-squared of the instrumenting equation is large enough for appropriate identification (R-sq > 0.5?)		
Are the instruments individually significant (p≤0.05)?		
For IV, If more than one instrument is used in the procedure, does the study include and report an overidentifying test (p≤0.05 is required to reject the null hypothesis)?		
Does the study qualitatively assess the exogeneity of the instrument/identifier (both externality as well as why the variable should not enter by itself in the outcome equation)?		
Ask questions below only for studies with censored outcome variables		
Do the authors use appropriate methods (e.g., Heckman selection models, tobit models, duration models) to account for the censoring of the data?	1 = Yes 2 = No 9 = Unclear 99 = Not applicable	Comment: Open answer
For Heckman models; is there is a variable that is statistically significant in the first stage of the selection equation and excluded from the second stage		

Code description	Code	Comment
Overall Assessment		
Assessment Selection Bias	Low risk of bias Medium risk of bias High risk of bias Unclear risk of bias	Comment: Open answer
Assessment Spillovers and Contamination Bias		
Assessment Outcome and Analysis Reporting Bias		
Assessment Other biases		

Risk of bias assessment for included quantitative intervention studies

	Selection Bias and Confounding	Performance Bias: Assessment, Spillovers, and Contamination	Outcome and Analysis Reporting Biases	Other Biases
Adrogue & Orlicki (2013)	**High risk of selection bias and confounding** This study uses a nonexperimental difference-in-difference design to determine the effect of the program on school-level early grade reading outcomes. We rated the study as high risk of selection bias because the study considers outcome variables 4 months after the start of the study as baseline values. However, these values may well have been affected by the program at that time. In addition, it is unclear whether the comparison group was similar to the beneficiary schools at the time of the start of the intervention.	**Medium Risk of Performance Bias** The authors do not account for the possibility of crossover effects. Students in the comparison schools may have switched to treatment schools because of the school feeding program. This behavioral change may result in spillovers to the comparison group.	**High risk of outcome and analysis reporting bias** The authors use an unusual difference-in-difference approach in which outcome measures after the start of the intervention are used as baseline values. This approach can result in considerable bias.	**Low risk of other biases.** There is no evidence for significant other risks of bias.
	Medium risk of selection bias and confounding	**Low risk of performance bias**	**Low risk of outcome and analysis reporting bias.**	**Low risk of other biases.**

	Selection Bias and Confounding	Performance Bias: Assessment, Spillovers, and Contamination	Outcome and Analysis Reporting Biases	Other Biases
Bando (2010)	This study uses a regression analysis that includes school and state-year fixed effects to determine the effect of a school governance program on early grade reading outcomes. Although this method does not fully account for the risk of selection bias, the risk of selection bias is only medium.	The analysis compares beneficiary schools with comparison schools that appear to be sufficiently isolated from the beneficiary schools.	There are no significant outcome and analysis reporting biases. The study uses a number of robustness checks to assess the validity of the results.	There is no evidence for significant other risks of bias.
	Low risk of selection bias and confounding	**Low risk of performance bias**	**Low risk of outcome and analysis reporting bias**	**Low risk of other biases.**
Barrera-Osorio & Linden (2009)	This study uses a cluster-randomized controlled trial to determine the impact of the distribution of computers on early grade reading outcomes. Although attrition was high, the authors were able to credibly account for this in the analysis. Thus the risk of selection bias and confounding was low.	The analysis compares beneficiary schools with comparison schools that appear to be sufficiently isolated from the beneficiary schools to prevent performance bias.	There are no significant outcome and analysis reporting biases.	There is no evidence for significant other risks of bias.
	Low risk of selection bias and confounding.	**Low risk of performance bias**	**Low risk of outcome and analysis reporting bias**	**Low risk of other biases**

