Comparing evidence on the effectiveness of reading resources from expert ratings, practitioner judgements, and research repositories

Fiona M. Hollands, fmh7@tc.columbia.edu
Teachers College, Columbia University, USA

Yilin Pan, ypan@worldbank.org
World Bank, USA

Michael J. Kieffer, michael.kieffer@nyu.edu
New York University, USA

Venita R. Holmes, vholmes@houstonisd.org
Houston Independent School District, USA

Yixin Wang, yixwang@unicef.org
UNICEF, New York, USA

Maya Escueta, mme17@tc.columbia.edu
Laura Head, leh2179@tc.columbia.edu
Atsuko Muroga, am3994@tc.columbia.edu
Teachers College, Columbia University, USA

Background: Education decision makers are increasingly expected to use evidence to inform their actions. However, the majority of educational interventions have not yet been studied and it is challenging to produce high quality research evidence quickly enough to influence policy questions.

Aims and objectives: We set out to gather evidence on the efficacy of reading resources implemented at 23 struggling elementary schools in a large, urban district in the US. The schools were at risk of closure by the state.

Methods: For each reading resource, we searched for existing effectiveness studies and collected professional judgements by surveying practitioners. We also used an expert survey to collect judgements from three reading experts. We compared the ratings among experts and between practitioners and experts. We also compared practitioner and expert judgements to evidence summaries from research repositories.

Findings: We found evidence summaries in research repositories for only five of 23 reading resources used in the 23 schools. Experts showed poor to good agreement on ten questions about each resource. Agreement between practitioners and experts was low with practitioners generally rating resources more positively than reading experts.

Discussion and conclusions: Practitioners may be overly optimistic about the efficacy of educational materials while experts have difficulty assessing how well the materials serve an unfamiliar population. In the absence of rigorous evaluations of locally-implemented programmes, district
decision makers can review the consistency of evidence collected from practitioners and experts, along with external research evidence, to inform actions towards supporting and guiding struggling schools.

Key words  evidence use in school districts • practitioner judgements • expert survey • reading proficiency

Key messages
• Rigorous research evidence is lacking for many of the programmes and practices implemented in schools.
• Districts can produce internal evidence on the effectiveness of reading resources by surveying practitioners.
• Expert ratings of the same resources are generally less optimistic about their efficacy.
• Decision makers should consider consistency across sources of evidence and relevance to local context.

To cite this article: Hollands, F., Pan, Y., Kieffer, M., Holmes, V., Wang, Y., Escueta, M., Head, L. and Muroga, A. (2021) Comparing evidence on the effectiveness of reading resources from expert ratings, practitioner judgements, and research repositories, Evidence & Policy, vol XX, no XX, 1–20, DOI: 10.1332/174426421X16366418828079

Introduction

Evidence use in education

Education decision makers are increasingly expected to use rigorous evidence to inform decisions about the programmes, practices and interventions implemented in schools (National Commission on Excellence in Education, 1983; No Child Left Behind Act of 2002; Cabinet Office, 2013; Sharpley, 2013; British Educational Research Association, 2014; Every Student Succeeds Act [ESSA] of 2015; US Department of Education, 2016). What counts as evidence in education and to what extent it is used are topics that have been widely studied. Weiss (1977) and other scholars since (King and Pechman, 1984; Nutley et al, 2007; Tseng, 2012; Penuel et al, 2017) demonstrated that educational research findings have been used conceptually to influence decision makers’ understanding of decision problems, symbolically or politically as tools of persuasion to justify decisions already made, and instrumentally to directly guide and shape decision making. Researchers have also documented ‘imposed’ use of research due to government funding tied to the adoption of evidence-based programmes, and ‘process’ use of research which refers to what practitioners learn from participating in evidence production (Weiss et al, 2005; Tseng, 2012).

Equally well documented are the low incidence of research use to inform education practice (Penuel et al, 2017; Gamoran, 2018; Walker et al, 2019) and the array of obstacles preventing the use of research evidence in education decision making (Whitty, 2006; Oliver et al, 2014; Farley-Ripple et al, 2018). A chief culprit is the difficulty of producing high-quality research evidence quickly or decisively enough to influence policy questions (Resnick, 2015). Additionally, the majority of educational strategies employed in schools have not yet been evaluated (OECD, 2015), much less the multitude of curricula and instructional materials that embody particular practices.
Even when evidence is available, it may not meet expected methodological standards (Malouf and Taymans, 2016) or point to a policy solution, resulting in solutions that often stray beyond the existing evidence into the realm of opinion (Hanushek, 2015). Despite initial enthusiasm for the increased rigour of educational research over the past 20 years, concerns abound that evidence prioritising internal validity, for example, from randomised controlled trials (RCTs), has too often failed to generalise across school contexts and student populations (Cartwright and Hardie, 2012; Joyce and Cartwright, 2020). The perceived lack of applicability of RCT findings to local situations leads to low credibility of such evidence among educators (Finnigan et al, 2013; Mills et al, 2020).

