READING RANDOMISED CONTROLLED TRIALS

OPENING THE BOOK

ROBERT SAVAGE, AMY FOX, ANNEKA DAWSON,
HELEN GRAY & CLARE HUXLEY

UCL PRESS

Reading Randomised Controlled Trials

Reading Randomised Controlled Trials

Opening the book

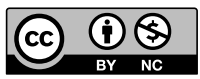
Robert Savage, Amy Fox, Anneka Dawson,
Helen Gray and Clare Huxley

UCL PRESS

First published in 2025 by
UCL Press
University College London
Gower Street
London WC1E 6BT

Available to download free: www.uclpress.co.uk

A CIP catalogue record for this book is available from The British Library.



ISBN: 978-1-80008-280-9 (Hbk.)
ISBN: 978-1-80008-279-3 (Pbk.)
ISBN: 978-1-80008-278-6 (PDF)
ISBN: 978-1-80008-281-6 (epub)
DOI: https://doi.org/10.14324/111.9781800082786

Contents

List of figures

List of boxes

Glossary and abbreviations

DfE	Department for Education: the government department responsible for education in England between 1992 and 1995 and from 2010 to date of publication
EAL	English as an Additional Language
EEF	Educational Endowment Foundation: a body responsible for evidenced school improvement in England and the funder of the Flexible Phonics trial
EV	extraneous variable: any variables that might interfere with establishing a causal link between two named variables
GDPR	the European Union's General Data Protection Regulation
GPC	grapheme-phoneme correspondence: a grapheme is a letter or letter cluster that represents an individual phoneme in print. Phonemes are the smallest units of spoken language. For example, the letter *g* in the word *go* and the letters *sh* in the word *shout* are both graphemes that correspond to the first phoneme sound in each word
ICC	intra-class correlation: here, the degree to which scores of individual children in the same schools are more similar than those in other schools – a school or neighbourhood effect
IDEA	intervention delivery evaluation analysis
IES	Institute for Employment Studies: the home institution of the trial evaluation team during the Flexible Phonics trial
MDES	minimum detectable effect size: the smallest effect of intervention that is statistically significant given the study sample size

NELI	The Nuffield Early Language Intervention
Ofsted	Office for Standards in Education, Children's Services and Skills
PHAST	Phonic and Strategy Training
RCT	randomised controlled trial: a design whereby units are assigned at random to a treatment condition, where the 'treated' receive an intervention, and a control condition, where the intervention is not received
SEND	special educational needs and disabilities
SMART	sequential, multiple assignment, randomised trials
TIDieR	Template for Intervention Description and Replication
UCL	University College London: the home institution of the trial delivery team for the Flexible Phonics trial
YARC	York Assessment of Reading for Comprehension

Preface

This book describes a single large-scale randomised controlled trial (RCT) of a novel early reading intervention called Flexible Phonics. The Flexible Phonics trial consisted of over 120 schools, with nearly 3,000 children recruited. This book is thus in part a case study of the Flexible Phonics trial and its wider implications for research and practice. Our focus on early children's literacy is a major educational concern at any time, but even more so in the context of school closures in England in 2020 and 2021 due to the Covid-19 pandemic. More broadly, however, the book illuminates how evidence-based policy in education is currently realised through the funded trial methodologies of the Education Endowment Foundation (EEF) in England. Importantly, the book speaks to the complexities of carrying out such large RCTs in education more generally, but also as an important response to the unprecedented Covid-19 pandemic.

A central thesis of this book is that educational RCTs can be viewed as a particularly complex form of cultural activity shaped around an embodied and rigorous research design. This varied human activity can usefully be foregrounded to emphasise what such scaled educational trials have in common with other trials (such as classic pharmaceutical trials) and how they depart from them. Our work emphasises variation as an expected principle of intervention delivery and of student response. Such foregrounding, we feel, provides useful insights into success in larger community-based educational trials, success that has sometimes proved elusive. Our approach here also surfaces the (very) human operations, decision-making and actions that necessarily drive and influence such trials. The book thus thoroughly explores the theory and methodology of RCTs, our data and our theory-driven logic model, but also the politics, policy and decision-making at all levels, including the UK government, that situates our work.