	Selection Bias and Confounding	Performance Bias: Assessment, Spillovers, and Contamination	Outcome and Analysis Reporting Biases	Other Biases
Beuermann et al. (2015)	**Low risk of selection bias**	**Low risk of performance bias**	**Low risk of outcome and analysis reporting bias**	**Low risk of other biases**
	The study uses a cluster-randomized controlled trial to determine the impact of the distribution of laptops to children on early grade reading outcomes. There are no major concerns about selection bias.	The study uses a credible social network analysis to determine the spillover effects of the program.	There are no significant outcome and analysis reporting biases.	There is no evidence for significant other risks of bias.
Campos et al. (2011)	**High risk of selection bias and confounding**	**Low risk of performance bias**	**High risk of outcome and analysis reporting bias**	**Medium risk of other biases**
	The study uses hierarchical regression analysis to determine the impact of participation in preschool on early grade reading outcomes. This methodology is not a well-established method to account for selection bias without a clear identification strategy.	The analysis compares beneficiary schools with comparison schools that appear to be sufficiently isolated from the beneficiary schools.	The study only reports statistically significant effects in the tables. However, the narrative suggests that not all results were statistically significant. This is an indication for outcome and analysis reporting bias.	The study does not account for clustering in the estimation of the standard errors.

	Selection Bias and Confounding	Performance Bias: Assessment, Spillovers, and Contamination	Outcome and Analysis Reporting Biases	Other Biases
Cardoso-Martins et al. (2011) Experiment 1	**Medium risk of selection bias**	**High risk of performance bias**	**Low risk of outcome and analysis reporting bias**	**Low risk of other biases**
	The study uses a randomized controlled trial to determine the impact of the program on early grade reading outcomes. However, the sample only consisted of 32 students. This sample size is not sufficient to ensure equivalence in observable and unobservable characteristics.	The study used randomization at the student-level within the same school. There is thus a lot of interaction between beneficiary and control students. This interaction creates a major risk of performance bias.	There are no significant outcome and analysis reporting biases.	There is no evidence for significant other risks of bias.
Cardoso-Martins et al. (2011) Experiment 2	**Medium risk of selection bias**	**High risk of performance bias**	**Low risk of outcome and analysis reporting bias**	**Low risk of other biases**
	The study uses a randomized controlled trial to determine the impact of the program on early grade reading outcomes. However, the sample only consisted of 20 students. This sample size is not sufficient to ensure equivalence in observable and unobservable characteristics.	The study used randomization at the student-level within the same school. There is thus a lot of interaction between beneficiary and control students. This interaction creates a major risk of performance bias.	There are no significant outcome and analysis reporting biases.	There is no evidence for significant other risks of bias.

	Selection Bias and Confounding	Performance Bias: Assessment, Spillovers, and Contamination	Outcome and Analysis Reporting Biases	Other Biases
Cristia et al. (2012)	**Low risk of selection bias**	**Low risk of performance bias**	**Low risk of outcome and analysis reporting bias**	**Low risk of other biases**
	The study uses a cluster-randomized controlled trial to determine the impact of the distribution of laptops to children on early grade reading outcomes. There are no major concerns about selection bias.	The analysis compares beneficiary schools with control schools that appear to be sufficiently isolated from the beneficiary schools.	There are no significant outcome and analysis reporting biases.	There is no evidence for significant other risks of bias.
De Felicio, Terra, and Zoghbi (2012)	**Medium risk of selection bias**	**Medium risk of performance bias**	**High risk of outcome and analysis reporting bias**	**Low risk of other biases**
	The study uses a propensity score matching design to assess the effects of participation in preschool on early grade reading outcomes. This design enables the researchers to correct for selection-bias from observable characteristics. However, the selection bias is still medium because the methodology does not allow the researchers to account for unobservable characteristics from self-selection into preschool.	The study compares beneficiaries with non-beneficiaries in the same municipality. Thus, there is a risk of interaction between beneficiary and comparison students, which we interpret as a medium risk of performance bias.	The authors fail to report statistically insignificant effects. However, the narrative indicates that the results are not statistically significant in all specifications. This discrepancy in reporting indicates a high risk of outcome and analysis reporting bias.	There is no evidence for significant other risks of bias.