The limited use of externally-produced rigorous evidence has led to debate among the research community about how to enhance current research designs or produce more locally-relevant research. Defendants of RCTs argue for ways in which to improve their value and generalisability. For example, Kaplan et al (2020) recommend incorporating elements of qualitative research to help shift the goal of RCTs from ‘… generalising findings to the population to generalising findings to a robust theoretical understanding of the contextualised causal mechanisms underlying the phenomenon’. A more common call is to find a balance between rigour and ‘timely relevance’ (Feuer, 2015: 99).

To address concerns about the relevance of research to local decisions, many researchers emphasise the importance of collaboration between researchers and practitioners. Penuel et al (2015) advocate for ‘boundary practices’, activities characterised as hybrids between the practices of researchers and practitioners. Penuel and Farrell (2017) assert that participation in the research process is important for utilisation of evaluations, and suggest that conducting local evaluations of programmes may produce results that are more timely, credible and responsive to the needs of school district leaders. In turn, this increases the likelihood that the decision makers will act upon the findings. Irrespective of the supply of relevant research, Farrell and Coburn (2017) argue that districts’ capacity to use research evidence depends on organisational features which include prior knowledge and expertise, communication pathways, and strategic knowledge leadership.

In addition to research evidence, other forms of information may be as or more important to education practitioners. Coburn and Turner (2011a; 2011b), Daly et al (2014); Farrell (2014); and Marsh and Farrell (2015) focus on use of data as evidence to inform action in schools. Finnigan et al (2013) found that school personnel mostly used data such as standardised test scores and grades rather than research as evidence for decision making. Local values, interests and practical circumstances may be of equal or greater importance than objective evidence (Edwards et al, 1975; Weiss, 1995; Honig and Coburn, 2008; Asen et al, 2013; Tseng and Nutley, 2014; Hollands et al, 2019; Huguet et al, 2021). Boaz et al (2019) add issues of cost, implementation and stakeholder engagement to the list of considerations.

Spencer et al (2012) argue that evidence-based practice in education requires professional judgements in addition to research evidence and a consideration of client values and context. Feuer (2015) similarly stresses the importance of practitioner knowledge, suggesting that experiential evidence, that is, professional practice and experience, is ‘essential to the construction of useful theory’ (Feuer, 2015: 97). However, the extent to which professional educator judgements regarding the effectiveness of educational programmes align with rigorous evidence produced by
formal research studies does not appear to have been studied. Expert judgements can also provide evidence on effectiveness of instructional programmes, practices and personnel (Arreola, 2000; Berk, 2005, Berk and Theall, 2006) and contribute to education decision making (Clayton, 1997). Similar to practitioner judgements, it has not been established whether expert judgements align well with existing research evidence in education, or even if experts’ assessments agree with each other or with practitioners’ judgements. In the context of a school district’s efforts to use evidence to guide its response to state pressure to improve reading performance in its underperforming elementary schools, the present study investigates whether practitioner and expert judgements can serve as reliable sources of evidence on the effectiveness of reading resources, and how they compare with existing research evidence.

**Setting**

In 2017–18, a large, urban school district in the southern US (‘SUD’) was required by the state education agency to improve reading achievement in 23 elementary schools or risk forced closure of these schools. SUD served over 200,000 students, of which 86% were African-American or Hispanic. Almost half the students were English language learners (ELLs) and 79% were economically disadvantaged.

At the 23 struggling schools, 98% of the students were African-American or Hispanic, 32% were ELLs, 92% were economically disadvantaged, 75% were deemed at risk of dropping out of school, and only 24% met grade level in state standardised reading tests. Among the 20 school principals for whom we were able to obtain demographic information, 75% were female, 60% were Black, 30% were Hispanic, and 10% were White. They averaged 14.8 years of experience and 60% had earned a Master’s or higher-level degree. A Teacher Development Specialist (TDS) worked at each school (one school had two TDSs) observing teachers’ instructional practices, providing feedback, and coaching them on how to increase student achievement. Among the 15 TDSs for whom we were able to obtain demographic information, 93% were female, 60% were Black, 20% were Hispanic, 13% were White, and one was Other. They averaged 15.8 years of experience and 47% had earned a Master’s degree.

SUD is a decentralised district: schools may implement any resources from the district’s list of approved vendors. The district office provides support for certain programmes and may encourage school uptake of specific resources by purchasing district-wide services or offering to pay for licenses, materials or professional development.

**Research production and use in SUD**

In 2017–18, the SUD research department consisted of 22 researchers. All have Master’s degrees and 14 have doctorates. The department conducts approximately 20 programme evaluations each year to produce local evidence on district initiatives, for example, on bilingual education, early childhood education, migrant education, state compensatory education, college readiness and success, professional development models, and the use of specific programmes or resources, including instructional software, to improve achievement. Some evaluations are mandated by state and federal funding while others are initiated by curricular departments to investigate locally relevant issues. In other cases, community partners that contribute resources and
funds request evaluations to assess the benefits resulting from their investments. These evaluations are publicly available on SUD’s website and some are in the Educational Resources Information Center’s repository of educational research. In addition to preparing written reports, the researchers meet with those requesting the evaluations to present and discuss the results and how they can be applied to modify programme implementation or instruction. Results are also shared with the superintendent and, where relevant, with organisation boards and local universities. Monthly principals’ meetings and school board meetings, which are attended by community partners, are used as opportunities to share and discuss student data and evidence-based practices. SUD researchers conduct literature reviews on existing programmes and practices to provide external evidence to inform SUD’s own actions. In addition, SUD collaborates with external researchers to engage in a variety of research, development and evaluation activities such as the present study.

