THE EFFECTS OF ABRACADABRA ON READING OUTCOMES: AN UPDATED META-ANALYSIS AND LANDSCAPE REVIEW OF APPLIED FIELD RESEARCH

Abrami, P. C., Lysenko, L., & Borokhovski, E.

;

IDRC Grant: 108356-001-Scaling digital learning in Kenya

This is the pre-peer reviewed version of the following article: Abrami, P. C., Lysenko, L. & Borokhovski, E. (2020). The effects of ABRACADABRA on reading outcomes: An updated meta-analysis and landscape review of applied field research. Journal of Computer-Assisted Learning. 36, 260-279 , which has been published in final form at https://doi.org/10.1111/jcal.12417. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions

THE EFFECTS OF ABRACADABRA ON READING OUTCOMES: AN UPDATED META-ANALYSIS AND LANDSCAPE REVIEW OF APPLIED FIELD RESEARCH

Running head: THE EFFECTS OF ABRA ON READING OUTCOMES

Abstract

ABRACADABRA (ABRA) is an evidence-based suite of interactive multimedia that engages learners in the development of core reading skills. This meta-analysis presents an update of the research evidence about the effectiveness of ABRA for elementary students. It reports 91 effect sizes in six reading-related outcomes for a sample of 7,388 students. Regardless of context and measurement type, the studies yielded positive effects of ABRA, ranging in magnitude from $g+=0.080$ for *Vocabulary Knowledge* to $g+=0.378$ for *Phonemic Awareness*, and reaching statistical significance in four outcome categories. This meta-analysis adds to our understanding of the effectiveness of ABRA-based reading instruction by exploring factors of research design, ABRA design and implementation contexts, and various student characteristics and offers implications for instructional practice.

Keywords: early reading instruction, meta-analysis, interactive software for learning, elementary education, reading outcomes

THE EFFECTS OF ABRACADABRA ON READING OUTCOMES: AN UPDATED META-ANALYSIS AND LANDSCAPE REVIEW OF APPLIED FIELD RESEARCH

Introduction

Contemporary education practices are hard to imagine without the use of computer technology. As a second-order meta-analysis by Tamim, Bernard, Borokhovski, Abrami, and Schmid (2011) has shown, technology use in education can have a positive influence on learning outcomes. Its impacts, however, can vary substantially, especially if technology-based interventions are not carefully designed and well-implemented (e.g., Tacacz, Swart & Bus, 2015; Schmid et al., 2014). Thus, special attention has to be paid to those instructional interventions that incorporate and promote evidence-based practices of "what works" in education (e.g., Cheung & Slavin, 2012; Abrami, Bernard, Bures, Borokhovski, & Tamim, 2011). Early literacy instruction is one of numerous curricular areas that could substantially benefit from technological innovations (e.g., Cheung & Slavin, 2012; Slavin, Lake, Chambers, Cheung & Davis, 2009). In their most recent overview of early literacy software, Wood, Grant, Gottardo, Savage and Evans (2017) offered a thorough evaluation of offline and online computer applications targeting basic reading skills and named ABRACADABRA (A Balanced Reading Approach for Children Designed to Achieve Best Results for All, ABRA for short) the most highly-rated online literacy tool. Over the years ABRA has also been studied extensively in applied field educational research.

Abrami, Borokhovski & Lysenko (2015) synthesized a collection of 9 experimental field studies on ABRA in a systematic review of empirical research conducted on a sample of 2,739 of elementary students (N_{exp} =1,443; $N_{control}$ =1,296). The review documented the effectiveness of the tool on elementary school students' basic reading skills and competencies in a variety of contexts using a range of standardized measures and strong field research designs. Taken together, the results provided positive evidence of the overall value of ABRA as a tool to promote the development of early literacy skills. All average effect sizes were positive. Of a total of seventy-three individual effect sizes, fifty-seven were positive and only sixteen were negative. Only two sets of outcome measures revealed statistically heterogeneous results implying that ABRA-based early literacy interventions may work consistently in the different contexts represented in that review. In most of the ABRA interventions reviewed, student exposure to the tool and its activities was about two hours per week for several months totaling nearly two dozen hours. This implies that with quite limited but targeted exposure, literacy gains were still noticeable.

Since 2015, ABRA software has continued to attract the interest of the research community generating new empirical evidence; seven new studies were conducted internationally. Therefore, it has become important to ensure the review is current by updating its findings and expanding them by relying on several important, new ways to examine the findings. Adding new research to the existing collection from Abrami et al., (2015) resulted in a new total of 91 effect sizes based on the data from 3,341 students in experimental groups, where various forms of ABRA-based instruction unfolded, and 4,047 students in control conditions, who had no ABRA exposure; this nearly tripled the 2015 sample size. This update increases the statistical power of the overall analyses and provides an opportunity to introduce moderator variables to the analytical model, thus strengthening the reliability of findings and broadening their impact.

The added studies feature a broader international context including research completed in the UK, Mainland China and Hong Kong in addition to Canada, Australia and Kenya. Additional independent evaluations conducted by McNally, Ruiz-Valenzuella, and Rolfe (2016) and Bailey, Arciuli and Stancliff (2016) have also contributed to the research knowledgebase about ABRA impacts. In addition, because several new studies added to the pool of ABRA research used randomized cluster designs and accounted for class effects in the analyses, we also explored the potential of applying the clustering logic to the existing ABRA research where students were the units of analysis.

A Balanced Approach to Early Literacy Instruction

Educational research has clearly established a set of skills and sub-skills related to the development of emerging literacy. The following five overarching skills form the proposed taxonomy of reading abilities (National Reading Panel, 2000; Pressley et al., 2001; Taylor, Pearson, Clark, & Walpole, 2000; Savage & Pompey, 2008):

- *phonics* – ability to relate specific written letter(s) to specific sound(s);
- *phonological/ phonemic awareness* – ability to hear and manipulate individual sounds;
- *fluency* – ability to read text effortlessly and expressively;
- *vocabulary knowledge* – ability to recognize spoken and written form of the word

meaning; and

- *reading comprehension* – ability to understand and interpret printed text.

As the ability to understand and interpret text presented aurally, *listening comprehension*, is also related to the reading taxonomy although peripherally. The NRP report (2000) emphasizes active listening as a strategy to promote reading comprehension. In multilingual contexts where English is not the mother tongue, listening comprehension becomes a corner stone for building oral English proficiency, which is critical for literacy development of students who are below a threshold of linguistic competence in English (Bunyi, 2005; August & Shanahan, 2006). It is on this basis that listening comprehension was included as another outcome of ABRA interventions.

ABRA is a free software application that provides an interactive environment for learning literacy among young, school-aged children and its design is consistent with the taxonomy of reading skills outlined above in addition to well-established knowledge about reading instruction. The evidence-based nature of ABRA ensures the systematic integration of cumulative research on the major skills and associated sub-skills in the areas of alphabetics (i.e., phonics and phonemic awareness), fluency, comprehension and writing needed by successful readers (NRP, 2000). In line with the tenets of a balanced literacy philosophy, ABRA emphasizes a harmonious balance between code-emphasis and a literature-rich context. This approach allows students to explore their interests by applying a large repertoire of strategies that can be readily accessed when meaning breaks down (Pressley, 2002). Offering distinct environments (named modules) for students, teachers and parents, ABRA is neither linear in use nor prescriptive of a single concept or method of teaching and learning to read. This flexible and modular design of the software also grants access to a rich pedagogical resource. The instructional components can be repurposed based on teaching preferences and students' needs, allowing teachers to use it when, how, and with whom they see fit.

ABRA's Student Module contains a variety of instructional materials including 33 alphabetics, fluency, comprehension, and writing activities, many at different levels of difficulty and complexity. These instructional activities are linked to 20 interactive stories of various genres and 15 stories written by Australian and Canadian students. The gaming elements of ABRA are features designed to engage children in reading and writing and to increase their motivation. In each ABRA activity, children progress towards a goal following a set of simple rules. When this goal is reached, students are rewarded with a mini-game. Examples include being challenged to free a picture caught in seaweed before the time runs out, landing an ABRA character on a target with a parachute, protecting a whale from obstacles as they bring it to the ocean, or lighting up stars following the sequence they were shown. At times, the game is at the core of the pedagogical structure of the activities, such as in Word Matching, in which children have to find cards with similar sounds at the beginning or end of the words. ABRA characters also add to its game-like feel. Each character has a personal story the children can read or listen to that reinforces the purpose and context of what students have to do in each activity. This underlying narrative thread also helps create a gaming experience in ABRA. The embedded support within ABRA tailors the degree of learner scaffolding offered as students interact with the tool. If students answer incorrectly, they are provided with suggestions or can seek help.

Primarily designed as a student environment, ABRA also offers environments for teachers and parents to encourage their engagement and support for students learning. The teacher environment consolidates teaching material embedded in the tool. The Teacher Module provides just-in-time support for teachers and resources for classroom use including explanations, lesson plans, embedded video teaching vignettes, and printable resources. In addition, the teachers can access the teachers' manual available both electronically and in print form (Abrami, Meyer, et al., 2014). As part of the teacher environment, the Assessment Module enables teachers to review student and class performance on instructional activities for any period of time. Reports generated by this module communicate individual students' and class' learning needs in order for the teacher to focus on areas of instructional need and make decisions about the balance of instruction, that is the order in which these activities should be delivered.

The Parent Module of ABRA allows parents access to multimedia resources and tips on how to support student use of ABRA in the home. Finally, ABRA is available in both English and French, however the research summarized here examines only the English version of the tool.

ABRA implementation fidelity requirements have seemingly been set to ensure the optimal integration of the tool in teaching. In addition to adherence to a balanced teaching of key literacy components based on differentiation and progression of instruction to meet students' learning needs, the fidelity criteria also include time guidelines for student exposure to ABRA instruction. These guidelines recommend use of ABRA for about 2 hours per week per student for no less than 13 weeks.

Connected to ABRA is the Repository for E-Books and Digital Stories (READS), which is a searchable collection of multi-lingual stories available online and linked to ABRA. READS has been developed to supplement the stories within ABRA to help develop students' fluency and comprehension skills. READS contains over 500 free stories, many available in several languages, for instance Kiswahili, Mandarin and Hindi. The stories are geared primarily to emerging readers from Kindergarten through Grade 3. To allow readers from different cultural backgrounds and instructional contexts to enjoy the stories, READS offers a variety of themes, genres, country of origins, etc. ABRA teacher support materials such as lesson plans also employ READS books and stories to further improve students' fluency and comprehension skills.

Naturally, instructional application of ABRA varied in the degree and scope of use of all these features across countries, grade levels, etc. Accumulated applied research has now allowed for the estimation of overall effects of the tool on major reading outcomes. In addition, this research has explored the differential contribution of specific instructional conditions, including age of learners, duration of treatment, fidelity of implementation and other substantive study characteristics to successful implementation of ABRA.