This approach, we hope, will bring RCTs to life for people who either do not know or understand them or even currently oppose them. We will have succeeded if this book makes RCTs easier to understand for

more people. To this end, a helpful metaphor for reading this text is that of behind-the-scenes documentaries, which commonly illuminate the production aspects of famous films. No other text on RCTs does this, as far as we are aware. Aiding wider public understanding of RCTs is one of the most important potential outcomes of this book. Drawing on cutting-edge basic science, as it does, we also hope the book engages the reading research community and university students worldwide.

Professor Robert Savage

Foreword

RCTs are undoubtedly seen as the gold standard when it comes to the science of treatment evaluation. This is most evident in the medical domain – nowadays we wouldn't consider trying a new treatment unless it had been through the rigours of scientific testing, with RCTs being at the heart of these regimes. One may wonder, then, why such a rigorous evaluation approach is not so readily applied to educational contexts – the use of RCTs in education is relatively new.

RCTs are notoriously expensive (in terms of time and money) to conduct. In England, the EEF was set up in 2010 specifically to address inequality in education by rigorously evaluating 'what works' in educational contexts and it has chosen RCTs as its principal methodology. These trials take considerable effort and expertise to bring to fruition. It is therefore imperative that the treatment is underpinned by a strong evidence base of data and theory and is implemented with fidelity. Unlike in medicine, where the treatment is often straightforwardly delivered as a medicine, in education – although the treatment may have been carefully designed by scientists – delivery depends on a plethora of factors. These include the intervention's acceptability to the educational context and its integration into current practice, and also the plethora of variation that comes with working in the education system. This, perhaps, is where educational researchers are starting to move towards an understanding that, as researchers, we would do best to work with educational practitioners on how to implement treatments in schools, rather than trying to impose a treatment on them.

Conducting gold-standard research in education begins with a strong evidence base and a research question – in this case, is this new approach to phonics (Flexible Phonics) better than current practice? However, this is just where the journey begins. Beyond lies a myriad of obstacles to be negotiated, including securing the funding, persuading schools that participating in the trial is worth their while, and delivering the new approach in schools, all within the confines of best scientific

practice. Bringing it to fruition requires a complex interplay of all stakeholders, each with a pivotal role. At the heart of this process is the education community, on whom we rely for participation. This is no mean feat. For RCTs to become accepted as part of the research and development process in education, we need a more widespread understanding of how RCTs work in education, because although there are similarities with medical trials, there are differences too.

Reading Randomised Controlled Trials: Opening the book is unique in that it takes the reader on an intimate journey of how to implement RCTs in educational practice. It opens up access to each and every stage of the procedure, including the decision-making processes engaged in by an exceptional team of scientists with over 95 years of experience in literacy and educational research between them. As a former president of the most prestigious society for reading research in the world – and having conducted RCTs across multiple continents – Professor Robert Savage leads this equally impressive team in 'opening the book'. Access to a detailed walk-through of a real-world RCT in education by a team of this calibre is unprecedented in the field.

Opening up the process of RCT science in this way has many highlights for me. First, it acts as an excellent primer for anyone involved in education research who wants to learn best practice for conducting their own research, as no stone is left unturned in the process of describing the Flexible Phonics RCT. The book also acts as an admirable demonstration of how to engage the community that you wish to recruit for participation in a meaningful dialogue to optimise implementation without compromising integrity. Recognising and integrating the considerable expertise that teaching professionals have to contribute is impressive indeed. The responsiveness of the Flexible Phonics team in this respect is highly commendable, given the trial operated during the Covid-19 pandemic.

Overall, this book opens the science of RCTs in education to the public domain, with a clear and easy-to-follow narrative that answers the 'why' as well as the 'how to' questions of best practice. Broad in its appeal, *Reading Randomised Controlled Trials* will be of value to anyone with an interest in discovering how the science of RCTs can be harnessed in educational research, be they policymakers, scientists, teaching professionals, or students with an interest in education research.