	Selection Bias and Confounding	Performance Bias: Assessment, Spillovers, and Contamination	Outcome and Analysis Reporting Biases	Other Biases
Ferrando et al. (2011)	**Medium risk of selection bias**	**Medium risk of performance bias**	**Medium risk of outcome and analysis reporting bias**	**Medium risk of other biases**
	The study uses a propensity score matching design to assess the effects of the distribution of laptops to children on early grade reading outcomes. This design enables the researchers to correct for selection-bias from observable characteristics. However, the selection bias is still medium because the methodology does not allow the researchers to account for unobservable characteristics.	The analysis compares beneficiary schools with comparison schools that appear to be sufficiently isolated from the beneficiary schools.	The study only uses a subset of available control variables for the propensity score matching. It is unclear why the authors do not include the other potential control variables. This approach may be an indication for outcome and analysis reporting bias	The study does not account for clustering in the estimation of the standard errors.
Gomez Franco (2014)	**High Risk of Selection Bias** The study uses a cluster-randomized controlled trial to determine the impact of the program on early grade reading outcomes. However, the study analyses data for beneficiaries that comply with the instructions during the training. This non-random sample significantly increases the risk of selection bias. In addition, the authors use several potentially endogenous characteristics as control variables.	**Low Risk of Performance Bias** The analysis compares beneficiary schools with control schools that appear to be sufficiently isolated from the beneficiary schools.	**High Risk of Outcome and analysis Reporting Bias** The study uses several potentially endogenous characteristics in the estimation of the impact of the program. This approach is an indication for outcome and analysis reporting bias.	**High Risk of Other Biases** The study does not account for clustering in the estimation of the standard errors.

	Selection Bias and Confounding	Performance Bias: Assessment, Spillovers, and Contamination	Outcome and Analysis Reporting Biases	Other Biases
Ismail et al. (2014)	**Medium Risk of Selection Bias** The study uses a propensity score matching design to assess the effects of a school feeding program on early grade reading outcomes. This design enables the researchers to correct for selection-bias from observable characteristics. However, the selection bias is still medium because the methodology does not allow the researchers to account for unobservable characteristics.	**Medium Risk of Performance Bias** The analysis suggests that comparison schools may not be sufficiently isolated from the beneficiary schools. Thus, there is a medium risk of performance bias.	**Low Risk of Outcome and Analysis Reporting Bias** There are no significant outcome and analysis reporting biases.	**Low Risk of Other Biases** There is no evidence for significant other risks of bias.
Larrain et al. (2012) Study 1	**High Risk of Selection Bias** The study randomly assigns two classrooms to the treatment group and two classrooms to the control group. This sample size is too small to ensure equivalence in observable and unobservable characteristics. In addition, the authors do not present evidence for equivalence in observable characteristics. Balance tables are not reported.	**Medium Risk of Performance Bias** The program is randomly assigned at the classroom level within the same school. Thus, there may be interaction between beneficiary students and control students, which may result in spillovers.	**Outcome and Analysis Reporting Bias** There are no significant outcome and analysis reporting biases.	**Low Risk of Other Biases** There is no evidence for significant other risks of bias.