Research Questions

The primary objective of this meta-analysis is to estimate the effectiveness of ABRA-based early literacy instruction on six basic reading skills outlined in the NRP (2000) guidelines: (1) Phonemic Awareness; (2) Phonics; (3) Reading Fluency; (4) Reading Comprehension; (5) Listening Comprehension; and (6) Vocabulary Knowledge – in comparison with regular instruction in reading. Six weighted average point estimates, representing synthesized available effect sizes in each category, are to be produced to inform this focal area of interest and answer the research question of comparative effectiveness of ABRA-based reading interventions.

In addition to this major research question, the study will explore and report whether any substantial changes in average effect sizes occurred with the addition of seven studies compared to the findings of a previous meta-analysis of ABRA effectiveness (Abrami et al., 2015). Finally, given the increase in diversity of empirical studies in the updated collection, this meta-analysis also addresses a set of secondary research questions in the subsequent analyses of moderator variables: Under what circumstances (i.e., substantive and demographic study characteristics) do

ABRA-based instructional interventions tend to be more or less effective in each category of reading outcomes?

Method

Data Source

The data for the review were studies carried out between 2008 (first ABRA research publication) and 2017 in formal educational settings (pre-K through grade 3) in various geographical locations. The systematic review team carefully reviewed all published and unpublished research reports in order to filter out possible sample overlaps and data duplicates. Table 1 below summarizes the descriptive details of the seventeen studies identified for inclusion in the current meta-analysis. It should be noted that studies 1 - 11 were reviewed in the 2015 meta-analysis (Abrami et al., 2015); the current review added studies 12 – 17.

Insert Table 1 about here

Inclusion/Exclusion Criteria

To be included in the meta-analysis, a study needed to meet the following criteria:

(1) Be conducted in a formal educational setting with students in pre-K – Grade 3;

(2) Feature some form of ABRA-based early literacy intervention;

(3) Compare independent samples of the experimental and control [1] students on at least one of six relevant reading outcomes;

(4) Report sufficient statistical information to enable effect size extraction; and

(5) Control for major threats to internal validity (i.e., use randomized control trials or quasi-experimental research design).

Review Procedures

Two researchers reviewed all available studies to ensure that inclusion criteria were met. They resolved disagreements through joint discussions, when necessary inviting a third opinion. Special attention was paid to establishing the independence of study samples and to categorizing outcomes according to the NRP (2000) classification scheme. In addition, the two reviewers

[1] ABRA exposure was the only distinction between experimental and control conditions. The latter varied in form of reading instruction which might also have included use of learning technologies but was not ABRA-based.

coded each study's research design. Data extraction was also carried out by these reviewers and results compared and finalized for admission to the subsequent analyses. Inter-rater agreement rates for all stages of the review are reported in the results section.

Occasionally, the same outcome type was addressed by more than one assessment tool for the same group comparison. Only one effect size of the same type for each independent sample was retained. The decision was made in favor of the most frequently used assessment for the purposes of more consistent representation of data sources across studies.

In rare cases where more than one experimental condition was compared with the same control, the sample size of the group used twice was reduced proportionally. For example, Savage, Abrami, Hipps, and Deault (2009) featured two types of ABRA-based instruction (Synthetic and Analytic approaches) in comparison with the same non-ABRA control condition. Both comparisons were retained, but the control group sample size in each of them was split in half. Finally, in one study (Wolgemuth et al., 2010), two of the six reported settings lacked pre-test data for the control groups and so the results for these settings were discarded.

Measures Used

This section offers a brief description of instruments that studies used to measure the reading outcomes in the categories of phonics, phonemic awareness, reading fluency, reading comprehension, listening comprehension, and vocabulary knowledge as suggested by the National Reading Panel report (2000). Measures of reading readiness and pre-reading skills (i.e., print awareness auditory and visual skills, rapid naming, etc.), as well as oral proficiency skills were excluded.

Phonemic Awareness was measured by the subtests selected from a number of the standardized assessment tools. These include the *Group Reading Assessment and Evaluation* (GRADE; Williams, 2001) *Phonological Awareness* subtest including sound matching and rhyming tasks that assess the child's ability to hear like sounds (sound discrimination and sound matching) and ability to hear matching common monographs (rhyming). The *Comprehensive Test of Phonological Processing* (CTOPP; Wagner, Torgesen & Rashotte, 1999) subtests of Elision and Blending Words were used to assess awareness and access to phonological structure of oral language. Elision measures the analysis component of phonological awareness or the extent to which a student can say a word, repeat it and then say what is left after dropping out designated sounds. Blending Words targets the synthesis component of phonological awareness and assesses a student's ability to combine sounds to form words. Two subtests of the *Dynamic Indicators of Basic Early Literacy Skills* (DIBELS; Good & Kaminski, 2002) standardized test were used and included Phonemic Segmentation Fluency, assessing a student's ability to produce the individual sounds within a given word, and Initial Sound Fluency, testing student ability to hear and produce the initial sounds in words. *Preschool and Primary Inventory of Phonological Awareness* (PIPA; Dodd, Crosbie, McIntosh, Teitzel & Ozanne, 2000) was also used in ABRA studies and included the subtests of *Phoneme Segmentation* to measure skills of breaking down words into smaller parts and *Letter Sound* to assess ability to identify the sound (phoneme) that corresponds to each letter. *Letter-Sound Test* (LeST; Larsen, Cohnen, McArthur & Nickels, 2015) was employed to measure ability to sound out single letters and letter combinations.

A selection of subtests was used to assess Phonics outcomes. The *Letter-Sound Knowledge (LSK)* measure asked students to say the corresponding sound of each of the 26 letters of the English alphabet. The *Phoneme-Grapheme Correspondence* subtest of GRADE assessed the ability to recognize a phoneme or sound and the ability to match a symbol to that sound. *Word Attack* and *Letter-Word Identification* scales from *Woodcock-Johnson III Tests of Achievement* (Woodcock, McGrew, Mather, 2001) were also used. *Word Attack* employs grapheme-to-phoneme translation of pseudo-words and measures the ability to assemble the pronunciation of a letter string by applying knowledge of typical correspondences between graphemes and sounds, whereas *Letter-Word Identification* is a measure of the ability to identify letters and words. The *Wide Range Achievement Test IV* (WRAT 4; Wilkinson & Robertson, 2006) subtest of *Word Reading* measured decoding through letter identification and word recognition. The *Spelling* subtest assessed students' capacity to encode sounds into written form. *Diagnostic Test of Word Reading Processes* (DTWRP; Forum for Research in Literacy and Language, 2012) was used to assess the processes that underlie recognition and understanding of written words: regular words, exception words and non-words. *The DIBELS* (2002) *Nonsense Word Fluency* subtest assessed ability to blend letters into words in which letters represent their most common sounds.

Reading Fluency outcomes were also measured by four tests. The *Reading Fluency* subtest battery from the Woodcock-Johnson III Tests of Achievement (Woodcock, McGrew, Mather, 2001) focuses on reading speed by measuring the automaticity of access to words and their meaning in the mental lexicon as well as comprehension of simple sentences. DIBELS (Good &

Kaminski, 2002) *Oral Reading Fluency* measures a student's skill at reading a connected text correctly and fluently. Students' ability to recognize these words by sight as a prerequisite to reading fluency was tested with the *Fry Words* list (Fry, 1980), which contains words in reading and writing divided by frequency of use and difficulty. Text-reading accuracy as an aspect of fluency was tested with the Neale Analysis of Reading Ability scale (NARA-3; Neale, 1999) subtest of *Passage Reading Accuracy* measuring accuracy, comprehension and rate of reading.

The change in students' Reading Comprehension skills was measured primarily by the GRADE (Williams, 2001) subtests of *Sentence and Passage Comprehension*. Sentence Comprehension measures student ability to comprehend a sentence as a whole by using contextual cues, knowledge of grammar and vocabulary. Passage Comprehension employs a variety of multiple-choice comprehension questions about each of 24 text passages of different types, topics, and lengths. The subtest of *Passage Reading Accuracy* of the Neale Analysis of Reading Ability scale (NARA-3; Neale, 1999) was used to measure comprehension, although the main focus of this subtest is on text-reading accuracy.

The outcomes of Vocabulary Knowledge were assessed by the vocabulary subtests of GRADE (Williams, 2001) that measure student's ability to both decode regularly spelled words and recognize sight words (*Word Reading*) as well as understanding of early-reading vocabulary (*Word Meaning*). The *British Picture Vocabulary Scale* (BPVS-3; Dunn, Dunn, Styles, & Sewell, 2009) was also used to test students' receptive vocabulary.

GRADE's *Listening Comprehension* scale (Williams, 2001) was the only measure of Listening Comprehension outcomes used in the included ABRA studies. This subtest assessed students' linguistic comprehension without presenting printed cues.

Statistical Procedures

Effect Size Extraction

A *d*-type (Cohen's *d*) standardized mean difference effect size (Cohen, 1988) was used as the common metric. Equation 1 expresses this method of calculation when all descriptive information is available.

$$d = \frac{\bar{X}_E - \bar{X}_C}{SD_{Pooled}} \quad (1)$$

A modification of this basic equation was used for studies reporting pre-test and post-test data for both experimental and control groups, where the numerator utilized the difference between the corresponding gain scores with the value of the post-test pooled standard deviation in the denominator (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009). When descriptive statistics were not available, effect sizes were extracted from inferential statistics, such as *t*-tests, *F*-tests, or exact *p*-values, using conversion equations from Glass, McGaw and Smith (1981) and Hedges, Shymansky and Woodworth (1989). To correct for small sample bias, *d* was converted to the unbiased estimate *g* (Hedges & Olkin, 1985), as in Equation 2.

$$g = d\left(1 - \frac{3}{4N - 9}\right) \qquad (2)$$

Regression-based Adjustment

Three of the included studies employed randomized cluster research design and utilized the Hierarchical Linear Modeling (HLM) procedure to analyze the data. The strength of such an approach was two-fold. First, it allowed the researchers to encompass more representative and diverse samples across multiple schools and school districts while employing random assignment of students to experimental and control conditions. Second, it permitted accounting for extraneous sources of variance (e.g., socio-economic status of students' families) when evaluating the impact of ABRA-based instruction on reading outcomes. Arguably, this research strategy is more robust and sensitive to details with more error variance removed, and thus capable of producing more reliable findings than typical experimental studies that rely solely on treating individual students as the unit of analysis.