Dr Janet Vousden
Nottingham Trent University

Acknowledgements

Thanks are due to many. Firstly, to the external evaluators and subsequently co-authors – Anneka Dawson, Helen Gray and Clare Huxley, all at the Institute for Employment Studies – for expert, willing collaboration throughout. Thanks to co-author Amy Fox, who is an exceptional research manager and co-author. All have provided unique insight, writing and reviewing for many hours beyond any contracted time for the work. Thank you for so readily accepting the idea of a book on this shared project.

Thanks go to the Education Endowment Foundation for funding the work we describe and to Joe Collin, Diotima Rapp and Sarah Tillotson there for supporting it so strongly. Thanks go to our extremely able research assistants who operated as school liaison partners for intervention schools: Alice Robinson, Clare Whalley, Denise Amankwah, Greta Boldrini, Ameena Khan Sullivan, Sophia Gowers and Sam Dexter. Thanks also to Diotima Rapp, Celeste Cheung, Maria Pomoni, Andrea Mills, and Sarah Tillotson at the EEF for their guidance and continued support. This project has benefitted from the expertise of Dafni Papoutsaki for contributing to the original design and early impact work (including the randomisation); we are also grateful to Georgie Akehurst at IES for carrying out observations and interviews for the project, Jade Talbot at IES for arranging interviews, Seemanti Ghosh and Ben Brindle for additional data analysis, and Mandi Ramshaw and Rebecca Duffy at IES for formatting the report.

Turning to the book and its production, many thanks are due to Pat Gordon-Smith, commissioning editor at UCL Press, for supporting this project and seeing its potential and for expert advice and support throughout the production process. We thank the anonymous reviewers who commented on and then supported the book, and those who offered comments to improve the final draft. Your generous sharing of expertise is greatly appreciated.

Finally, warm thanks go to all head teachers, classroom teachers and teaching assistants in our research schools, and especially to the very many who attempted to modify practice with the goal of improving learning for young children, despite all the challenges that the Covid-19 pandemic brought to their professional lives. Without such commitment, none of this work, or indeed any school-based educational RCT, would be impossible. We have not banged pots and pans in public support for teachers, teaching assistants and all educators specifically. We should.

1
Introduction to randomised controlled trials

This book is about how to teach children to read. It is about how to teach them the foundations of language and literacy so that more children can become fluent readers. Most questions of practical interest occur in complex situations where a bewildering number of events co-occur. Schools are no exception to this complexity. Questions about how to aid literacy can be answered through the use of quality methodologies sophisticated enough to manage this complexity. Any attempt to improve literacy thus involves a consideration of methodology.

This chapter introduces a particular methodology used in educational research to help improve reading attainment: randomised controlled trials (RCTs). RCTs seek to minimise the impact of complexity when evaluating interventions. They do so through random allocation either to a 'treatment', as it is referred to in such trials in the natural sciences, or to a comparison 'control' condition, which does not involve delivering the treatment of interest. In our educational work, we prefer the term 'intervention' over 'treatment' to indicate a principled change to teaching activity aimed at interrupting risk of low attainment. Any differences after the intervention – compared to a control condition in an outcome of interest, such as attainment – likely reflect the impact of the intervention. We argue that appropriately designed and executed RCTs are a vital element in establishing the most effective practice, resulting in improvements in attainment through evidenced teaching and learning. However, we also argue that to do more to improve literacy, we may need to consider subtle modifications to basic RCT methodology.

We first consider some overarching questions: what are RCTs? Can they help us improve education? Why are they used in education? What are their strengths and limitations? In answering these questions, we also provide an overview of the design features of RCTs and the way these are

applied in practice by funders of educational research. We focus on the approach of the Education Endowment Foundation (EEF), an organisation in England that funds and shares educational research involving RCTs. This sets the stage for our discussion, in Chapter 2, of the Flexible Phonics intervention in the UK. In the chapters that follow, we consider questions about the role of theory, evidence, delivery and wider implementation and uptake of interventions by schools in multi-agency working.