	Selection Bias and Confounding	Performance Bias: Assessment, Spillovers, and Contamination	Outcome and Analysis Reporting Biases	Other Biases
Larrain et al. (2012) Study 2	**High Risk of Selection Bias** The study randomly assigns two classrooms to the treatment group and two classrooms to the control group. This sample size is too small to ensure equivalence in observable and unobservable characteristics. In addition, the authors do not present evidence for equivalence in observable characteristics. Balance tables are not reported.	**Medium Risk of Performance Bias** The program is randomly assigned at the classroom level within the same school. Thus, there may be interaction between beneficiary students and control students, which may result in spillovers.	**Low Risk of Outcome and Analysis Reporting Bias** There are no significant outcome and analysis reporting biases.	**Low Risk of Other Biases** There is no evidence for significant other risks of bias.
Lockheed, Harris, & Jayasundera (2010)	**Medium Risk of Selection Bias** The study uses a propensity score matching design to determine the impact of a school governance program on early grade reading outcomes. This methodology enables the researchers to control for observable characteristics. However, the risk of selection bias remains medium because the design does not allow for controlling for unobservable characteristics.	**High Risk of Performance Bias** The study reports that the comparison schools also often received the program but does not account for this in the analysis.	**Low Risk of Outcome and Analysis Reporting Bias** There are no significant outcome and analysis reporting biases	**Medium Risk of Other Biases** The study does not account for clustering in the estimation of the standard errors.

	Selection Bias and Confounding	Performance Bias: Assessment, Spillovers, and Contamination	Outcome and Analysis Reporting Biases	Other Biases
Maluccio et al. (2009)	**Medium Risk of Selection Bias** The study uses a cluster-randomized controlled to determine the impact of a nutrition program on early grade reading outcomes. However, the sample size is very small and does not ensure equivalence in observable characteristics between treatment and control villages. The authors account for this concern by showing descriptive statistics, but there is nonetheless a medium risk of selection bias.	**Low Risk of Performance Bias** The study uses village-level randomization to determine the impact of the program on early grade reading outcomes. The villages appear to be sufficiently isolated, which limits the potential for bias from spillovers or contamination.	**Low Risk of Outcome and Analysis Reporting Bias** There are no significant outcome and analysis reporting biases	**Low Risk of Other Biases** There is no evidence for significant other risks of bias.
Mendive et al. (2016)	**High Risk of Selection Bias** The study uses hierarchical least squares regression analysis to determine the effect of the compliance with a teacher training program on early grade reading outcomes. Compliance is determined by self-selection. The use of regression analysis does not enable controlling for this self-selection. Thus, the risk of selection-bias is high.	**Low Risk of Performance Bias** The program is randomly assigned as the school-level. This approach limits the interaction between beneficiary and control students, which reduces the risk of bias from spillovers or contamination.	**High Risk of Outcome and Analysis Reporting Bias** The authors use arbitrary thresholds for determining whether the program was implemented with sufficient adherence and dosage. In addition, the authors use OLS regression analysis as opposed to instrumental variable regression analysis. The authors should have used the randomization as an instrument for compliance in order to appropriately estimate the impact of compliance with the program.	**Medium Risk of Other Biases** The use of videos to measure teacher behavior could have resulted in Hawthorne Effects, which could bias the impact of the program.

	Selection Bias and Confounding	Performance Bias: Assessment, Spillovers, and Contamination	Outcome and Analysis Reporting Biases	Other Biases
Pallante & Kim (2013)	**Low Risk of selection Bias** The study uses a cluster-randomized controlled trial to determine the impact of the program on early grade reading outcomes. The authors also present evidence for balance in observable characteristics across treatment and control conditions. Thus, the risk of selection bias is low.	**Medium Risk of Performance Bias** The study used random assignment at the classroom level. Thus, there may have been interaction between beneficiary and control students. This interaction results in a medium risk of bias from spillovers or contamination.	**Low Risk of Outcome and Analysis Reporting Bias** There are no significant outcome and analysis reporting biases	**Low Risk of Other Biases** There is no evidence for significant other risks of bias.
Powell et al. (1998)	**Low Risk of selection Bias** The study uses student-level randomization to determine the impact of the program on early grade reading outcomes. The study also shows evidence for balance in observable characteristics. Thus, we consider this study as low risk of selection bias.	**High Risk of Performance Bias** The study compares beneficiary and control students within the same classroom. This approach significantly increases the risk of spillovers and contamination.	**High Risk of Outcome and Analysis Reporting Bias** The study reports statistically significant effects. However, our effect size calculations suggest that the results are not statistically significant.	**Low Risk of Other Biases** There is no evidence for significant other risks of bias.
Simeon, Grantham-McGregor, & Wong (1995)	**Low Risk of selection Bias** The study uses student-level randomization to determine the impact of the program on early grade reading outcomes. The study also shows evidence for balance in observable characteristics. Thus, we consider this study as low risk of selection bias.	**High Risk of Performance Bias** The study compares beneficiary and control students within the same classroom. This approach significantly increases the risk of spillovers and contamination.	**Low Risk of Outcome and Analysis Reporting Bias** There are no significant outcome and analysis reporting biases	**Low Risk of Other Biases** There is no evidence for significant other risks of bias.