We were interested in estimating the results that this approach would yield when applied to the entire collection of ABRA-based research. Three studies (McNally et al., 2016; Piquette, Savage & Abrami, 2014; Savage et al., 2013) together produced 20 pairs of effect sizes across all six categories of reading outcomes. Each study allowed for two different ways of calculating effect sizes: (1) based on reported HLM coefficients and (2) based on descriptive statistics of the corresponding samples. Derived both ways, a set of these 20 paired effect sizes was subject to a regression analysis to examine how the HLM-based effects could then be used to predict the respective effect sizes from the non-HLM studies.

Since the obtained regression model was statistically significant ($p = .021$), we applied the resulting equation to estimate effect sizes in the 14 non-HML studies with their 71 effects. The "adjusted" effects were then aggregated the same way the actual effect sizes were. In the Results section we report both non-adjusted and adjusted weighted average effect sizes and associated statistics by outcome category.

Effect Size Aggregation and Analysis

With seven additional studies, diversity in ABRA-based interventions and conditions of implementation reached the point beyond which applying only the fixed effect model was no longer a viable option. Instead, effect size aggregation was done within the paradigm of the Random Effects Model, which weighs individual effect sizes by the inverse of within-study variance with the addition of the average between-study variance. The Fixed Effect Model, however, was employed to estimate total variability (Q_{Total}) and test for heterogeneity of the effect size distribution. The corresponding values of I^2 (i.e., percentage of heterogeneity in effect sizes exceeding chance sampling expectations; Higgins, Thompson, Deeks & Altman, 2003) are also reported. All analyses, including sensitivity and publication bias analysis, were performed in *Comprehensive Meta-Analysis*™ 3.3 (Borenstein, Hedges, Higgins & Rothstein, 2014).

Because there was a limited number of studies overall, there was a low number of independent effect sizes per outcome category. This low number rendered the statistical power of the standard moderator variable analysis insufficient. Therefore, our study opted to use a "vote count" descriptive technique as a means to explore the variability of effects. Individual effect sizes for both within-study (e.g., gender, learners' preexisting reading abilities) and between-study (e.g., duration of ABRA intervention, average teacher-student ratio) categories of moderators were divided into three levels of magnitude. These three levels of magnitude were: (1) positive effects (above 0.1); (2) 'trivial' or zero-like effects (within the boundaries of -0.1 and +0.1); and (3) negative effects (under -0.1). For each moderator variable the total number of effects at each level of magnitude was reported, while also indicating to which outcome type they belonged. Estimates of the non-independent average (i.e., across outcomes from some repeated samples) per comparison category were also presented.

Results

Overall, seventeen studies addressing the effectiveness of ABRA-based early literacy interventions in comparison with regular instruction in reading were identified and reviewed in this updated and expanded meta-analysis. Reviewers extracted relevant effects sizes in six outcome categories. The final collection contained 91 independent effect sizes.

At each stage of the review, decisions were made independently by two researchers, who then met to discuss and resolve disagreements. Established in this way, inter-rater agreement rates were:

- Inclusion/Exclusion decisions – 94.4 % (Cohen's kappa = 0.89);
- Classification of outcomes – 88.7% (Cohen's kappa = 0.77);
- Effect size extraction and adjustment – 89.5% (Cohen's kappa = 0.79);
- Coding of moderator variables – 85.2% (Cohen's kappa = 0.70)

Sensitivity and Publication Bias Analyses

Visual examination of the data distribution in each outcome category and a "one study removed" CMA routine (Borenstein et al., 2014) revealed no obvious outliers. The original meta-analysis (Abrami et al., 2015) identified and excluded one outlier among effect sizes in the Listening Comprehension category, and it was not retained and used in this review either.

Publication bias analysis is intended to determine if some effects might have been missed by the meta-analysis and if so, how they would have affected the findings (Rothstein, Sutton, & Borenstein, 2005). Non-significant (i.e., inconclusive) results may not require this scrutiny, however the significant ones do. As shown in the section below, four out of six weighted average effect sizes were statistically significant. Three of those--Phonics, Phonemic Awareness, and Listening Comprehension--produced balanced distributions of effect sizes, as reflected both in the respective funnel plots and in the results of several types of fail-safe CMA routines. They affirmed the robustness of these distributions as well. Specifically, Classic Fail-Safe analyses (Rosenthal, 1979) established that to render effect sizes in these categories statistically insignificant, 66, 455, and 69 potentially missing "null-effects" would need to be added to the respective distributions. According to the Duval and Tweedie's Trim and Fill analytical procedure (Duval & Tweedie, 2000), only one outcome type, Reading Comprehension, yielded a relatively unstable weighted average effect size. Introducing one "null-effect" would result in the reduction of its weighted average from the observed $g+ = 0.180$ to an adjusted $g+ = 0.167$. However, the average effect would remain statistically significant if a nonsignificant study was added.

Main Analysis Results by the NRP (2000) Outcome Categories

The collection of ABRA-based interventions and the conditions of their implementation were not consistent across studies, varying rather substantially in duration, geographical locations, students' mother tongues and some key delivery components (e.g., with a teacher or a research assistant administering ABRA). Therefore, we aggregated effect sizes for each outcome type category according to the Random Effects Model. Results of these analyses are presented in six consecutive summary tables below for both unadjusted and adjusted effects.

Eighteen out of twenty-three effects in the *Phonics* outcome category (Table 2) were positive and only five were negative. They were based on a sample of 3,273 students and produced an average unadjusted effect size of $g+ = 0.187$ ($p = .006$), with individual effect sizes widely ranging from –0.780 to +0.716. The heterogeneity index was statistically significant at $p = .003$. The average adjusted effect size was larger, g+ = 0.263.

Insert Table 2 about here

All twenty-three effect sizes in the outcome category of *Phonemic Awareness* (Table 3) were positive, together resulting in the overall weighted average of $g+ = 0.378$ ($p < .001$). The associated total sample size was 3,384. The magnitude of individual effect sizes ranged from +0.101 to +1.038. Despite the substantial number of independent effects (including three added since 2015), this collection was homogeneous: $Q_{Total} = 25.45$ ($df = 22$, $p = .28$). The average adjusted effect size was somewhat smaller, g+ = 0.299.

Insert Table 3 about here

The *Reading Fluency* outcome category (Table 4) contains only seven effect sizes based on a total sample of 1745 students. The overall weighted average effect size of $g+ = 0.088$ was not statistically significant ($p = .38$), with individual effect sizes ranging from –0.134 to +0.471. Despite the small number of effect sizes in it, this collection was significantly heterogeneous $Q_{Total} = 18.95$ ($df = 6$, $p = .004$). The average adjusted effect size was larger, g+ = 0.181.

Insert Table 4 about here

All but one out of eleven effects in the *Reading Comprehension* outcome category (Table 5) were positive and ranged in magnitude from –0.111 to +0.603. The overall weighted average effect size based on a sample of 4,593 students was statistically significant: *g*+ = 0.180, *p* = .01, as well as the heterogeneity statistics (Q_{Total} = 37.04, p < .001). The average adjusted effect size was larger, g+ = 0.240.

Insert Table 5 about here

The *Listening Comprehension* outcome category contained nine individual effects (Table 6) from 2,165 students. These effect sizes ranged from –0.074 (the only negative effect) to +0.686. The overall weighted average effect size of *g*+ = 0.274 was statistically significant (*p* = .02). The collection was significantly heterogeneous (*p* < .001). The average adjusted effect size was larger, g+ = 0.313.

Insert Tables 6 about here

Eighteen effect sizes (total sample size 3,310) in the *Vocabulary Knowledge* outcome category (Table 7) ranged from –0.774 to +0.655 with roughly equal number of negative (*k* = 8) and positive (*k* = 10) individual effects. This collection was significantly heterogeneous (p < .001), resulting in a non-significant overall average effect size of *g*+ = 0.080 (p = .401). The average adjusted effect size was markedly larger, g+ = 0.183.

Insert Table 7 about here

Across outcome categories, the non-independent weighted overall average unadjusted effect sizes were *g*+ = 0.200 and *g*+ = 0.174 for the Random Effects and Fixed Effect Models, respectively (*k* = 91, *p* < .001). As is evident from Tables 2 – 7, the effects of ABRA-based instruction on six major reading outcomes were predominantly positive, though low or low-to-moderate in magnitude and in two cases not reaching the level of statistical significance. In those instances, average effect sizes resulted from very close differences in the gain scores between experimental and control groups, whereas the magnitude of actual gain scores was quite substantial. For example, similarly to the findings of Abrami et al (2015), the average effect size in the *Vocabulary Knowledge* outcome category (*g*+ = 0.08, *p* = .401 in the Random Effects

Model) may reflect the fact that experimental and control conditions gained in learning new words almost equally – the difference between post-test and pre-test scores, expressed in un-weighted standard deviation units was close to +0.49 for both groups.

Finally, the non-independent weighted overall average adjusted (through regression based on available HLM studies, as described earlier) effect size was $g+ = 0.256$, ($k = 91$, $p < .001$).

Supplementary Analyses

This section summarizes the results of the analyses we completed to explore the within-study and between-study factors that might mediate the results of the main analysis using the 'vote count' approach.

Within-Study Factors

By their design, some studies included into this review accounted for factors that might have contributed to ABRA effects on students' reading skills. These mainly encompassed student characteristics such as baseline reading ability, gender, economic and indigenous status, and prior ABRA experience. ABRA implementation fidelity was the only instruction-related factor. Whenever studies under review reported data for these characteristics separately by their levels (e.g., gender), the corresponding effect sizes (e.g., ABRA-based instruction vs. control for male and female students) were extracted. A factor might have been a feature of a single study or shared by several studies.

After sorting the list of within-study factors, we calculated effect sizes by reading outcome on the basis of the available descriptive statistics and counted "votes" by dividing the effect sizes in three categories, as described earlier in the Method section. These were non-trivial positive (higher than 0.1), non-trivial negative (lower than -0.1) and trivial effects (ranging between -0.1 and 0.1). Table 8 displays effect sizes by the levels of each factor and also offers total effect sizes and sample sizes split between experimental and control conditions. Forty-four positive, non-trivial effects suggest that across all within-study factors the ABRA groups consistently outperformed students in the control conditions. Nineteen trivial effects, suggesting no difference between experimental and control conditions, also surfaced for all available comparisons. Negative effects were found for only four comparisons.

Insert Table 8 about here

Specifically, students' reading ability at the baseline was accounted for in three studies (Abrami, Wade, et al., 2014; McNally et al., 2016; Lysenko, Abrami, & Wade, under review) but after the ABRA intervention, the low-reading students benefit the most. In the context of the Kenyan studies, using ABRA was also found to be beneficial for students with higher reading ability (Abrami et al., 2014; Lysenko et al., under review). On the contrary, McNally et al. (2016) reported that ABRA intervention did not affect above-median pupils compared to controls.