In order to provide timely guidance and support to the 23 elementary schools, SUD’s Office of Elementary Curriculum and Development immediately sought to identify which reading resources (reading programmes, assessments, online tools, book collections, professional development supports) were available at each school, and to assess their effectiveness at improving student reading proficiency. Our research-practice team explored a variety of options for procuring suitable evidence to inform SUD’s actions. Because SUD needed to act swiftly to avert school closures, it was not feasible to plan and execute multiple experiments or quasi-experiments. We considered quantitative analyses using SUD’s administrative student data, but two obstacles arose. First, each school used multiple reading resources with the same students, making it hard to isolate the contribution of any one resource. We examined the possibility of applying value-added regression models to investigate the relative contribution of each resource to student reading performance using an approach similar to Betts et al (2005) and Shand et al (2020). This plan met with the second obstacle which is that SUD’s schools do not routinely document which students engage with which resources.

**Current study**

Given the inability to conduct experimental or quasi-experimental studies to evaluate the schools’ reading resources, we explored three strategies to collect evidence on their effectiveness. We developed two surveys to elicit judgements on the effectiveness of the resources for improving reading proficiency, one from professional educators in the schools and one from external reading experts. We also searched for existing external research evidence on the effectiveness of each resource in reputable repositories of education research (Donaldson et al, 2009). We compared the experts’ judgements with each other to determine the extent to which they agreed in their assessments of each reading resource. Subsequently, we compared the expert judgements with the practitioner judgements to evaluate the consistency of assessments between the two groups. We also compared the expert and practitioner judgements with existing evidence from research repositories. We addressed three research questions:

1. To what extent do reading experts agree with each other when rating the effectiveness of a set of reading resources?
2. To what extent do practitioners’ ratings of reading resources implemented in their schools coincide with ratings provided by reading experts?
3. Do ratings from experts and practitioners align with evidence summaries from external research repositories?

Methods

Searching for existing studies in research repositories

For each reading resource, we searched two US-based repositories of research on educational programmes to identify evidence summaries and studies on its effectiveness for improving reading: the What Works Clearinghouse (WWC) and Evidence for ESSA (E4E). We also searched the evidence summaries provided by the Education Endowment Fund (EEF) in the UK.

Collecting practitioners’ professional judgements

We designed a Qualtrics survey in which the TDS at each school indicated which reading resources were being used at their school. Because TDSs routinely use student assessment data to guide their coaching strategies, they were considered to be a suitable source of judgements on the effectiveness of the resources. For each resource used, the TDS answered the following prompt: ‘Please indicate how effective you believe [resource] is for improving student reading proficiency’. Response choices were: ‘hinders reading proficiency’, ‘makes no difference’, ‘improves reading proficiency a little’, ‘improves reading proficiency a lot’, or ‘not sure’.

We created individual links to the survey for each of the 23 schools and our SUD liaison emailed the links to the relevant TDSs (23 surveys total) at the end of November 2017. Reminders were sent several times to those who did not respond or who only partially completed the survey. The TDSs were also invited to complete the survey during a routine meeting at the district office in January 2018. As a result, we collected responses from TDSs at all 23 schools, which we used to create a master list of resources. A total of 23 reading resources were being used at one or more of the 23 schools. We documented how many schools used each resource and how it was rated by each TDS. We did not calculate ICCs (intra-class correlations) for the practitioners because they were each exposed to a different instantiation of the programme so a difference in rating could reflect different approaches or different levels of implementation fidelity at each school, not simply differences in judgement of the same intended programme.

Collecting expert judgements

We recruited three reading experts and, using an expert survey (Schiefelbein et al., 1998), elicited their professional opinions regarding the effectiveness of the subset of reading resources which were being used as instructional programmes. The goal in recruiting reading experts was to find individuals with expertise in effective reading instruction, based in both research and practice. The reading experts were a convenience sample of professionals familiar with both research and practice in teaching reading, but not necessarily with knowledge of the same resources. The experts were selected on the basis of having a doctorate in literacy, multiple years of experience in K-12 literacy teaching, and multiple years of experience in university-based teacher education training teachers of reading. In our judgement, these
Comparing evidence on the effectiveness of reading resources from expert ratings

Criteria are similar to those that a school district might look for in university-based experts. The first expert is the second author who had collaborated previously with other members of the author team. The other two experts were in the first expert’s professional network and were invited to participate by the first expert. The first expert is an Associate Professor of Literacy Education who studies the language and literacy development of students from linguistically diverse backgrounds, especially in urban schools, and who was formerly a middle-school teacher. A second expert is a Clinical Professor of Literacy Education who studies the effective instruction of reading in the elementary years, especially in urban settings, and who taught for over 25 years in high-poverty, public elementary and middle-school classrooms. The third expert is a former Professor of Literacy and Special Education and charter member of the Reading League who taught Special Education in public schools for 33 years.