	Selection Bias and Confounding	Performance Bias: Assessment, Spillovers, and Contamination	Outcome and Analysis Reporting Biases	Other Biases
Tapia & Benitez (2013)	**High Risk of Selection Bias** The study uses random assignment with a sample of 10 treatment and 10 control students. This sample size is too low to ensure equivalence in observable characteristics between treatment and control students.	**High Risk of Performance Bias** The study compares beneficiary and control students within the same school. This approach significantly increases the risk of spillovers and contamination.	**Medium Risk of Outcome and Analysis Reporting Bias** The study reports statistically significant effects with a very small sample size However, the authors do not report the results of the outcome equation. Instead, the results are presented through graphs. Thus, we consider this study as medium risk of outcome and analysis reporting bias.	**Medium Risk of Other Biases** The study suggests that researchers were heavily involved in the data collection, which may have resulted in Hawthorne Effects.
Murad & Topping (2000)	**High Risk of Selection Bias** The study uses random assignment but the sample size is too small to ensure equivalence in observable characteristics. In addition, treatment students were switched to the control group because they could not comply with the intervention. Together, these constraints result in a high risk of selection bias.	**High Risk of Performance Bias** The study compares beneficiary and control students within the same school. This approach significantly increases the risk of spillovers and contamination.	**Medium Risk of Outcome and Analysis Reporting Bias** The authors exclude outliers from their analysis for unclear reasons. The exclusion of these outliers may have affected the statistical significance of the impact estimates.	**Medium Risk of Other Biases** The study suggests that researchers were heavily involved in the data collection, which may have resulted in Hawthorne Effects.

	Selection Bias and Confounding	Performance Bias: Assessment, Spillovers, and Contamination	Outcome and Analysis Reporting Biases	Other Biases
Vivas (1996) Experiment 1	**Medium Risk of Selection Bias** The study uses random assignment but the sample size is too small to ensure equivalence in observable characteristics.	**Low Risk of Performance Bias** The program is randomly assigned as the school-level. This approach limits the interaction between beneficiary and control students, which reduces the risk of bias from spillovers or contamination.	**Medium Risk of Outcome and Analysis Reporting Bias** The study estimates the impact of the program by comparing the median value of early grade reading outcomes between beneficiary and control students. This approach is unusual and may be an indication for outcome and analysis reporting bias.	**Low Risk of Other Biases** There is no evidence for significant other risks of bias.
Vivas (1996) Experiment 2	**Medium Risk of Selection Bias** The study uses random assignment but the sample size is too small to ensure equivalence in observable characteristics.	**Low Risk of Performance Bias** The program is randomly assigned as the school-level. This approach limits the interaction between beneficiary and control students, which reduces the risk of bias from spillovers or contamination.	**Medium Risk of Outcome and Analysis Reporting Bias** The study estimates the impact of the program by comparing the median value of early grade reading outcomes between beneficiary and control students. This approach is unusual and may be an indication for outcome and analysis reporting bias.	**Low Risk of Other Biases** There is no evidence for significant other risks of bias.