The same three studies accounted for gender in their analyses of ABRA effects (Abrami et al., 2014; McNally et al., 2016; Lysenko et al., under review). The results of the studies suggest that both male and female students significantly benefit from using ABRA and outperformed students of both genders in the control condition. In McNally et al. (2016) the effects of ABRA were equivalent for both genders whereas the Kenya studies reported somewhat higher student reading gains for girls compared to boys.

Economic disadvantage is a student demographic characteristic accounted for by McNally et al. (2016). Based on government-provided information, the study identified as economically disadvantaged those students who were eligible for free school lunches. The findings indicate higher benefits of ABRA for disadvantaged children whose reading ability increased by almost 37% of a standard deviation compared to their peers from the control group.

Wolgemuth et al. (2011, 2013) examined whether the effects of ABRA instruction varied for indigenous versus non-indigenous students in the Northern Territory of Australia. Their research revealed that both indigenous and non-indigenous ABRA students' reading gains were significantly higher compared to the control group. Wolgemuth et al. (2013) also reported that ABRA has potential for preventing lags in foundational literacy experienced by indigenous students; they gained significantly more that non-indigenous students on phonemic awareness.

Student prior experience with ABRA was studied by Lysenko et al. (under review) in the context of a project that unfolded in Kenya public schools over a few years, where some students were exposed to the programme longer than one year. Although the results suggested that new students in the ABRA group gained the most, longer experience with ABRA allowed the students to continue maintaining their superiority.

Two studies accounted for the effects of different levels of ABRA fidelity of implementation on students' reading gains (Savage et al., 2010; Savage et al., 2013). Their findings suggest that implementing ABRA in reading instruction does make a difference – stronger implementation of ABRA produced higher gains.

Between-Study Factors

Where observable, we also analysed a set of between-study factors, although these characteristics were not directly controlled within each study's design but varied from study to study. These characteristics pertained to context, instruction, and students.

The *Country* where the study was completed was the only context-related study feature that was analysed. Country relates to the educational context where the study occurred including the overall level of advancement of the educational system and those aspects of the local educational environment that relate to access to resources. For instance, Kenya, the Northern Territory of Australia and rural regions in Mainland China were the contexts that relied on sparse educational resources versus more affluent locations, such as in Canada, the UK, Australian provincial schools, and Hong Kong where resources, including technology and technology support, were more readily available.

Instruction-related features included ABRA intervention delivery, exposure to ABRA instruction, and the student-teacher ratio.

ABRA intervention delivery. As noticed in the earlier meta-analysis (Abrami et al., 2015), ABRA effectiveness was related to the authenticity of the delivery context. Authenticity indicates how close to realistic learning environments the conditions of ABRA intervention delivery were. If ABRA-based interventions were delivered by ABRA-trained classroom teachers, special educators, tutors, or professionals who knew well their students' learning capacities and needs, ABRA integration was viewed as superior to those interventions delivered either by a research assistant or a teacher assistant solely for the purpose and only for the duration of the research study. Hence, we dichotomized between two levels of this moderator as either more or less authentic ABRA delivery.

ABRA treatment duration. Whenever reported, exposure to ABRA instruction expressed in hours (both – per week and total) was used in the 'vote count' analyses. This indicator of total exposure to ABRA varied considerably ranging from as low as 3.5 hours in the study by Cheung, Mak, Abrami, Wade & Lysenko (2016) to 20 (Lysenko et al., under review) and exceeding 30 hours (Wolgemuth et al., 2011). The descriptions offered in the reports indicated whether these estimates included time spent on demonstrating ABRA activities to the whole class, teaching

ABRA-related content, doing extension activities or if it only included student-ABRA interactions. We used the threshold of about 1.5-2 hours per week per student for no less than 13 weeks as a cut-off point. Therefore, we distinguished between sufficiently long interventions of 20 and more hours and relatively short interventions lasting fewer than 20 hours.

Student-teacher ratio. The number of students per individual instructor (whether it was a regular classroom teacher, teaching assistant, or a researcher implementing ABRA-based interventions) was used at two categorical levels to separate types of ABRA delivery conditions. More targeted delivery unfolded in the contexts where the instructor taught 10 or fewer students. In less targeted delivery conditions the instructor's attention had to be split among a relatively large numbers of students (more than 10 and sometimes 40 or more). Similar to other factors, student-teacher ratio varied among the 17 studies ranging from one-to-one in the context of ABRA instruction to children with Autistic Spectrum Disorder (Bailey et al., 2016; 2017) to one-to-forty or more in Kenyan classrooms (e.g., Lysenko et al., under review).

Finally, the only student-related feature observable in all ABRA studies was the student *grade level*. The majority of the studies dealt with samples representing several grade levels and grade splits (for instance, Anderson et al., 2011). In one of the studies (Bailey et al., 2016) ABRA-based remedial intervention was administered to special needs students ranging from grade one to grade four. Our approach was to distinguish and group the samples by grade level including kindergarten, and grades 1, 2 and 3.

Similar to the approach we took to summarize within-study factors, Table 9 presents the effect sizes calculated by reading outcomes as well as vote counts in each of the three levels of direction and magnitude.

Insert Table 9 about here

In general, the prevalence of positive effects and their non-independent weighted averages per outcome category suggest overall positive impacts of ABRA on student reading skills – independent of the country, fashion in which ABRA instruction was delivered, and student grade level. The highest effects have been noted for grade 2 students, in the context of one-on-one instruction, also emphasizing the importance of the teacher's role in high-quality ABRA implementation.

Discussion

This meta-analysis presents an update of the available research evidence about the effectiveness of ABRACADABRA early literacy software in teaching basic reading skills to elementary school students. A previous meta-analysis reported 65 effect sizes in six outcome categories based on a sample of 2,739 students, whereas the current study contained 91 effect sizes and the sample more than doubled, reaching 7,388 students. The average unadjusted effect sizes for all six outcome categories were positive, ranging from *g*+ = 0.080 for the outcome category *Vocabulary Knowledge* to g+ = 0.378 for the outcome category *of Phonemic Awareness*. The non-independent weighted average unadjusted effect size was g+ = 0.200. The adjusted effect sizes in the respective categories, with one exception, were larger in magnitude, with the overall non-independent weighted average adjusted effect size of g+ = 0.256. Unadjusted effects were statistically significant for *Phonics*, *Phonemic Awareness, Reading* and *Listening Comprehension*, whereas unadjusted effects for *Vocabulary Knowledge* and *Reading Fluency*, though positive, were relatively low in magnitude and statistically non-significance. Only one collection of effect sizes in the category of *Phonemic Awareness* was homogeneous, thus emphasizing consistency of ABRA-based influence on this particular outcome category across studies. To elucidate these findings, we discuss what this meta-analysis adds to our understanding of the effectiveness of ABRA-based reading instruction and then consider and summarize a variety of explanations for the pattern of results, including issues of research design, ABRA design and implementation challenges, and various student characteristics.

Overall Impacts

The first important purpose of this meta-analysis was to determine whether and to what extent ABRA has an impact on students' learning to read. To date, there have been 17 studies of ABRA, including high-quality RCTs and quasi-experiments, some undertaken as third-party independent evaluation. ABRA has not only been studied in a variety of contexts globally, ranging from North America to Africa and Asia, but with a range of internationally recognized standardized measures of literacy that tap a variety of literacy skills. All these studies of ABRA, regardless of context and measurement type, have found positive effects. The cumulative evidence shows ABRA-based instruction does work.

What is less obvious from individual studies are the conditions that lead to maximum impact. Subsequently, the second important purpose of this meta-analysis was to explore whether the effects vary as a factor of features either accounted for by the inherent design of the study or presented as part of each study's description. To that end, we ran a set of supplementary analyses that helped us to capture the nuances in implementation and other study features that may influence ABRA effects on elementary students' reading skills development.

ABRA Implementation

In the context of authentic classroom instruction where ABRA is in the hands of teachers, the issue of fidelity of implementation continues to be critical. Although the fidelity of implementation data is collected inconsistently and with varying quality, the evidence from ABRA efficacy trials (e.g., Savage et al., 2010; Savage et al., 2013) strongly suggests that quality implementation means teaching various aspects of early literacy in a balanced way and tailoring instruction to meet the learning needs of students. The high-implementing teacher has a set of skills that includes the ability to integrate technology in their teaching, as well as a range of pedagogical abilities such as good lesson planning, clear instructional differentiation, and a capacity to provide students with adequate exposure to ABRA. The pedagogical context in her classroom is student-centered, where students use ABRA in pairs and small groups and literacy instruction is marked by systematic attention to decoding and text-comprehension. Consequently, in the future it will be important to explore whether systematic use of ABRA's embedded support and teacher resources enhance the fidelity of classroom implementation. In addition, it is important to explore whether and how teacher professional development of ABRA instructional skills can be brought to scale and sustained via enhanced training.

Software cost-effectiveness is an important factor affecting implementation beyond just research. The studies under review did not pay much attention to the issue; however, McNally et al. (2016) examined the cost efficiency of ABRA. The average cost per pupil per year over three years was £8.52. This cost included training teaching assistants, cover during training, and travel costs. The Education Endowment Fund, who funded the research project, rated the ABRA intervention "1", or very low cost, on a five-point cost scale.

Student Characteristics

In addition to diversity in ABRA implementation, findings continue to suggest that ABRA effects may vary as a function of the range of student characteristics such as gender, reading ability, attention, etc. In this regard, an important pattern emerges suggesting that ABRA may have potential to help the most vulnerable groups of the elementary student population, those in the greatest need of adequate, targeted reading instruction. These include disadvantaged students (McNally et.al, 2016), poor readers (e.g., McNally et al., 2016; Abrami, Wade, et al., 2014) as well as indigenous students (Wolgemuth et al., 2011; 2013). In their study of inattention and reading development in a non-clinical sample, Deault, Savage and Abrami (2009) found that ABRA software can effectively compensate for variation in grade-one student attention skills. Indeed, the findings suggest that ABRA instruction can diminish the gap, well-known as the “Matthew effect” (Merton, 1968) where the differences between high-ability and low-ability students increase when they progress through the formal system of education. Moreover, new research by Bailey and his team (2016) has added to knowledge about ABRA effectiveness for children with special needs. Specifically, their research examined the impact of ABRA instruction on literacy with a diverse group of children with autism spectrum disorder (ASD) and demonstrated important experimental student gains in reading accuracy, comprehension and spelling accuracy in comparison to those of the wait-list control students. Individual effect sizes in this study were among the strongest in the entire collection. The reported findings imply that numerous features within ABRA could benefit children with a range of learning difficulties. ABRA uses auditory and visual processing, components of a well-designed multimedia that increase the effectiveness of the cognitive processing of information (Mayer, 2008) and thus may successfully captivate young students’ attention and maintain their focus on building fundamental reading skills. Further, the tool can reduce the difficulties the ASD population may have accessing the instructional content and engaging with it, as well as increase these students’ capacity to generalize the skills they learned with ABRA across various instructional contexts (Bailey et al., 2016).