What are RCTs? Can they assist in education?

Precursors of the thinking behind RCTs can be found throughout human history, arguably because the thinking behind RCTs is in some senses comparable to our everyday reasoning about causes. Pearl and Mackenzie (2018, p. 135) attribute the first documented case of an RCT to the biblical story of Daniel, who attempted to establish the effect of diet on a person's suitability to be a court advisor to Nebuchadnezzar in 597 BCE. Most scholars agree that the modern RCT has its origins in the pioneering work of the English scientist, statistician and polymath Sir Ronald Aylmer Fisher. (See Box 1.1 for more context on Fisher.) Fisher's research in agriculture during the 1920s and 1930s yielded methodological insights that underpin modern social and natural science. His primary concern was to find a mathematical solution to the challenge of quantifying uncertainty when we evaluate any putative cause. His solution was to assign units, at random, to either a treatment condition, where the 'treated' receive an intervention, or a control condition, where the intervention is not received. His insight was that randomly assigning units to treatment and control conditions means we can expect that, at the beginning of the experiment, the treatment and control groups are on average similar as regards background characteristics. This follows because treatment assignment is based on a random process which is, on average, unrelated to any observed or unobserved background characteristics.[1] These expectations mean that an RCT design can provide an unbiased estimate of the error associated with assessing any causal effect statistically. Statistics can be understood as the mathematics of informing rational decision-making in situations of unavoidable uncertainty. Most real-world situations, including RCTs, involve uncertainty because chance events sometimes occur. RCTs do not therefore provide 'truth' or certainty (Deaton & Cartwright, 2018).

Box 1.1 Ronald Aylmer Fisher, a deeply controversial figure

In a book such as this, which describes the wider context of RCTs, it cannot go unmentioned that the intellectual context of RCTs is troublingly complex. On the one hand, Ronald Aylmer Fisher almost single-handedly founded many modern statistics and the fundamentals of trial methodology. The scientific insights gained by those using his methods, particularly around food production and in developmental economics of fighting poverty, have had huge positive impacts worldwide. Nevertheless, Fisher himself is deeply controversial because of the racist, colonialist and eugenicist attitudes he held throughout his life. Therefore, despite his scholarly reputation, UCL's Department of Genetics, Evolution and Environment chose in 2020 to rename centres previously named after Fisher. (More detail and further reading on all these matters can be found at UCL Division of Biosciences, 2024; see, notably, Bodmer et al., 2021).

As Pearl and Mackenzie (2018, p. 147) point out, quality RCTs maximise our confidence that scientific research 'asks nature the right question', meaning that they reduce the risk of scientists inadvertently testing the effect of some other unacknowledged factor on outcomes. RCTs provide answers to the question of how to establish causes in contexts where a bewildering number of background factors occur alongside the focal intervention. An RCT allows us to evaluate the impact of the intervention on the outcomes of interest, because the risk of intrusion by extraneous factors on outcomes should, on balance, be equal across treatment and control conditions, at least if samples sizes are sufficiently large. Balance is likely to be most evident and thus most reliable in a meta-analysis of a series of individual RCTs.

The control condition in RCT designs serves as a baseline from which to assess the impact of any given intervention. This is important, because the distribution of scores on a given outcome measure can change over time for extraneous reasons. For example, reading scores might decline over a summer holiday; they might be affected by other events in a classroom or by measurement inaccuracies on reading tests. In the absence of a control group, there is a risk that changes in outcomes experienced by the treatment group over time will be erroneously attributed to the impact of the intervention. Observing outcomes for a control group allows a fair(er) test of an intervention. Control conditions in RCTs are sometimes termed 'counterfactuals', as outcomes for the control group provide an indication of the outcomes that the treatment group would have attained *if* they had not received the treatment.