	Selection Bias and Confounding	Performance Bias: Assessment, Spillovers, and Contamination	Outcome and Analysis Reporting Biases	Other Biases
Yoshikawa et al. (2015)	**Low Risk of Selection Bias** The study uses a cluster-randomized controlled trial to determine the impact of the program on early grade reading outcomes. The authors also present evidence for balance in observable characteristics across treatment and control conditions. Thus, the risk of selection bias is low.	**Low Risk of Performance Bias** The program is randomly assigned as the school-level. This approach limits the interaction between beneficiary and control students, which reduces the risk of bias from spillovers or contamination.	**Low Risk of Outcome and Analysis Reporting Bias** There are no significant outcome and analysis reporting biases	**Medium Risk of Other Biases** The use of videos to measure teacher behavior could have resulted in Hawthorne Effects, which could bias the impact of the program.

Appendix D. Quantitative Nonintervention Quality Review Protocol

Quantitative Nonintervention Protocol		
Study Title:	**Reviewer Name:**	
Quality Review Questions	**Yes/No/ unknown/NA**	**Justification for your answer. Please include text from the article as support when possible (include pg. #'s where appropriate)**
Did the outcome measure include some measure of reading or a reading sub-skill (e.g., fluency, phonological awareness, language, decoding, letter knowledge, comprehensions etc.)?		
If the study did not include a measurement of reading or a reading sub-skill, is literacy measured in a different manner?		
Is the sample selection criteria/justification provided?		
Is there data reported on covariates?		
Is there information on training test administrators?		
Are outcomes collected through self-reports?		
How was language of reading data collection determined?		
Did the study report data collection procedures (quiet room, during school hours, possible fatigue effects etc.)		
Was the unit of allocation and the unit of analysis the same?		
Do all students targeted by the study take the reading test/answer the survey questions?		
Does the study take into consideration potential data collection implementation failures?		
Does the study have a strong conceptual or theoretical framework?		
Do the authors generalize only to the reading outcome, and population applicable from the sample		
Do the authors argue convincingly that it is not likely that being monitored influences the behavior of study participants?		
Are there appropriate reliability scores for all tests?		

Does the study describe the analysis method?		
Does the study justify the analysis method (is the analysis method appropriate for the research question/objective)		
Were any participants not included in the analysis? If so, is there justification for why?		
Summarize the main findings of this article in regards to how it might affect our main stakeholder groups (policy makers, Intl NGOs, Teacher training institutes, researchers, etc.)		
Please list other potentially relevant references that should be checked		
1.)		
2.)		
3.)		
4.)		
5.)		

Appendix E. Qualitative Intervention and Nonintervention Quality Review Protocol

Directions: Please list the title of the article and your name as reviewer in the appropriate rows. After reading the article, please rate each criteria as either high, medium, low or unclear by placing an "X" in the appropriate box. For any of the quality criteria that do not apply to the research in question, please place an "X" under the NA column. If you are unable to rate a particular criteria for low, medium or high levels of evidence because none is provided, then please place an "X" in the Not mentioned column. Whenever possible, provide the justification for your choices in the final column listing both strengths and weaknesses and supplying quotes from the article with page numbers.

High = Level of evidence provided is strong
Medium = Level of evidence provided is adequate but not sufficient
Low = The evidence provided is weak
NA = The criteria is not applicable to this research
Not Mentioned = No evidence is provided for the criteria

Study Name:		**Evidence Rating**					**Reviewer Name:**
Qualitative Review Questions	**Consider** Green = Highly Important, Yellow = Moderately Important	**High**	**Med**	**Low**	**NA**	**Not Mentioned**	**Reasoning behind selection. Please support your answers with text from the article and pg. numbers and comment on both strengths and weaknesses where applicable.**
1. Clear statement of research.	• the goal of the research						
	• why it is important						
2. Appropriateness of qualitative methodology	• Does the research interpret or illuminate the actions and/or subjective experiences of research participants						
3. Research design addresses the	• Is the research guided by research questions or hypotheses?						