ABRA provides access to gender-sensitive reading content and activities that appear to equally advantage both boys and girls. By design, the ABRA characters, their actions, and the activities that students engage in, help to dispel gender role stereotypes. Therefore, the finding suggesting that ABRA is about equally effective for students of both genders was not unexpected. Girls and boys who learned with ABRA show enhanced performance compared to students learning to read in the traditional manner (McNally et al., 2016; Lysenko et al., under review).

Further ABRA Design

Although ABRA is intended to be a balanced and comprehensive tool for teaching emerging literacy skills, refinement and expansion of the tool in some areas may be called for to further strengthen the effects, especially in fluency and comprehension.

The addition of READS, a collection of e-books and digital stories, may be an important supplement to ABRA as it provides a rich resource of reading materials (in multiple languages) for students to practice and strengthen their skills. This includes the development of new vocabulary and the improvement of reading fluidity and automaticity through reading practice with a variety of texts. Changing the nature of ABRA fluency activities to offer more feedback is another aspect that could be explored in future designs. For instance, Wood et al. (2016) suggest that placing students in groups by difficulty level as a function of their performance and achieving consistency in adequate scaffolding/feedback might be an important direction to take.

Another area of possible expansion for ABRA pertains to comprehension, and especially vocabulary knowledge, where the two ABRA activities focusing on vocabulary may not be enough to drive more substantial improvements. It may also be important to expand the traditional taxonomy of skills targeted by ABRA components of alphabetics, fluency, comprehension and writing based on NRP (2000). For instance, adding activities that target skills that are cutting-edge trends in literacy research (e.g., morphological awareness – Spencer et al., 2015) may help provide a more complete approach to teaching literacy and, thus, achieve higher literacy gains in reading comprehension.

Types of ABRA Studies and Research Design

It would not be an exaggeration to say that a meta-analysis is as good as the primary studies synthesized. With the new ABRA studies, research based on random assignment designs continues to be well-represented in the updated synthesis. Indeed, more than half of the studies in the current collection are randomized control trials including those where random assignment took place at the student level (for instance, Bailey et al., 2016), class level (for instance, Cheung et al., 2016) or multiple levels. The example of the latter is reported in McNally et al. (2016) where schools were assigned to treatment or control group and then students in treatment schools were randomly allocated to one of the three treatment conditions. Only three studies from the reviewed set accounted for the nested structure of the data that were generated by the designs in which intact groups of individuals were randomized to receive different interventions. These three rely on the Hierarchical Linear Model (HLM) analytical logic (Raudenbush & Bryk, 2002) allowing analysis of variance in the students' reading-related outcomes by taking into account a range of important factors, ABRA implementation being one of them (Savage et al., 2013).

Studies where randomized cluster research design is combined with the HLM analytical approach are known for their potential to generate more representative and reliable results. Based on this assumption, we used the outcomes of the three HLM studies in this meta-analysis to project potential results for the rest of the studies if they would have employed the same research paradigm.

Conclusion

ABRA is an evidence-based and evidence-proven approach to reading instruction that relies on the use of interactive multimedia to engage learners in the development of core reading skills. This research integration summarized the evidence collected to date from 17 high-quality field studies of the impact of ABRA in very different locations around the globe and on a wide range of quality reading measures. Each of the 17 studies found positive effects for ABRA compared to control conditions where ABRA was not used. Furthermore, ABRA's effects were generalizable across country contexts and measurement approaches.

Our findings suggest that ABRA benefits both boys and girls about equally and that low performing students and struggling readers were often able to learn the most and retain that learning beyond the initial intervention. Finally, ABRA was shown to be a cost effective solution to enhance the literacy skills of young children.

References

** Marked with asterisks are studies included in the meta-analysis*

Abrami, P. C., Bernard, R. M., Bures, E. M., Borokhovski, E., & Tamim, R. (2011). Interaction in distance education and online learning: Using evidence and theory to improve practice. *Journal of Computing in Higher Education, 23*(2-3), 83-103. http://dx.doi.org/10.1007/s12528-011-9043-x

Abrami, P. C., Borokhovski, E., & Lysenko, L. (2015). The effects of ABRACADABRA on reading outcomes: A meta-analysis of applied field research data. *Journal of Interactive Learning Research, 26*(4), 337-367. Retrieved from http://learntechlib.org/d/147396

Abrami, P. C., Meyer, E., Wade, A., Pillay, V., Lysenko, L., Lebel, C., & Venkatesh, V. (2014). *The learning toolkit: A teacher's guide*. Montreal, QC: Centre for the Study of Learning and Performance (CSLP).

*Abrami, P. C., Wade, A., Lysenko, L. V., Marsh, J., & Gioko, A. (2014). Using educational technology to develop early literacy skills in Sub-Saharan Africa. *Education and Information Technologies, 21*(4), 945-964. http://dx.doi.org/10.1007/s10639-014-9362-4

*Anderson, A., Wood, E., Piquette-Tomei, N., Savage, R., & Mueller, J. (2011). Evaluating teachers' support requests when just-in-time instructional support is provided to introduce a primary level web-based reading program. *Journal of Technology and Teacher Education, 19*, 499-525. Retrieved from http://www.editlib.org/p/34599

August, D., & Shanahan, T. (2006). *Developing literacy in second-language learners: Report of the national literacy panel on language minority children and youth*. Mahwah, NJ: Lawrence Erlbaum Associates.

*Bailey, B., Arciuli, J., & Stancliffe, R. J. (2016). Effects of ABRACADABRA literacy instruction on children with Autism Spectrum Disorder. *Journal of Educational Psychology, 109*(2), 257-268. http://dx.doi.org/10.1037/edu0000138

*Bailey, B., Arciuli, J., & Stancliffe, R. J. (2017). Effects of ABRACADABRA instruction on spelling in children with Autism Spectrum Disorder. *Scientific Studies of Reading, 21*(2), 146-164. http://dx.doi.org/10.1080/10888438.2016.1276183

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2014). *Comprehensive Meta-analysis Version 3*. Englewood, NJ: Biostat.

Bunyi, G. W. (2005). Language classroom practices in Kenya. In A. M. Y. Lin & P. W. Martin (Eds.), *Decolonisation, Globalisation: Language-in-Education Policy and Practice* (pp. 131-152). Clevedon: Multilingual Matters. Retrieved from https://eric.ed.gov/?id=ED489857

*Cheung, A., Mak, B. S. Y., Abrami, P. C., Wade, A., & Lysenko, L. (2016). The effectiveness of the ABRACADABRA (ABRA) web-based literacy program on primary school students in Hong Kong. *Journal of Interactive Learning Research, 27*(3), 219-245. Retrieved from https://www.learntechlib.org/p/151937

Cheung, A., & Slavin, R. E. (2012). How features of educational technology programs affect student reading outcomes: A meta-analysis. *Educational Research Review, 7*, 198-215. http://dx.doi.org/10.1016/j.bbr.2011.03.031

*Cheung, A. C. K., & Guo, X. (2017). *Examining the effects of ABRACADABRA (ABRA), a web-based literacy program on primary school students in rural China* (Digital Learning for Development (DL4L) final research report). Retrieved from DL4D website: http://dl4d.org/portfolio-items/examining-the-effects-of-abracadabra/

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Earlbaum.

*Comaskey, E., Savage, R., & Abrami, P. C. (2009). A randomized efficacy study of web-based synthetic and analytic programmes among disadvantaged urban kindergarten children. *Journal of Research in Reading, 32*(1), 92-108. http://dx.doi.org/10.1111/j.1467-9817.2008.01383.x

*Deault, L., Savage, R., & Abrami, P. C. (2009). Inattention and response to the ABRACADABRA web-based literacy intervention. *Journal of Research on Educational Effectiveness, 2*, 250-286. http://dx.doi.org/10.1080/19345740902979371

*Di Stasio, M., Savage, R., & Abrami, P. C. (2012). A follow-up study of the ABRACADABRA web-based literacy intervention in grade1. *Journal of Research in Reading, 35*(1), 69-86. http://dx.doi.org/10.1111/j.1467-9817.2010.01469.x

Dodd, B., Crosbie, S., McIntosh, B., Teitzel, T., & Ozanne, A. (2000). *Preschool and Primary Inventory of Phonological Awareness (PIPA)*. Oxford, UK: Pearson Assessment.

Dunn, L., Dunn, L., Styles, B., & Sewell, J. (2009). *British Picture Vocabulary Scales* (3rd ed.). Windsor, UK: Granada.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455-463. http://dx.doi.org/10.1111/j.0006-341X.2000.00455.x

Forum for Research in Literacy and Language. (2012). *Diagnostic test of reading processes* (measurement instrument). London: GL Assessment.

Fry, E. (1980). The new instant word list. *The Reading Teacher, 34*(3), 284-289.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

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

Hedges, L. V., Shymansky, J. A., & Woodworth, G. (1989). *A practical guide to modern methods of meta-analysis*.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*(7414), 557-560. http://dx.doi.org/10.1136/bmj.327.7414.557

Larsen, L., Kohnen, S., Nickels, L., & McArthur, G. (2015). The Letter-Sound Test (LeST): a reliable and valid comprehensive measure of grapheme–phoneme knowledge. *Australian Journal of Learning Difficulties, 20*(2), 129-142. http://dx.doi.org/10.1080/19404158.2015.1037323

Lysenko, L, Abrami, P. C., & Wade, A. (Under review). Promoting young Kenyans' growth in literacy with educational technology: A tale of two years of implementation. *International Journal of Educational Research*.

*Mak, B. S. Y., Cheung, A., Guo, X., Abrami, P. C., & Wade, A. (2017). Examining the impact of of the ABRACADABRA (ABRA) web-based literacy program on primary school students in Hong Kong. *Education and Information Technoogies, 22*(6), 2671-2691. http://dx.doi.org/10.1007/s10639-017-9620-3

Mayer, R. E. (2008). Applying the science of learning: Evidence-based principles for the design of multimedia instruction. *American Psychologist, 63*(8), 760-769. http://dx.doi.org/10.1037/0003-066X.63.8.760

*McNally, S., Ruiz-Valenzuela, J., & Rolfe, H. (2016). *ABRA: Online reading support. Evaluation report and executive summary. Addendum added March 2018*. Retrieved from Education Endowment Foundation website: https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Project_Report_ABRA.pdf

Merton, R. K. (1968). The Matthew Effect in science. *Science*, *159*(3810), 56-63. http://dx.doi.org/10.1126/science.159.3810.56

National Reading Panel (NRP). (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Pub. No.00-4754). Retrieved from National Institute of Child Health & Human Development website website: http://www.nichd.nih.gov/publications/pubs/nrp/pages/report.aspx

Neale, M. D. (1999). *Neale Analysis of Reading Ability* (3rd ed.). Melbourne, Vic: Australian Council for Educational Research.