Counterfactuals fit the everyday notion of a cause as something that makes a difference by its presence and that is noticeable by its absence, and are also consistent with some of the earliest claims as to the nature of causality (Hume, 1748/2007).[2] Counterfactual-type thinking – 'What would have happened if . . .?' – in the general sense allows an exploration of our notions of causes. Indeed, counterfactual reasoning underpins our everyday moral and interpersonal reflections and attempted self-improvements ('What would have happened if I had not done/said x?' or 'How could I do x differently next time?', and so forth). In our everyday attributions of causes, a host of well-documented systematic biases operate on our reasoning. These include self-serving bias, hindsight bias and even fundamental attribution error – a tendency to erroneously attribute the cause to individual human agents over other less visible but at least equally plausible causal forces (see, for example, Weiner, 1985; Williams et al., 1993). In some disciplines, such as academic history, counterfactuals – like 'What would have happened if Cleopatra's nose was a different shape?' or 'What if a certain technology (such as railways) had not been invented?' – are speculative (albeit fascinating and controversial) mental exercises in our models of historical causality. As we noted above, if counterfactuals in RCT designs are accompanied by strong sampling of participants, they provide a statistical base for the estimate of

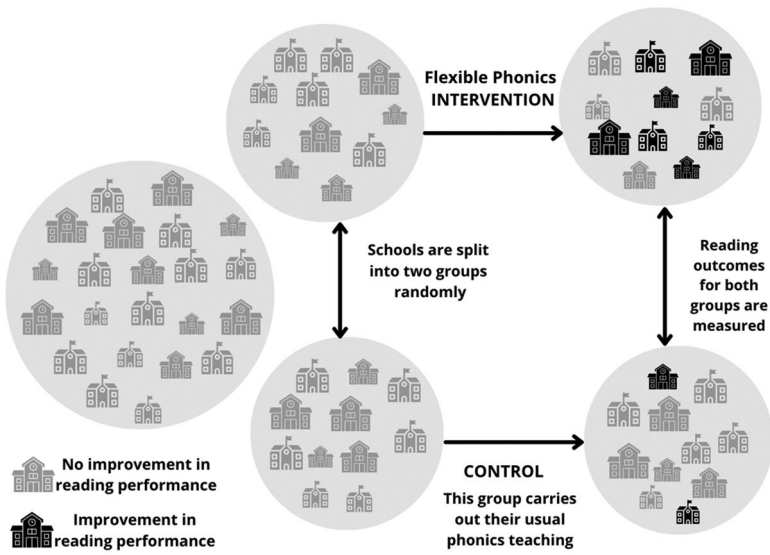


Figure 1.1 Randomised controlled trials as illustrated to teachers and teaching assistants. © Authors.

error attached to any reported effect and, thereby, the robustness of the confidence one can attach to any given effect of intervention reported. See Figure 1.1 for an illustration of how we explained RCTs to teachers in the Flexible Phonics intervention in the UK.

In many RCTs, including this one on Flexible Phonics, the control group receive business-as-usual provision. This means that teachers are asked to continue with their usual teaching practices for the control group. Any effect of the intervention above and beyond standard practice then provides an insight into the impact of replacing the status quo with a new policy. However, if those allocated to the control group become aware that they are not receiving the intervention, there is a danger that this knowledge will affect the outcomes that they experience and bias the estimate of impact.

One possibility is that outcomes for the treatment group are affected by participants' motivation rather than reflecting the intervention content per se. This is known as the Hawthorne effect. There have also been cases of control groups increasing their efforts to avoid being beaten by an intervention condition, which they construed as a rival (the John Henry effect). While protections against such effects are possible even in research designs with business-as-usual controls, studies without such design features are exposed to threats to the validity of causal claims. In some disciplines, most notably medicine, the placebo control serves as a way of 'blinding' the participant (and often, in double-blind trials, the administering researcher as well) to whether the individual has been allocated to the treatment or control group. For example, placebo Covid-19 vaccines were administered in key blinded control studies of the Pfizer Covid-19 vaccine (Polack et al., 2020). Such double-blind approaches are generally impossible in educational research.