Study Name:		Evidence Rating					Reviewer Name:
Qualitative Review Questions	**Consider** Green = Highly Important, Yellow = Moderately Important	**High**	**Med**	**Low**	**NA**	**Not Mentioned**	**Reasoning behind selection. Please support your answers with text from the article and pg. numbers and comment on both strengths and weaknesses where applicable.**
aims of the research	• If the researcher has justified the research design (i.e., have they discussed how they decided which methods to use)?						
4. Appropriate recruitment strategy	• if the researcher has explained how the participants were selected;						
	• if they explained why the participants they selected were the most appropriate to provide access to the type of knowledge sought by the study						
5. Was the data collected in a way that addressed the research issue?	• if the setting for data collection was justified;						
	• if it is clear how data were collected (e.g., focus group, semi-structured interview etc.)						
	• if the researcher has justified the methods chosen						

Study Name:		Evidence Rating					Reviewer Name:
Qualitative Review Questions	**Consider** Green = Highly Important, Yellow = Moderately Important	**High**	**Med**	**Low**	**NA**	**Not Mentioned**	**Reasoning behind selection. Please support your answers with text from the article and pg. numbers and comment on both strengths and weaknesses where applicable.**
	• if the researcher has made the methods explicit (e.g., for interview method, is there an indication of how interviews were conducted, did they used a topic guide?)						
	• if methods were modified during the study. If so, has the researcher explained how and why?						
	• if the form of data is clear (e.g., tape recordings, video material, notes etc.)						
	• if the researcher has discussed saturation of data						
6. Has the relationship between researcher and participants been adequately considered?	• Consider if the researcher critically examined their own role, potential bias and influence during: • formulation of research questions and research instruments (e.g., asking leading questions)						
	• data collection, including sample recruitment and choice of location						

Study Name:		Evidence Rating					Reviewer Name:
Qualitative Review Questions	**Consider** Green = Highly Important, Yellow = Moderately Important	**High**	**Med**	**Low**	**NA**	**Not Mentioned**	**Reasoning behind selection. Please support your answers with text from the article and pg. numbers and comment on both strengths and weaknesses where applicable.**
7. Have ethical issues been taken into consideration?	• if there are sufficient details of how the research was explained to participants for the reader to assess whether ethical standards were maintained						
	• if the researcher has discussed issues raised by the study if on sensitive issues (e.g., issues around informed consent or confidentiality or how they have handled the effects of the study on the participants during and after the study)						
	• if approval has been sought from an ethics committee						
8. Was the data analysis sufficiently rigorous?	• if there is a thorough description of the analysis process						
	• if thematic analysis is used. If so, is it clear how the categories/themes were derived from the data?						

Study Name:		Evidence Rating					Reviewer Name:
Qualitative Review Questions	**Consider** Green = Highly Important, Yellow = Moderately Important	**High**	**Med**	**Low**	**NA**	**Not Mentioned**	**Reasoning behind selection. Please support your answers with text from the article and pg. numbers and comment on both strengths and weaknesses where applicable.**
	• whether the researcher explains how the data presented were selected from the original sample to demonstrate the analysis process (ex. I chose this because 90% of the participants said something similar)						
	• if sufficient data are presented to support the findings						
	• to what extent contradictory data are taken into account						
	• whether the researcher critically examined their own role, potential bias and influence during analysis and selection of data for presentation						
	• if the researcher considered contextual factors which may have influenced the research results (if you do a study in Peru, you must take into consideration context of Peru) Urban vs. Rural, etc.						