*Piquette, N. A., Savage, R. S., & Abrami, P. (2014). A cluster randomized control field trial of the ABRACADABRA web-based literacy intervention: Replication and extension of basic findings. *Frontiers in Psychology, 5*, Article 1413. http://dx.doi.org/10.3389/fpsyg.2014.01413

Pressley, M. (2002). Effective beginning reading instruction. *Journal of Literacy Research, 34*, 165-188. http://dx.doi.org/10.1207/s15548430jlr3402_3

Pressley, M., Wharton-McDonald, R., Allington, R., Block, C. C., Morrow, L., Tracey, D., . . . Woo, D. (2001). A study of effective first grade literacy instruction. *Scientific Studies of Reading, 5*, 35-58. http://dx.doi.org/10.1207/S1532799XSSR0501_2

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638-664. http://dx.doi.org/10.1037/0033-2909.86.3.638

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester, UK: Wiley.

*Savage, R., Abrami, P. C., Hipps, G., & Deault, L. (2009). A randomized control trial study of the ABRACADABRA reading intervention program in Grade 1. *Journal of Educational Psychology, 101*, 590-604. http://dx.doi.org/10.1037/a0014700

*Savage, R., Abrami, P. C., Piquette-Tomei, N., Wood, E., Deleveaux, G., Sanghera-Sidhu, B., & Burgos, G. (2013). A (pan-Canadian) cluster randomised control effectiveness trial of the ABRACADABRA web-based literacy program. *Journal of Educational Psychology, 105*(2), 310-328. http://dx.doi.org/10.1037/a0031025

*Savage, R., Erten, O., Abrami, P., Hipps, G., Comaskey, E., & van Lierop, D. (2010). ABRACADABRA in the hands of teachers: The effectiveness of a web-based literacy intervention in grade 1 language arts programs. *Computers & Education, 55*(2), 911-922. http://dx.doi.org/10.1016/j.compedu.2010.04.002

Savage, R. S., & Pompey, Y. (2008). What does the evidence really say about effective literacy teaching? *Educational and Child Psychology, 25*, 21-30.

Schmid, R. F., Bernard, R. M., Borokhovski, E., Tamim, R. M., Abrami, P. C., Surkes, M. A., Wade, C. A., Woods, J. . (2014). The effects of technology use in postsecondary education: A meta-analysis of classroom applications. *Computers & Education, 72*, 271-291. http://dx.doi.org/10.1016/j.compedu.2013.11.002

Slavin, R. E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). Effective reading programs for the elementary grades: A best-evidence synthesis. *Review of Educational Research, 79*(4), 1391-1465.

Spencer, M., Muse, A., Wagner, R. K., Foorman, B., Petscher, Y., Schatschneider, C., ... & Bishop, M. D. (2015). Examining the underlying dimensions of morphological awareness and vocabulary knowledge. *Reading and Writing*, 28(7), 959-988. https://doi.org/10.1007/s11145-015-9557-0

Takacs, Z. K., Swart, E. K., & Bus, A. G. (2015). Benefits and pitfalls of multimedia and interactive features in technology-enhanced storybooks: A meta-analysis. *Review of Educational Research, 85*, 698-739. http://dx.doi.org/10.3102/0034654314566989

Tamim, R., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning: A second-order meta-analysis and validation study. *Review of Educational Research, 81*(1), 4-28. http://dx.doi.org/10.3102/0034654310393361

Taylor, B. M., Pearson, P. D., Clark, K., & Walpole, S. (2000). Effective Schools and accomplished teachers: Lessons about primary grade reading instruction in low income schools. *Elementary School Journal, 101*(2), 121-165. http://dx.doi.org/10.1086/499662

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *The Comprehensive Test of Phonological Processing (CTOPP)*. Austin, TX: Pro-Ed.

Wilkinson, G. S., & Robertson, G. J. (2006). *Wide Range Achievement Test (WRAT4)*. Lutz, FL: Psychological Assessment Resources.

Williams, K. T. (2001). *Group Reading Assessment and Diagnostic Evaluation (GRADE)*. Circle Pines, MN: American Guidance Service (AGS, Inc.).

Wolgemuth, J., Ehrich, J., Emmett, S., Helmer, J., Bartlett, C., Smith, H., . . . Deleveaux, G. (2009). *ABRACADABRA! (ABRA) Early Childhood Literacy Project: A pilot study of the feasibility of the ABRA literacy software in Northern Territory indigenous classrooms* (Annual Report No.1). Retrieved from Charles Darwin University, Darwin, NT. website: https://trove.nla.gov.au/work/198852924?q&versionId=217791122

Wolgemuth, J., Helmer, J., Emmett, S., Bottrell, C., Lea, T., Bartlett, C., . . . Savage, R. (2010). *ABRACADABRA! (ABRA) early childhood literacy project: A quasi-experimental study of the ABRA literacy software in Northern Territory indigenous classrooms* (Annual Report No.2). Retrieved from Charles Darwin University website: https://trove.nla.gov.au/work/198852920?q&versionId=217791105

*Wolgemuth, J., Savage, R., Helmer, J., Harper, H., Lea, T., Abrami, P. C., . . . Louden, W. (2013). ABRACADABRA aids indigenous and non-indigenous early literacy in Australia: Evidence from a multisite randomized controlled trial. *Computers & Education, 67*, 250-264. http://dx.doi.org/10.1016/j.compedu.2013.04.002

*Wolgemuth, J., Savage, R., Helmer, J., Lea, T., Harper, H., Chalkiti, K., . . . Abrami, P. (2011). Using computer-based instruction to improve Indigenous early literacy in Northern Australia: A quasi-experimental study. *Australasian Journal of Educational Technology, 27*(727-750). Retrieved from http://www.ascilite.org.au/ajet/ajet27/wolgemuth.pdf

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock Johnson III Tests of Achievement*. Rolling Meadows, IL: Riverside

Table 1.
Basic Description of ABRA Studies Reviewed for the Meta-Analysis

Study number	Design	Length of intervention	N	Grade level	Sample characteristics	ABRA delivery	Outcome type	Publication
1	Cluster randomized control trial	10-12 weeks (120 min per week)	1067 students (549E, 518C)	K (316 students) 1(616 students) 2(135 students)	Schools in Alberta, Ontario, Quebec, Canada	ELA teacher	• Phonics • Phonemic awareness • Fluency • Reading comprehension • Listening comprehension	Savage, Abrami, Piquette-Tomei et al. (2013)
2	Randomized control trials	10 weeks (80 min per week)	143 students (86E, 57C)	1	Urban schools in Quebec, Canada	Researcher	• Phonics • Phonemic awareness • Fluency • Reading comprehension • Listening comprehension • Vocabulary	Deault, Savage & Abrami (2009) Savage, Abrami, Hipps & Deault (2009)

3	Quasi-experiment	8 weeks (120 min per week)	60 students (49E, 11C)	1	Suburban schools in Quebec, Canada	ELA teacher and facilitator	• Phonics • Phonemic awareness • Vocabulary • Reading comprehension • Listening comprehension	Savage et al. (2010)
4	Quasi-experiment		312 students (119E, 193C)	K(4 classes) 1(4 classes) Split 1-2 (2 classes)	Schools in Alberta, Ontario, Quebec, Canada	ELA Teacher	• Phonemic awareness • Fluency	Anderson et al. (2011)
5	Quasi-experiment	16 weeks (120 min per week)	166 students (118E, 48C)	Transition (4 classes) Transition to year 2 (1 class) 1(3 classes) Split 1-2 (5 classes) Split 2-3(1 class)	Schools in Northern Territories, Australia	ELA Teacher	• Phonics • Phonemic awareness • Vocabulary	Wolgemuth et al. (2011) Wolgemuth et al. (2010)

6	Randomized control trial	16 weeks (120 min per week)	308 students (163E, 145C)	Transition (2 classes) Transition to year 1 (9 classes) 1(1 class) Split 1-2 (5 classes)	4 provincial and 2 remote schools in Northern Territories, Australia	Tutor	• Phonics • Phonemic awareness • Vocabulary	Wolgemuth, Savage, Helmer et al. (2013)
7	Cluster randomized control trial	14 weeks (90 min per week)	358 students (180E, 178C)	Standard 2 (12 classes)	6 urban schools, Kenya	ELA Teacher	• Vocabulary • Reading comprehension • Listening comprehension	Abrami, Wade, Lysenko, Marsh, & Gioko (2014)
8	Pre-experiment	10 weeks (120 min per week)	98 students	Transition (2 classes) 1 (9 classes) 2 (1 class) 3 (5 classes)	1 provincial and 2 remote schools in Northern Territories, Australia	ELA Teacher	• Phonics • Phonemic awareness • Vocabulary • Reading comprehension • Listening comprehension	Wolgemuth, Enrich, Helmer et al. (2009)[1]

9	Two-year randomized control trial	Year 1: 10 hours Year 2: 10 hours	43 (analytic phonics: 27; synthetic phonics: 26) Both groups used ABRA	K (43 students) 1 (43 students)	One urban school in Quebec, Canada	Facilitator	• Phonics • Phonemic awareness • Fluency • Reading comprehension • Listening comprehension	Comaskey, Savage & Abrami (2009)[2] Di Stasio, Savage, Abrami (2012)[2]
10	Cluster randomized control trial	20 hours of exposure	203 students (105E, 98C)	K (107 students) 1 (96 students) (24 classes)	Schools in a rural northern district in Alberta, Canada	ELA Teacher	• Phonics • Phonemic awareness • Fluency • Vocabulary • Listening comprehension	Piquette, Savage, & Abrami (2014)
11	Cluster randomized control trial	14 weeks	122 students (74E, 48C)	2 (5 classes)	Hong Kong English school	Three local English teachers and one native English teacher	• Phonics • Reading comprehension • Listening comprehension	Cheung, Mak, Abrami, Wade, & Lysenko (2016)
12	Randomized control trial	13 weeks (26 hours of exposure)	20 students (11E, 9C)	5-11 years old	Australia, students diagnosed with autistic spectrum disorder	One speech pathologist, one-to-one instruction	• Phonics • Fluency • Reading comprehension	Bailey, Arciuli & Stancliff (2016); Bailey, Arciuli & Stancliff (2017)[3]

13	Quasi-experiment	20 weeks	249 students (145E,104C)	1 (10 classes)	Hong Kong, English school	ESL teachers	• Phonics • Phonemic awareness • Listening comprehension	Mac, Cheung, Abrami &Wade (2017)
14	Nested cluster randomization	20 weeks (20 hours of exposure)	2,241 students (710E, 1158C in control schools + 373 in exp. schools)	1-3	The United Kingdom, 10 areas	Trained teacher assistants	• Phonics • Phonemic awareness • Vocabulary • Reading comprehension	McNally, Ruiz-Valenzuella, & Rolfe (2016, 2018)[4]
15	Quasi-experiment	20 weeks	749 students (307E, 393C)	1-3 (22 classes)	Kenya, ESL context	EL teachers	• Vocabulary • Reading comprehension • Listening comprehension	Lysenko, Abrami, & Wade (under review, phase 1)
16	Quasi-experiment	20 weeks	1,015 students (498E, 474C)	3 (27 classes)	Kenya, ESL context	EL teachers	• Vocabulary • Reading comprehension • Listening comprehension	Lysenko, Abrami, & Wade (under review, phase 2)
17	Quasi-experiment	30 weeks	375 students (227E, 148C)	3	Rural China	EFL teachers	• Phonics • Phonemic awareness	Cheung & Xin Guo (2017)

[1] Results are not included as the study was designed as one-group pretest/posttest comparison (pre-experimental design)

[2] Results are excluded from the analyses as in both conditions the instruction was ABRA-based (synthetic vs. analytic). No non-ABRA control group was used.