An alternative to business-as-usual controls is an active control, which receives an alternate intervention. Such an intervention sets the bar at a different level, as the counterfactual can reflect the best of current practices, for example, rather than a distribution of existent practices of varying quality. We will discuss one such study in Chapter 2 (also see Savage et al., 2018). Chief among the advantages of active control approaches is that they potentially mitigate Hawthorne and related effects. A variant design here is to increase the number of arms of a trial, beyond a classic two-arm treatment–control design, to include passive and active control contrasts to a treatment in a three-arm trial. A good example of this is a recent study of a web-based literacy programme called ABRACADABRA that we have been developing and testing for many years through RCTs (Savage et al., 2013; Bell et al., 2022). In a

study funded by the EEF, Bell et al. (2022) used a typical business-as-usual arm in a multi-site RCT trial, a technology-based literacy intervention supported by classroom assistants. They also included a second intervention, identical to the first except that it was delivered by classroom assistants in a paper-based, rather than technology-based, format. The latter intervention controls for the effects of technology and allows separation of the medium of delivery (technology versus paper) from the 'message' of the intervention content. Results indicated that the paper-based approach was superior, suggesting the mode of delivery was important. Such studies require correspondingly larger samples and are thus also more expensive to run.

Hawthorne effects illustrate a wider challenge for RCTs in education and in all behavioural disciplines – unlike the corn plants in Fisher's original agricultural studies, people are active agents with beliefs, attitudes and behaviours that might not be compatible with the smooth running of a planned RCT. For ethical reasons, all participants in RCTs must have the right to give or refuse informed consent to any research undertaken on them or their children. As a result, parents or carers may seek out alternative interventions or refuse interventions altogether. Children, and indeed their teachers, can also refuse to take part, or they may move to another school or be absent, producing sample attrition. These challenges to the internal validity – non-adherence to intervention and attrition – have been well understood for some time (see, for example, Campbell & Stanley, 1963). Certain measures can reduce or even sometimes eliminate potential design flaws and build confidence that causal links exist. These include mechanisms such as intention-to-treat (ITT), which assesses all participants regardless of whether they received a given intervention, or formal treatment integrity/fidelity assessment, alongside assessing selective non-availability of participants (referred to as 'experimental mortality'), appropriate sampling, and replication (repeat of intervention in RCTs by other teams).

Why are RCTs used in education?

A primary purpose of an education RCT is to ask: does this educational intervention work? In other words, does the intervention produce a reliable, measurable effect on a valid outcome? Does it work at a scale likely to be relevant to policy? These are very good reasons for running RCT trials, as they can innovate evidence-based practice and wider policy for the public good and also evaluate the value for public money of any

proposed new intervention (Haynes et al., 2012). RCTs have been adopted in England by the EEF, which funds the current work we describe, as part of the 'what works' movement to evidence effective English policy and practice since 2011 (Dawson et al., 2018). RCTs are seen – by some at least – as the gold standard for evaluating the effectiveness of an intervention or a teaching practice. The EEF seeks to use research to find ways to improve attainment and life chances for young people across England, particularly children from lower socioeconomic contexts.[3] This is typically achieved by funding scaled RCT efficacy trials in a single site, for example 60–100 schools in one region. If the first trial proves efficacious, this may lead to the funding of further studies of the same intervention scaled at a multi-site level – for example, multiple regions – in subsequent effectiveness trials. The EEF efficacy trial is typically premised on other promising evidence for the intervention being available.