The subset of ten resources evaluated by the experts, listed in Table 1, were those used by SUD as core reading programmes, web-based supplemental literacy programmes, guided reading programmes, or intervention programmes. We did not ask the experts to rate the other 13 resources. Eleven of these were not instructional programmes: four were assessments (Renaissance 360, Developmental Reading Assessment 2, Fountas and Pinnell Assessment Kits, Evaluación del Desarrollo de la Lectura); five were book collections (Scholastic Classroom Libraries, Scholastic Read Aloud Libraries, Scholastic Extended Bookroom, MyON, Storia); and two were professional development supports (Neuhaus Education and Writers in the Schools). Finally, two resources would have qualified as instructional programmes but were Spanish-only (Tesoros and Benchmark Spanish Leveled Libraries) and, because the experts were not fluent Spanish speakers, they could not fully evaluate their quality.

For each of the ten resources, we conducted an extensive internet search to collect the information outlined in Table 2. In addition to searching the three aforementioned

<table>
<thead>
<tr>
<th>Programme (publisher)</th>
<th>Type of reading resource (as used in SUD)</th>
<th># of schools using this resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrective Reading (McGraw-Hill)</td>
<td>Guided reading/intervention programme: small group remedial</td>
<td>8</td>
</tr>
<tr>
<td>Frog Street (Frog Street Press)</td>
<td>Core reading programme: bilingual</td>
<td>14</td>
</tr>
<tr>
<td>Imagine Language and Literacy (Imagine Learning)</td>
<td>Web-based supplemental literacy programme</td>
<td>22</td>
</tr>
<tr>
<td>Istation Reading (Istation)</td>
<td>Web-based supplemental literacy programme</td>
<td>3</td>
</tr>
<tr>
<td>Leveled Bookroom (Scholastic, Inc.)</td>
<td>Guided reading/intervention programme: small group non-remedial instruction</td>
<td>23</td>
</tr>
<tr>
<td>LitCamp (Scholastic, Inc.)</td>
<td>Guided reading/intervention programme: small group non-remedial instruction</td>
<td>6</td>
</tr>
<tr>
<td>Leveled Literacy Intervention (LLI) (Heinemann)</td>
<td>Guided reading/intervention programme: small group remedial instruction</td>
<td>1</td>
</tr>
<tr>
<td>Reading A-Z (Lazel)</td>
<td>Guided reading/intervention programme: small group non-remedial reading</td>
<td>2</td>
</tr>
<tr>
<td>Reading Mastery (McGraw-Hill)</td>
<td>Guided reading/intervention programme: small group remedial</td>
<td>8</td>
</tr>
<tr>
<td>Reading Street (Pearson Education, Inc.)</td>
<td>Core reading programme</td>
<td>20</td>
</tr>
</tbody>
</table>
research repositories, we entered the name of the resource in Google’s search engine, in Google Scholar, and in the vendor website search functions. We compiled the information in a database and made it available to the experts via a Google Sheet and a shared Google Drive folder of research studies and other materials. We also provided contextual details about SUD including student demographics and performance, and how each resource was used (for example, as a guided reading programme or intervention).

Questions for the experts, listed in Table 3, were embedded in a Qualtrics survey that we designed. The questions elicited ten ratings on each of ten resources. Experts completed their review of the compiled materials and their survey responses independently within a few weeks of each other. We compared the ratings across the three experts. We also compared the experts’ responses to a similarly-worded question about expected effectiveness of each resource for students in SUD with the practitioners’ effectiveness ratings for the same resource. We expected these ratings would be correlated.

Assessing inter-rater reliability among the experts

Response choices to each of the questions in the expert survey are ordinal, that is, they yield categorical data where the variables have ordered categories but the distances between the categories is unknown (Agresti, 2013). To assess the inter-rater reliability (IRR) among the experts, we calculated the ICC between the ten programme

<table>
<thead>
<tr>
<th>Information provided</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name of resource and publisher</td>
</tr>
<tr>
<td>URLS for public websites with general information about each resource</td>
</tr>
<tr>
<td>A classification of type of resource: core reading programmes; guided reading/intervention programmes: small group non-remedial; guided reading/intervention programmes: small group remedial; web-based supplemental literacy programmes</td>
</tr>
<tr>
<td>Description of resource</td>
</tr>
<tr>
<td>Description of activities (what the resource requires students/teachers to do in the classroom)</td>
</tr>
<tr>
<td>Length of time resource has been in existence</td>
</tr>
<tr>
<td>Dosage (how often students use the resource; for how much time on each occasion; over how many weeks, months, or years)</td>
</tr>
<tr>
<td>Scope (aim of the programme and what skills students are taught)</td>
</tr>
<tr>
<td>Sequence (in what order these skills are taught)</td>
</tr>
<tr>
<td>Target audience: general Education/ELLs</td>
</tr>
<tr>
<td>Language (English/Spanish/bilingual)</td>
</tr>
<tr>
<td>Grades/levels served</td>
</tr>
<tr>
<td>Information and URLs to materials that address alignment of the resource with Common Core Standards</td>
</tr>
<tr>
<td>Information on whether the materials align with other standards</td>
</tr>
<tr>
<td>All reports/studies/evaluations/reviews found by searching the WWC, E4E, EEF, Google Scholar, Google, and the vendor’s website using the name of each resource.</td>
</tr>
</tbody>
</table>
Comparing evidence on the effectiveness of reading resources from expert ratings