[3] Results are not included as the reported outcome types were not directly related to reading.
[4] Delayed post-test data reported in an addendum to the 2016 paper were not accounted for in the analyses.

Table 2
Overall Weighted Average Effect Size (Random Effects Model): Adjusted and Unadjusted Data for ***Phonics*** *Outcome Type and Heterogeneity Statistics (Fixed Effect Model, Unadjusted Effects Only)*

Population Estimates	*k*	*g*+	*SE*	Lower 95th	Upper 95th
Unadjusted Effects	23	0.187**	0.07	0.05	0.32
Adjusted Effects	23	0.263**	0.04	0.19	0.34
** $p = .006$					
Heterogeneity Analysis	$Q_T = 44.42$ ($df = 22$), $p = .003$, $I^2 = 50.47$				

Table 3
Overall Weighted Average Effect Size (Random Effects Model): Adjusted and Unadjusted Data for ***Phonemic Awareness*** *Outcome Type and Heterogeneity Statistics (Fixed Effect Model, Unadjusted Effects Only)*

Population Estimates	*k*	*g*+	*SE*	Lower 95th	Upper 95th
Unadjusted Effects	23	0.378***	0.04	0.29	0.46
Adjusted Effects	23	0.299***	0.04	0.23	0.37
*** $p < .001$					
Heterogeneity Analysis	$Q_T = 25.45$ ($df = 22$), $p = .28$ (ns), $I^2 = 13.56$				

Table 4
Overall Weighted Average Effect Size (Random Effects Model): Adjusted and Unadjusted Data for ***Reading Fluency*** *Outcome Type and Heterogeneity Statistics (Fixed Effect Model, Unadjusted Effects Only)*

Population Estimates	*k*	*g*+	*SE*	Lower 95th	Upper 95th
Unadjusted Effects	7	0.088 (ns)	0.10	-0.11	0.28
Adjusted Effects	7	0.181**	0.06	0.06	0.30
** $p = .004$					
Heterogeneity Analysis	$Q_T = 18.95$ ($df = 6$), $p = .004$, $I^2 = 68.34$				

Table 5
Overall Weighted Average Effect Size (Random Effects Model): Adjusted and Unadjusted Data for ***Reading Comprehension*** *Outcome Type and Heterogeneity Statistics (Fixed Effect Model, Unadjusted Effects Only)*

Population Estimates	*k*	*g*+	*SE*	Lower 95th	Upper 95th

Unadjusted Effects	11	0.180**	0.07	0.04	0.32
Adjusted Effects	11	0.244**	0.04	0.16	0.33
** p = .01; *** p < .001					
Heterogeneity Analysis	Q_T = 37.04 (df = 10), p < .001, I^2 = 73.00				

Table 6

Overall Weighted Average Effect Size (Random Effects Model): Adjusted and Unadjusted Data for ***Listening Comprehension*** *Outcome Type and Heterogeneity Statistics (Fixed Effect Model, Unadjusted Effects Only)*

Population Estimates	*k*	*g+*	*SE*	Lower 95th	Upper 95th
Unadjusted Effects	9	0.274*	0.12	0.05	0.50
Adjusted Effects	9	0.313**	0.08	0.16	0.47
* p = .02; *** p < .001;					
Heterogeneity Analysis	Q_T = 60.72 (df = 8), p < .001, I^2 = 89.47				

Table 7

Overall Weighted Average Effect Size (Random Effects Model): Adjusted and Unadjusted Data for ***Vocabulary Knowledge*** *Outcome Type and Heterogeneity Statistics (Fixed Effect Model, Unadjusted Effects Only)*

Population Estimates	*k*	*g+*	*SE*	Lower 95th	Upper 95th
Unadjusted Effects	18	0.080 (ns)	0.10	-0.11	0.27
Adjusted Effects	18	0.183**	0.07	0.05	0.31
** p = .005;					
Heterogeneity Analysis	Q_T = 97.20 (df = 17), p < .001, I^2 = 82.51				

Table 8.

Vote count summary of effect sizes for the within-study factors

Within-study factors	Vote count (ABRA vs. control)			Source of the effects and effect sizes (g)[1]	Total weighted effect size and sample size
	Positive effect (>0.1)	Trivial effect	Negative effect (< -0.1)		
BASELINE READING ABILITY:					
Advanced readers: ABRA vs. control students	2	2		Abrami et al. (2014): d=0.09(VK); d=0.26 (RC); d=0.52(LC) *McNally et al. (2016): β=0.06 (PIRA)* [1, 3]	g=0.29 (N_E=61; N_C=61)
Low readers: ABRA vs. control students	5	1	1	Abrami et al. (2014): d=0.08 (VK); d=0.17 (RC); d=0.43 (LC) Lysenko et al. (under review, phase 2): d=-0.19 (VK); d=0.54 (RC); d=0.29 (LC) *McNally et al. (2016): β=0.22 (RC)* [1, 3]	g=0.21 (N_E=157; N_C=274)
GENDER:					
ABRA boys vs. control boys	9	1		Abrami et al. (2014): d=0.23 (VK); d=0.54 (RC); d=0.40 (LC) Lysenko et al. (under review, phase 1): d=0.41 (VK); d=0.35 (RC); d=0.34 (RC) Lysenko et al. (under review, phase 2): d=-0.08 (VK); d=0.17 (RC); d=0.14 (LC) *McNally et al. (2016): β=0.14 (RC)* [1, 3]	g=0.23 (N_E=407; N_C=520)
ABRA girls vs. control girls	6	3	1	Abrami et al. (2014): d=-0.03 (VK); d=0.13 (RC); d=0.48 (LC) Lysenko et al. (under review, phase 1): d=0.73 (VK); d=0.84 (RC); d=1.17 (RC) Lysenko et al. (under review, phase 2): d=-0.14 (VK); d=0.07 (RC); d=-0.06 (LC) *McNally et al. (2016): β=0.14 (RC)* [1, 3]	g=0.30 (N_E=578; N_C=521)
ECONOMIC STATUS:					
Eligible for free meal (EFM) ABRA vs. control students	1			*McNally et al. (2016): β=0.37 (RC)* [1, 3]	
Non-EFM ABRA vs. control students		1		*McNally et al. (2016): β=0.08 (RC)* [1, 3]	
INDIGENEOUS STATUS:					
Indigenous ABRA vs. indigenous control students	2	4		Wolgemuth et al. (2011): d=0.94 (PA); d=0.01(P); d=-0.03 (VK) Wolgemuth et al. (2013): d=0.80 (PA); d=-0.05 (P); d=0.08 (VK)	g=0.16 (N_E=103; N_C=72)
Non-indigenous ABRA vs. non-indigenous control students	5	1		Wolgemuth et al. (2011): d=0.64 (PA); d=0.06 (P); d=0.12 (VK) Wolgemuth et al. (2013): d=0.23 (PA); d=0.30 (P); d=0.24 (VK)	g= 0.26 (N_E=172; N_C=114)

PRIOR EXPERIENCE WITH ABRA:					
New ABRA vs. control students	2	1		Lysenko et al. (under review, phase 2): d=0.01 (VK); d=0.20 (RC); d=0.24 (LC)	g=0.15 (N_E=160; N_C=237)
Students from classes with prior ABRA exposure vs. control students	1	1	1	Lysenko et al. (under review, phase 2): d=-0.16 (VK); d=0.15 (RC); d=0.06 (LC)	g=0.02 (N_E=338; N_C=237)
FIDELITY OF ABRA IMPLEMENTATION:					
Higher implementation of ABRA vs. control	5	3		Savage et al. (2010)[2]: d=0.44 (PA); d=0.64 (VK); d=0.61 (RC) Savage et al. (2013): d=0.61(PA); d=0.33 (P); d= 0.01 (F); d=-0.06 (RC); d=-0.03 (LC)	g=0.21 (N_E=168; N_C=365)
Lower implementation of ABRA vs. control	6	1	1	Savage et al. (2010)[2]: d=0.34 (PA); d=0.52 (VK); d=0.37 (RC) Savage et al. (2013): d=0.23 (PA); d=0.21 (P); d= -0.19 (F); d=0.01 (RC); d=0.27 (LC)	g=0.14 (N_E=316; N_C=365)

PA – Phonemic Awareness
P – Phonics
VK – Vocabulary Knowledge
F – Fluency
RC – Reading Comprehension
LC – Listening Comprehension

[1] OLS (Ordinary Least Squares) based standardized coefficients as indicated in the respective reports; not included in the total weighted effect size calculation

[2] Effect sizes in the categories of P and LC are not included in this vote count due to the inconsistencies in data reporting