The EEF's process for commissioning and use of RCTs is entirely practical and atheoretical in nature. If the efficacy trial is underpowered, the subsequent effectiveness trial may fail to replicate due to a promising trial bias (Sims et al., 2020). This bias reflects the idea that smaller studies may produce effects that prove unreliable in larger replication studies. It is also possible that a given efficacy RCT result is valid due to general equilibrium effects (Deaton & Cartwright, 2018), with the result that the aggregate case does not go on to resemble the initial RCT trial. Deaton and Cartwright (2018) give an example of such effects from economics: an evidenced increase in the productivity of a given farm yields economic gain for that farmer, but applying the same methods at scale leads to excess supply, resulting in a fall in prices and profit for a nation's farmers. One could speculate whether education's equivalent of such equilibrium effects might be a rise in national reading standards leading to a recalibration of standardised tests. Alternatively, impacts on some outcome measures might be offset by negative effects on others. For example, a sizeable positive effect on reading after sustained teacher-led literacy work might be associated with a smaller negative effect distributed across mathematics, science, social skills, health, motor skills and so on in the same children, and the overall pattern of attainment across all school subjects reflecting only the pedagogical prioritisation a teacher has made. A highly nuanced science of reading intervention would assess such systems matters carefully.

The EEF's approach may be problematic if the wider science behind any funded RCT is weak. Traditionally, scientific funding bodies evaluate the wider science behind proposed RCTs to develop putative causal models. RCTs can, then, because of their potential to illuminate causes,

be used to test causal theories of development. For example, it was argued some years ago that phonological awareness (the ability to reflect on speech sounds in spoken words) was causal in early word-reading development (Bryant & Bradley, 1985). Bryant and Bradley (1985) reasoned that a claim to a causal link is supported if an intervention that taught phonological awareness improved reading ability. As it turned out, attempts to prove a causal link were unsuccessful. (However, based on subsequent work, it is very likely that phonological awareness plays a complex co-causal role in reading development; see, for example, Hulme et al., 2012; Savage et al., 2020). Implicit and explicit claims were later made about many other aspects of language and instruction, such as morphology (the distributed structure of word meanings, where a morpheme is the smallest unit of meaning in a language; Levesque et al., 2021). We make an analogous claim in Chapter 2 about Flexible Phonics, based on cumulative scholarship and research using diverse methods. If the skills taught in Flexible Phonics have an impact on reading ability, teaching them should improve reading over a counterfactual condition. Whether more complex and co-causal patterns occur for Flexible Phonics is also explored.

What are the limitations of RCTs?

RCTs have been the subject of ongoing debate in education and social science (Bonell et al., 2012; Dawson et al., 2018; Deaton & Cartwright, 2018; Lortie-Forgues & Inglis, 2019; Diener et al., 2022). We consider the importance of some frequently articulated critiques of RCTs below. Before we do, the first thing to note is that individual RCTs can vary dramatically in quality. There exist numerous quality indicators for RCTs, such as sampling bias, selective attrition (discussed above) or measurement weaknesses. Weakness in any one of these could potentially produce fatal flaws in the internal validity of any causal claim in any given RCT (see, for example, Altman et al., 2012). We have heard earnest consumers of evidence-based reading research say: 'I only read RCTs!' This is unfortunate, as randomisation is just one feature (albeit a crucial one) of a high-quality RCT. Thus, a well-designed and well-executed quasi-experimental design intervention can potentially produce a higher standard of evidence than a poor-quality RCT. However, a series of well-designed and well-executed RCTs are likely to produce the highest possible standard of evidence. We consider five common fallacies about RCTs below.

Fallacy 1: RCTs are suited only to some broad evaluation questions

RCTs arguably best fit evaluative questions ('Does it work?') or generalised causal theoretical questions about the link between intervention and outcomes (see, for example, Diener et al., 2022). We note here, too, the complex issues of RCT generalisation: the effect on a unit of the treatment may not apply in other contexts. RCTs do not often speak to *how* or *why* an intervention worked or failed in any given context. Nested mixed-method approaches, involving a discussion with teachers about the intervention and their experience of it, as incorporated into EEF trials, do potentially speak to this issue (Siddiqui et al., 2018). It is possible, however, by using other methods alongside a RCT design, to evaluate why, how and for whom a given effect occurred (see, for example, Deaton & Cartwright, 2018). Supplementary methods include mediation, which is the statistical exploration of the pathways through which a dependent variable of treatment versus control operates; moderation, which is the exploration of the features that attenuate a given independent variable outcome effect; and quantile or nested (cluster) causal path models to explore individual and sub-group response to intervention. Clear, theorised mediation within an RCT provides particularly compelling data on the possible cause. Bonell et al. (2012) refer to these latter approaches as realist RCTs that might also test theory and generalisation by exploring effects in a range of deliberately contrasting contexts. It should be clear that it is erroneous to claim that RCT studies can always only report average effects. We will pick up this theme on mediation again in Chapter 3, when we discuss the Theory of Change model in RCTs.