Comparing evidence on the effectiveness of reading resources from expert ratings

ratings provided by each expert for each item. Following Hallgren (2012), the ICC function is specified by three parameters. First, the ICC function depends on how coders are selected for the study: in a one-way model, a different set of coders is randomly selected from a larger population of coders for each subject (a subject is equivalent to a reading resource in our study). In a two-way model, all subjects are rated by all raters. The second determinant of the ICC function is whether agreement needs to be absolute or relative, that is, whether it is important for raters to provide scores that are similar in absolute value (agreement) or in rank order (consistency). The last determinant is whether ratings are provided by a single coder or averaged across multiple coders. We used a two-way model whereby all resources were rated by all expert raters and we estimated the consistency of ratings across coders. Higher ICC values signify greater IRR (Hallgren): an ICC estimate of 1 implies perfect agreement among raters while 0 suggests random agreement. Negative ICC values reflect systematic disagreement. ICCs of less than -1 may arise when there are three or more coders. Cicchetti (1994) provides benchmarks for ICC scores indicating that IRR is poor for ICC values below 0.4, fair for values of 0.4–0.59, good for values of 0.60–0.74, and excellent for values of 0.75–1.0.

Table 3: Intra-class correlation values for expert survey questions

<table>
<thead>
<tr>
<th>Question answered by experts for each reading resource</th>
<th>ICC value</th>
<th>Interpretation of ICC</th>
<th>P-value (H_0 = 0, H_1 &gt; 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>How clearly specified is the theory of action?</td>
<td>0.400</td>
<td>Fair</td>
<td>0.170</td>
</tr>
<tr>
<td>Given the targets of the programme (for example, student population, grade level), how well aligned is the content with existing research on effective reading instruction?</td>
<td>0.496</td>
<td>Fair</td>
<td>0.103</td>
</tr>
<tr>
<td>How strong is the evidence that [resource] is effective at improving reading skills?</td>
<td>0.740</td>
<td>Good</td>
<td>0.010</td>
</tr>
<tr>
<td>How much implementation support appears to be available from the vendor for teachers using [resource] under typical conditions?</td>
<td>-1.066</td>
<td>Poor</td>
<td>0.867</td>
</tr>
<tr>
<td>How effective do you think [resource] will be at improving reading proficiency for the target students in typical US elementary schools (that is, average demographics, average teacher qualifications and average school funding) if implemented as the publisher intends?</td>
<td>0.420</td>
<td>Fair</td>
<td>0.150</td>
</tr>
<tr>
<td>How much impact would you expect [resource] to have on each reading component?*</td>
<td>0.680</td>
<td>Good</td>
<td>0.020</td>
</tr>
<tr>
<td>Alphabets</td>
<td>0.400</td>
<td>Fair</td>
<td>0.170</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>0.740</td>
<td>Good</td>
<td>0.010</td>
</tr>
<tr>
<td>Fluency</td>
<td>0.577</td>
<td>Fair</td>
<td>0.058</td>
</tr>
<tr>
<td>Comprehension</td>
<td>-0.070</td>
<td>Poor</td>
<td>0.520</td>
</tr>
<tr>
<td>Based on the school profiles and student demographics provided, how effective do you think [resource] would be at improving reading proficiency for target students in under-served and underperforming SUD schools if implemented as the publisher intends?</td>
<td>-0.070</td>
<td>Poor</td>
<td>0.520</td>
</tr>
</tbody>
</table>

* Response choices were: Large negative, Medium negative, Small negative, No effect, Small positive, Medium positive, Large positive.
Correlation between experts’ ratings and practitioners’ ratings

To allow for a calculation of the correlation between experts’ ratings on the effectiveness of the resources for improving reading proficiency among SUD students and practitioners’ ratings, we changed the categorical rating of effectiveness to a continuous variable by assigning values to each of the categories as follows: hinders reading proficiency = 1; makes no difference = 2; not sure = 3; improves reading proficiency a little = 4; improves reading proficiency a lot = 5. This assumes that the distance between two adjacent categories is constant and that ‘not sure’ is neither a negative assessment nor a positive one. We used the values to calculate an average effectiveness rating for each resource across practitioners and separately across experts.

Results

Evidence from research repositories

Ten of the 23 reading resources being used in the SUD elementary schools were listed in WWC or in E4E, but qualifying studies were available for only half of these. WWC only includes studies that would qualify for the top two tiers of ESSA evidence (experimental or quasi-experimental studies) while E4E includes items that qualify for Tier 3 (correlational studies). Four resources were supported by evidence of effectiveness at improving one or more aspects of reading. For one item, the evidence was neither significantly positive or negative. All five resources appearing in at least one of the repositories with associated evidence were among the ten items evaluated by both practitioners and reading experts (see Table 4). No information on these US-developed literacy programmes was found in EEF’s evidence summaries.