[3] Sample size split is not reported

Table 9.
Vote count summary of effect sizes for the between-study factors

Between – study factors	Vote count (E vs. C) Positive effect (>0.1)	Trivial effect	Negative effect (< -0.1)	Source of the effects and effect sizes (Hedges' *g*)	Total weighted effect size and sample sizes
COUNTRY:					
Canada: ABRA students vs. control students	26	7	5	Savage et al. (2009): PA (0.23; 0.39; 0.19; 0.32); P (0.08; 0.12; 0.42; 0.14); VK (0.05; -0.04); F (-0.06; -0.13); RC (0.29; 0.16); LC (0.24; 0.34) Savage et al. (2010): PA (0.45; 0.67); P (-0.38; -0.13); VK (0.66); RC (0.60); LC (0.16) Anderson et al. (2011): PA (0.43); F (0.47) Savage et al. (2013): PA (0.36; 0.10); P (0.26); F (0.13; -0.10); RC (0.03; 0.05); LC (0.39) Piquette, Savage & Abrami (2014): PA (0.33); P (0.51); VK (0.66); F (0.04); LC (-0.44)	g=0.20 (N_E=440; N_C=493)
Kenya: ABRA students vs. control students	6	2	1	Abrami et al. (2014): VK (-0.01); RC (0.26); LC (0.54) Lysenko et al. (under review, phase 1): VK (0.49); RC (0.55); LC (0.69) Lysenko et al. (under review, phase 2): VK (-0.11); RC (0.11); LC (0.03)	g=0.24 (N_E= 985; N_C=1,041)
Australia: ABRA students vs. control students	22	5	6	Wolgemuth et al. (2011): PA (0.75; 0.73; 0.35; 0.25); P (0.16; 0.18; -0.78; 0.20); VK (0.04; -0.34; -0.77; 0.04) Wolgemuth et al. (2013): PA (0.16; 0.17; 1.04; 0.37; 0.24; 0.32); P (0.19; 0.35; 0.45; 0.03; 0.09; -0.26); VK (0.32; 0.08; -0.12; -0.27; 0.16; 0.38) Bailey, Arciuli & Stancliff (2017)[2]: P (0.28); F (0.29); RC (0.39)	g=0.17 (N_E= 292; N_C=202)
China and Hong Kong: ABRA students vs. control students	4	3	1	Cheung et al. (2016): P (-0.06); RC (-0.11); LC (0.03) Mak et al. (2017): PA (0.17); P (0.10); LC (-0.07) Cheung & Guo (2017): PA (0.46); P (0.72)	g=0.24 (N_E= 440; N_C=291)
The United Kingdom: ABRA students vs. control students	3	1		McNally et al. (2016): PA (0.20); P (0.18); VK (-0.05); RC (0.17)	g=0.13 (N_E= 307; N_C=1,103)
ABRA INSTRUCTION DELIVERY:					

ABRA students taught by teacher vs. control students	25	10	7	Wolgemuth et al. (2011): PA (0.75; 0.73; 0.35; 0.25); P (0.16; 0.18; -0.78; 0.20); VK (0.04; -0.34; -0.77; 0.04) Savage et al. (2013): PA (0.36; 0.10); P (0.26); F (0.13; -0.10); RC (0.03; 0.05); LC (0.39) Abrami et al. (2014): VK (-0.01); RC (0.26); LC (0.54) Piquette, Savage & Abrami (2014): PA (0.33); P (0.51); VK (0.66); F (0.04); LC (-0.44) Cheung et al. (2016): P (-0.06); RC (-0.11); LC (0.03) Mak et al. (2017): PA (0.17); P (0.10); LC (-0.07) Cheung & Guo (2017): PA (0.46); P (0.72) Lysenko et al. (under review, phase 1): VK (0.49); RC (0.55); LC (0.69) Lysenko et al. (under review, phase 2): VK (-0.11); RC (0.11); LC (0.03)	g=0.22 (N_E= 1,802; N_C=1,649)
ABRA students taught by other (e.g. researcher, tutor, teacher assistant, etc.) vs. control students	36	8	6	Savage et al. (2009): PA (0.23; 0.39; 0.19; 0.32); P (0.08; 0.12; 0.42; 0.14); VK (0.05; -0.04); F (-0.06; -0.13); RC (0.29; 0.16); LC (0.24; 0.34) Savage et al. (2010): PA (0.45; 0.67); P (-0.38; -0.13); VK (0.66); RC (0.60); LC (0.16) Anderson et al. (2011): PA (0.43); F (0.47) Wolgemuth et al. (2013): PA (0.16; 0.17; 1.04; 0.37; 0.24; 0.32); P (0.19; 0.35; 0.45; 0.03; 0.09; -0.26); VK (0.32; 0.08; -0.12; -0.27; 0.16; 0.38) McNally et al. (2016): PA (0.20); P (0.18); VK (-0.05); RC (0.17) Bailey, Arciuli & Stancliff (2017)[2]: P (0.28); F (0.29); RC (0.39)	g=0.17 (N_E= 662; N_C=1,490)
INSTRUCTOR-STUDENT RATIO:					
One-on-one ABRA vs. control students	3			Bailey, Arciuli & Stancliff (2017)[2]: P (0.28); F (0.29); RC (0.39)	g=0.32 (N_E= 11; N_C=9)
10 or fewer ABRA students per instructor vs. control students	21	10	7	Savage et al. (2009): PA (0.23; 0.39; 0.19; 0.32); P (0.08; 0.12; 0.42; 0.14); VK (0.05; -0.04); F (-0.06; -0.13); RC (0.29; 0.16); LC (0.24; 0.34) Anderson et al. (2011): PA (0.43); F (0.47) Wolgemuth et al. (2011): PA (0.35; 0.25); P (-0.78; 0.20); VK (-0.77; 0.04) Wolgemuth et al. (2013):	g= 0.17 (N_E= 637; N_C=1,465)

				PA (0.17; 1.04; 0.37; 0.24; 0.32); P (0.35; 0.45; 0.03; 0.09; -0.26); VK (0.08; -0.12; -0.27; 0.16; 0.38) Piquette, Savage & Abrami (2014): PA (0.33); P (0.51); VK (0.66); F (0.04); LC (-0.44) McNally et al. (2016): PA (0.20); P (0.18); VK (-0.05); RC (0.17)	
11 and more ABRA students per instructor vs. control students	27	8	6	Savage et al. (2010): PA (0.45; 0.67); P (-0.38; -0.13); VK (0.66); RC (0.60); LC (0.16) Wolgemuth et al. (2011): PA (0.75; 0.73); P (0.16; 0.18); VK (0.04; -0.34;) Savage et al. (2013): PA (0.36; 0.10); P (0.26); F (0.13; -0.10); RC (0.03; 0.05); LC (0.39) Wolgemuth et al. (2013): PA (0.16); P (0.19); VK (0.32) Abrami et al. (2014): VK (-0.01); RC (0.26); LC (0.54) Cheung et al. (2016): P (-0.06); RC (-0.11); LC (0.03) Mak et al. (2017): PA (0.17); P (0.10); LC (-0.07) Cheung & Guo (2017): PA (0.46); P (0.72) Lysenko et al. (under review, phase 1): VK (0.49); RC (0.55); LC (0.69) Lysenko et al. (under review, phase 2): VK (-0.11); RC (0.11); LC (0.03)	g=0.22 (N_E= 1,668; N_C=1,552)
STUDENT EXPOSURE TO ABRA INSTRUCTION:					
ABRA students exposed to ABRA less than 20 hrs vs. control students	24	9	6	Savage et al. (2009): PA (0.23; 0.39; 0.19; 0.32); P (0.08; 0.12; 0.42; 0.14); VK (0.05; -0.04); F (-0.06; -0.13); RC (0.29; 0.16); LC (0.24; 0.34) Savage et al. (2010): PA (0.45; 0.67); P (-0.38; -0.13); VK (0.66); RC (0.60); LC (0.16) Wolgemuth et al. (2013): PA (0.37); P (0.03); VK (-0.27) Abrami et al. (2014): VK (-0.01); RC (0.26); LC (0.54) Piquette, Savage & Abrami (2014): PA (0.33); P (0.51); VK (0.66); F (0.04); LC (-0.44) Cheung et al. (2016): P (-0.06); RC (-0.11); LC (0.03) Cheung & Guo (2017): PA (0.46); P (0.72)	g=0.25 (N_E= 681; N_C=531)

ABRA students exposed to ABRA more than 20 hrs vs. control students	38	10	8	Anderson et al. (2011): PA (0.43); F (0.47) Wolgemuth et al. (2011): PA (0.75; 0.73; 0.35; 0.25); P (0.16; 0.18; -0.78; 0.20); VK (0.04; -0.34; -0.77; 0.04) Savage et al. (2013): PA (0.36; 0.10); P (0.26); F (0.13; -0.10); RC (0.03; 0.05); LC (0.39) Wolgemuth et al. (2013): PA (0.16; 0.17; 1.04; 0.24; 0.32); P (0.19; 0.35; 0.45; 0.09; -0.26); VK (0.32; 0.08; -0.12; 0.16; 0.38) McNally et al. (2016): PA (0.20); P (0.18); VK (-0.05); RC (0.17) Mak et al. (2017): PA (0.17); P (0.10); LC (-0.07) Bailey, Arciuli & Stancliff (2017)[2]: P (0.28); F (0.29); RC (0.39) Lysenko et al. (under review, phase 1): VK (0.49); RC (0.55); LC (0.69) Lysenko et al. (under review, phase 2): VK (-0.11); RC (0.11); LC (0.03)	g=0.19 (N_E= 1,612; N_C=2,485)
GRADE LEVEL[2]:					
Kindergarten ABRA vs. Kindergarten control students	3	1		McNally et al. (2016): PA (0.20); P (0.18); VK (-0.05); RC (0.17)	g=0.13 (N_E= 307; N_C=1,103)
Grade 1 ABRA vs. Grade 1 control students	20	6	3	Savage et al. (2009): PA (0.23; 0.39; 0.19; 0.32); P (0.08; 0.12; 0.42; 0.14); VK (0.05; -0.04); F (-0.06; -0.13); RC (0.29; 0.16); LC (0.24; 0.34) Savage et al. (2010): PA (0.45; 0.67); P (-0.38; -0.13); VK (0.66); RC (0.60); LC (0.16) Mak et al. (2017): PA (0.17); P (0.10); LC (-0.07) Lysenko et al. (under review, phase 1): VK (0.78); RC (-0.02); LC (0.90)	g=0.22 (N_E= 290; N_C=242)
Grade 2 ABRA vs. Grade 2 control students	5	3	1	Abrami et al. (2014): VK (-0.01); RC (0.26); LC (0.54) Cheung et al. (2016): P (-0.06); RC (-0.11); LC (0.03) Lysenko et al. (under review, phase 1): VK (0.85); RC (1.46); LC (0.53)	g=0.43 (N_E= 378; N_C=373)
Grade 3 ABRA vs. Grade 3 control students	6	1	1	Cheung & Guo (2017): PA (0.46); P (0.72) Lysenko et al. (under review, phase 1): VK (0.36); RC (0.56); LC (1.22) Lysenko et al. (under review, phase 2): VK (-0.11); RC (0.11); LC (0.03)	g=0.20 (N_E= 819; N_C=800)

PA – Phonemic Awareness
P – Phonics
VK – Vocabulary Knowledge

F – Fluency
RC – Reading Comprehension
LC – Listening Comprehension

[1] ABRA mixed grade levels are not accounted for

[2] Special needs ASD-diagnosed students