Study Name:		Evidence Rating					Reviewer Name:
Qualitative Review Questions	**Consider** Green = Highly Important, Yellow = Moderately Important	**High**	**Med**	**Low**	**NA**	**Not Mentioned**	**Reasoning behind selection. Please support your answers with text from the article and pg. numbers and comment on both strengths and weaknesses where applicable.**
9. Is there a clear statement of findings?	• if the findings are explicit						
	• if there is adequate discussion of the evidence both for and against the researcher's interpretations						
	• if the researcher has discussed the credibility of their findings (e.g., triangulation, respondent validation, more than one analyst)						
	• if the findings are discussed in relation to the original research questions						
10. How valuable is the research?	• if the researcher discusses the contribution the study makes to existing knowledge or understanding (e.g., do they consider the findings in relation to current policy, or relevant research-based literature?)						
	• if they identify new areas where research is necessary						

Study Name:		Evidence Rating					Reviewer Name:
Qualitative Review Questions	**Consider** Green = Highly Important, Yellow = Moderately Important	**High**	**Med**	**Low**	**NA**	**Not Mentioned**	**Reasoning behind selection. Please support your answers with text from the article and pg. numbers and comment on both strengths and weaknesses where applicable.**
	• if the researchers have discussed whether or how the findings can be transferred to other populations or considered other ways the research may be used						
11. Information for stakeholders to assess replicability	• Does the paper provide adequate details on the design and implementation of the intervention to enable replication? Such as: a. Length of training b. Monitoring tools c. Training materials etc.						
Summarize the main findings of this article in regards to how it might affect our main stakeholder groups (policy makers, Intl NGOs, Teacher training institutes, researchers, etc.)							
Please list other potentially relevant references from the bibliography that should be checked							
1.)							
2.)							
3.)							
4.)							
5.)							

Appendix F. Effect Size Extraction Form

Effect Size Extraction/Coding
Study ID (sid): Need to Contact Authors (authors): Coders initials (coderid): Date coded (date): Country (cntry): Region in the world (region): Intervention type (inter):
Grade at start of the intervention (grade_st): Grade at time of impact estimate (grade_imp): Age of children at the start of the intervention (age_st): Age of children at time of impact estimate (grade_imp):
Methodology (method): Outcome measure (outcat): (1) reading comprehension (2) letter naming (3) letter sounds (4) time spent on reading (5) vocabulary (6)phoneme segmentation (7) letter-naming fluency (8) word reading (9) new word learning (10) fluency reading time together (11) comprehension (12) literacy scores (13) reading (14) spelling (15) English (16) letter word identification (17) early writing (18) language Outcome name (outname): With covariates (_covar): Effect size type (estype): (1) Standardized mean difference (2) other Other name (oth_name): Direction of effect (esdir): (1) effect favors treatment (2) effect favors comparison (3) effect favors neither (4) cannot tell
Effect is statistically significant (essig)?: (1) yes (2) no (3) cannot tell Treatment students sample size (trtss): Comparison students sample size (compss): Total students sample size (totals): Treatment cluster sample size (trtss_clus): Comparison cluster sample size (compss_clus): Total cluster sample size (totals_clus):
For continuous measures: Treatment group mean (txmean): Comparison group mean (compmean):
Are means reported above adjusted? (meanadj): (1) yes (2) no Treatment group standard deviation (txsd): Comparison group standard deviation (compsd): Treatment group standard error (txse): Comparison group standard error (compse): Mean difference (mdiff): Standard error mean difference (semdiff): Standard error in regression (seregress): Standard error in matching (sematching): *t*-value regression or single difference (est)

Pooled standard deviation (psd): Standardized mean difference (smd): Small sample size adjusted standardized mean difference (ssmd): Standard error Standardized mean difference (se_smd): *t*-value standardized mean difference (est_smd) Treatment time (trt_time):
Source: Wilson et al. (2014)

Appendix G. Articles Rejected—Breakdown by Inclusion Criteria

	Yes	No	Unclear	Unrated (because other criteria are not met)
Published after 1990?	1,138	5	5	
Study on the LAC Region?	458	560	6	124
Boys or girls birth through grade 3?	248	215	17	668
Focus on reading or literacy?	186	134	1	827
Is it research?	166	60	0	922
Does the research meet minimum criteria for the analysis?	124	42	0	982