Practitioner ratings

Table 4 summarises the number of schools using each of the ten resources for which we compared expert and practitioner ratings, and the percentage of responding TDSs who rated each resource in a particular category of effectiveness. No practitioner indicated that a resource hindered reading proficiency, but one indicated that one resource made no difference. For all ten resources, more than 50% of the TDSs believe that the resource either improves reading proficiency a little or a lot. However, for six of the resources, at least a quarter of respondents were unsure as to whether it was making any difference.

Expert ratings

The inter-rater reliability among the three experts, based on ICC values shown in Table 3, ranged from good to poor on questions regarding effectiveness of the ten resources at improving reading proficiency. The ICC value was higher for the experts’ judgements about the effectiveness of the resources for the average US student than for students at SUD. The ICC values were also higher when experts were asked about specific components of reading (alphabetics, fluency, vocabulary, comprehension) than when asked about effectiveness at improving reading proficiency overall.
### Table 4: Comparison of practitioners and experts' ratings on the effectiveness of reading resources

<table>
<thead>
<tr>
<th>Programme</th>
<th>Evidence from research repository</th>
<th>Number of raters</th>
<th>Hinders reading proficiency</th>
<th>Makes no difference</th>
<th>Not sure</th>
<th>Improves reading proficiency a little</th>
<th>Improves reading proficiency a lot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrective Reading</td>
<td>WWC: potentially positive effects on alphabets and fluency; no discernible effects on comprehension E4E: strong +0.06</td>
<td>8</td>
<td>3</td>
<td>0%</td>
<td>0%</td>
<td>38%</td>
<td>0%</td>
</tr>
<tr>
<td>Frog Street</td>
<td>WWC: no information E4E: no studies met inclusion requirements</td>
<td>14</td>
<td>3</td>
<td>0%</td>
<td>0%</td>
<td>33%</td>
<td>36%</td>
</tr>
<tr>
<td>Imagine Language and Literacy</td>
<td>WWC: no information E4E: promising +0.18</td>
<td>22</td>
<td>3</td>
<td>0%</td>
<td>0%</td>
<td>33%</td>
<td>45%</td>
</tr>
<tr>
<td>Istation Reading</td>
<td>WWC: no information E4E: no studies met inclusion requirements</td>
<td>3</td>
<td>3</td>
<td>0%</td>
<td>0%</td>
<td>33%</td>
<td>0%</td>
</tr>
<tr>
<td>Leveled Bookroom</td>
<td>WWC: no information E4E: no information</td>
<td>23</td>
<td>3</td>
<td>0%</td>
<td>0%</td>
<td>4%</td>
<td>67%</td>
</tr>
<tr>
<td>LitCamp</td>
<td>WWC: no information E4E: no studies met inclusion requirements</td>
<td>6</td>
<td>3</td>
<td>0%</td>
<td>0%</td>
<td>33%</td>
<td>33%</td>
</tr>
<tr>
<td>LLI</td>
<td>WWC: positive effects on general reading achievement, potentially positive effects on reading fluency; no discernible effects on alphabets E4E: strong +0.13</td>
<td>1</td>
<td>3</td>
<td>0%</td>
<td>0%</td>
<td>33%</td>
<td>0%</td>
</tr>
<tr>
<td>Reading A-Z</td>
<td>WWC: no information E4E: no studies met inclusion requirements</td>
<td>2</td>
<td>3</td>
<td>0%</td>
<td>0%</td>
<td>33%</td>
<td>33%</td>
</tr>
<tr>
<td>Reading Mastery</td>
<td>WWC: potentially positive effects on reading achievement for ELLs E4E: no studies met inclusion requirements</td>
<td>8</td>
<td>3</td>
<td>0%</td>
<td>0%</td>
<td>25%</td>
<td>0%</td>
</tr>
<tr>
<td>Reading Street</td>
<td>WWC: no information E4E: qualifying studies found no significant positive results</td>
<td>20</td>
<td>3</td>
<td>0%</td>
<td>33%</td>
<td>25%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Note: P = practitioner (TDS); E = reading expert
Comparison between ratings from experts and practitioners

Table 4 compares the percentage of practitioners and experts who rated each of the ten resources in each of the five response categories of perceived effectiveness at improving reading proficiency. While no practitioners indicated that a resource hinders proficiency, one expert did so for three resources. Only one resource was believed to make no difference to reading proficiency by the practitioners, but three resources were placed in this category by an expert. One or more practitioners were unsure about the effectiveness of seven resources. One or more experts were also unsure for seven resources, although only four resources overlapped between the two groups of raters. In general, the practitioners were more optimistic about the extent to which the resources improved reading proficiency. Specifically, for all ten resources, at least one practitioner indicated that the resource improved reading proficiency a lot. This was the case for only four resources based on the experts’ ratings.

Correlation between experts’ ratings and practitioners’ ratings

We found the correlation between expert and practitioner ratings to be low at 0.186. Because the experts expressed low confidence in their ratings of the resources for SUD students, we additionally calculated the correlation between the experts’ ratings on the effectiveness of these ten reading resources for the average US student and the practitioners’ ratings (which, by default, applied to SUD students). This correlation was slightly negative at -0.034. We also investigated the correlation between the scores of the experts and practitioners after dropping all ‘unsure’ ratings from both sets of raters. This produced a much higher correlation of 0.63 but, obviously, for fewer resources.

Comparison of expert and practitioner ratings with research repository evidence

Evidence summaries were only available for five of the ten resources rated by both experts and practitioners. For two of these, Corrective Reading and Reading Mastery, the majority of the practitioners and experts provided ratings that were positive and consistent with the evidence summaries. For two others, Imagine Language and Literacy and LLI, the majority of the practitioners provided ratings that were positive and consistent with the evidence summaries, but one expert rated the resource as hindering reading proficiency and one indicated ‘not sure’. For the fifth resource, Reading Street, the evidence summary indicated no significant positive or negative results, but the majority of both practitioners and experts rated it as improving reading proficiency.

Discussion

Our study demonstrates the difficulty of meeting demands to use evidence in educational decisions, especially when time is of the essence. Of 23 reading resources used in the elementary schools of a large, US public school district, the two best-recognised US repositories of education research evidence provided information on only five. This finding is consistent with Chang et al’s (2021) finding that there is either no or limited evidence on the effectiveness of many items in which schools invest Title I federal funds. Furthermore, the evidence summaries provided by the repositories are based on only one or very few studies, and often conducted with populations dissimilar to SUD’s. This may help explain why practitioners are failing
to select programmes and practices for which positive evidence is available in these repositories. One possible reason for the lack of studies on widely-used reading programmes is that research funders such as the US’s Institute of Education Sciences are more likely to fund studies that build theory on effective instruction using researcher-developed curricula that are not widely used and may never be published, as opposed to studies on curricula already in widespread use but for which the theory of action is less clear.

Neither is internally-produced evidence easy to come by. Accurate data on student participation in classroom-based programmes is not routinely collected by districts and schools, thereby limiting quantitative evaluations using existing data. Without the luxury of time or the data needed to conduct quantitative evaluations of SUD’s reading resources, we sought more easily available evidence on their effectiveness — practitioner and expert judgements — with the expectation that they would corroborate each other. Practitioners provided ratings on 23 reading resources, basing their judgements on their direct experiences observing the resources being used with students in their schools and their access to student assessment data. For the few resources for which external evidence was available, none of the practitioners’ judgements were contradicted by published studies. Although the reading experts were each given access to the same set of information about each of the ten reading resources, they showed varying levels of agreement in their ratings. This may be partly attributable to different starting levels of familiarity with the resources, in particular whether the expert had ever observed them being implemented in schools. The experts demonstrated fair agreement about the overall effects of each resource on reading proficiency for average US students, and fair to good agreement on individual components of reading. For two resources, the expert ratings did not comport with positive external evidence available in the research repositories. This may have been due to the fact that both repositories based their evidence summaries on only two studies, and these were conducted in contexts unlike SUD’s with student populations that were substantially less challenged.

The reading experts showed poor agreement on the overall effects of each resource on reading proficiency for students at SUD. The inconsistency in ratings for the SUD context may be due to experts focusing on different factors for the SUD population based on their experience and expertise. For example, one expert may have focused more on the percentage of ELLs, while another may have focused more on the percentage of minorities or economically-disadvantaged students. The low overall correlation between practitioner and expert ratings appears to be mostly attributable to practitioners rating the effectiveness of reading resources more favourably than experts. This optimism produced relatively limited variation in the practitioners’ ratings which, in turn, may have attenuated the correlation between practitioner and expert ratings. In sum, it appears that practitioners may be overly optimistic about the efficacy of educational materials, while experts have difficulty assessing how well the materials will serve a population with whom they are unfamiliar.

Limitations

Some of the reading resources evaluated were used in few schools so the number of practitioner ratings for these resources was correspondingly small. Ratings from more practitioners would provide a better picture of how well each resource was perceived by a range of practitioners. While these respondents were asked to comment on the resources as implemented in their schools, a fair number
indicated that they were not sure whether the resource had any impact on reading proficiency. It is unclear whether ‘not sure’ was a more diplomatic alternative to ‘makes no difference’ or whether the respondents genuinely felt they had inadequate knowledge of a resource’s impact. In addition, a larger number of experts may have yielded more consistent expert judgements: Schiefelbein et al (1998) recommend as many as 20 or 25 participants for expert surveys. We lacked the resources to recruit more than three.

Applying the results to decision making

Mills et al (2020) recommend that, to increase the use of research by practitioners, researchers should address how to implement their research findings. Accordingly, we demonstrate here how education decision makers can use the results of this study to guide schools towards or away from specific reading resources. We expect a similar approach could be applied to other types of resources, given parallel data. First, for each resource, the decision makers can review the consistency across the three sources of evidence in terms of reported effectiveness at improving reading outcomes. We also recommend they consider how many of their schools are using the resource to gauge how many might need support in implementing the resource more effectively, eliminating it altogether, or switching to a different resource. A further consideration in SUD’s situation would be the feasibility of implementing promising reading resources in additional schools quickly enough to help students within the desired timeframe.

We model this approach with four examples. Corrective Reading was used in one third of the 23 schools and the majority of TDSs indicated that it improves reading proficiency a lot. All three experts indicated that the programme improves reading proficiency. The external evidence in both repositories is positive, albeit based on only one study. In sum, the evidence is reasonably consistent and positive across sources, suggesting that this resource is worth considering for adoption by other schools. However, a substantial amount of training is needed for teachers to implement Corrective Reading. This might be feasible longer-term, but perhaps not in time to help the schools immediately. Reading Mastery was also used at eight schools. The TDSs and experts mostly rated the programme positively. WWC holds one study of Reading Mastery demonstrating potentially positive effects for ELL reading achievement. Given fewer training needs and materials than Corrective Reading, it might be more feasible to adopt in additional schools. Reading Street, used in 20 schools, earned quite disparate practitioner ratings, suggesting uneven implementation that the district office could better support to improve fidelity. Istation was rated positively by all three schools using it and by two of the experts. Being a computer-based programme, it could be implemented swiftly in additional schools with digital devices and connectivity.

This type of approach in which multiple factors are considered simultaneously could be implemented more systematically using a decision-making framework based on cost-utility analysis (Hollands et al, 2019). Resources needed to implement each programme would be compared with its utility. Utility would be determined by including the various assessments of effectiveness in addition to other considerations such as feasibility of implementation or scaling, with each consideration weighted by importance to the SUD stakeholders. One advantage of this approach is that it
Comparing evidence on the effectiveness of reading resources from expert ratings

allows decision makers to incorporate contradictory sources of evidence but assign different importance weights to each one.

**Future directions**

One interpretation of the low correlation found between expert and practitioner ratings is that they are providing valuable, complementary information, based on distinct sources about the materials’ efficacy. Although we cannot shed light on this hypothesis with the current data, future studies could consider an in-depth qualitative approach to better understand the bases for both groups’ evaluation process.

A possible future avenue for producing more rigorous local evidence on a timely basis would involve routinely collecting accurate data in schools on which students engage with which reading resources, in order to allow for regression analyses using value-added models. Betts et al (2005) and Shand et al (2020) demonstrate this approach for programmes or practices aiming to improve reading achievement. The information could be collected using districts’ student information systems.

Going forward, federal and state agencies could provide better information and support to help schools and districts use evidence-based practices. Boaz et al (2019) list a number of approaches for supporting evidence use in social programmes, all of which can be applied in education. Most of these involve capacity building for practitioners and policymakers to access and use evidence, often in collaboration with researchers. We additionally recommend that the US Department of Education or state education agencies survey schools to identify widely-used programmes and practices. The proposed School Pulse survey (Hart and Tromble, 2021; Schneider, 2020; 2021) could serve as a vehicle to gather the necessary information. Subsequently, they should commission studies on the effectiveness of the most common, being sure to sample a variety of student populations and contexts. We acknowledge that the value of such evidence may be limited to situations in which schools and districts are considering the adoption of new approaches. They are less likely to replace existing programmes regardless of effectiveness, due to resistance from entrenched interests (Moe, 2015) and the high costs of switching. For programmes found to be ineffective, it may be more pragmatic to identify modifications that current users can adopt to improve outcomes for students.

**Conclusion**

We conclude that there remains a substantial disconnect between the existing evidence base in education and the evidence needs of education decision makers (Kane, 2016; Malouf and Taymans, 2016). This applies to both external research and the data availability within education agencies that would allow for the production of local evidence. Even in a subject area as critical as reading, there is no rigorous evidence available to inform decision makers for most of the instructional resources being implemented. Until this can be rectified, a combination of professional and expert judgements regarding the effectiveness of the instructional resources, along with information about the feasibility of implementation and support needs, can provide some guidance to decision makers on where to concentrate their efforts to help students.
Research ethics statement
Institutional Review Board approval for this study was covered under protocol #17–372 from Teachers College, Columbia University, dated July 3, 2017. In addition, approval was granted by the Department of Research and Accountability at SUD on June 26, 2017.

Funding
This work was supported by the William T. Grant Foundation under Grant number 187095 awarded to Teachers College, Columbia University.

Contributor statement
FMH, YP, MJK, VRH, and MME conceptualised the study and, with AM, designed the data collection instruments. VRH collected data from the school district. MJK served as one of the reading experts and coordinated participation by the other two experts. LH investigated existing evidence and collected and prepared the materials for the experts. YP, YW, and AM analysed the data. FMH and YP drafted the initial manuscript. FMH, MJK and VRH made revisions. All Authors commented on the manuscript.

Conflict of interest
The authors declare that there is no conflict of interest.

References
Comparing evidence on the effectiveness of reading resources from expert ratings


Comparing evidence on the effectiveness of reading resources from expert ratings