In contemporary psychological theorising, the RCT is only one tool among many in building the research base for a candidate causal variable. Psychological research at its best is systematic and cumulative. A pattern reported in the work in Chapter 2 on the history of the abilities underlying Flexible Phonics is illustrative: here observation and conceptual work were followed by empirical correlational and then longitudinal and experimental studies, and only later by small-scale intervention work, eventually leading to larger scaled and replicated RCTs. Increasingly, such basic cognitive science work is augmented by computational, neuroscientific, case study and other interdisciplinary insights. RCTs are one tool among many in the establishment and testing of causal models.

Fallacy 2: RCTs evaluate only simple causal models

The classic design we describe above involves a two-condition RCT evaluating a single treatment–outcome link. It would be a mistake to assume, as some do (see, for example, Diener et al., 2022), that RCTs are limited to such cases. Complex co- and multicausal patterns that likely occur in many real-life contexts can be explored in more complex RCT designs. RCTs with multiple treatments within the same trial can estimate a response surface – a range of response dynamics to distinct aspects of a treatment (Shadish et al., 2002). Methodologies such as multiphase optimisation strategy (MOST) can be used to build, optimise and evaluate such multicomponent interventions (Collins et al., 2007). This approach contrasts different permutations of multiple intervention components with a single control group. To our knowledge, MOST and related designs have occurred in engineering and life science research but have not yet been taken up in education contexts such as reading. Another approach is sequential, multiple assignment, *randomised trials* (SMART), which involve randomisation at different points in a trial, allowing for the nuancing of intervention and response based on pre-RCT specifications of active processes (Collins et al., 2007). Finally, a recent development is mega-studies. A mega-study is a massive field experiment in which the effects of many interventions, which are often only subtly different, are compared in the same population on the same objectively measured outcome at the same time (Milkman et al., 2021; Royer, 2021). Such studies can provide data on the optimal components of complex interventions, but are extremely large and expensive.

None of these approaches are pursued here. Flexible Phonics is a multicomponent intervention with specific theorised active processes, which we elaborate on in Chapter 2. However, the predicted pattern of main effects and interactions and mediation, alongside the distribution of effects across attainment levels (and against known differentiation of activity), should be sufficient to confirm the active ingredients and where and for whom they have an impact.

There is another important sense in which RCTs may lead to oversimplification. Their very strength in controlling, through randomisation, for the effects of extraneous variables might be their greatest potential weakness in that it may lead the unwary to conclude that these other variables are unimportant. As a result there may be a deceptive simplification of the landscape, with an overly strong focus solely on RCTs and other robust forms of enquiry leading to these wider variables being ignored. The unwary are particularly likely to be seduced

by RCTs where positive short-term effects are evident in RCTs that seem to offer powerful and direct pedagogical 'quick fixes' to long-term structural social problems, such as low achievement and social class. Failure to move the needle on social inclusion through policy implementation, based on promising RCTs, could indicate such issues at play. We see no contradiction, however, between using RCTs judiciously to improve education, on the one hand, and the pursuit of equity-focused, inclusive and democratic school systems on the other.

Sometimes one detects the concern that RCTs will take over and dominate the educational world. This is not well founded. RCTs remain rare in education and account for a small fraction of the total educational research output. For example, we and others who undertake meta-analyses systematically seek out all RCTs on a given question. Routinely, we find thousands of potentially relevant articles, but when selecting carefully for methodology we often find this reduced to only a dozen or so studies. Key fields where the evidence base is surprisingly modest include phonics, comprehension, reading technology and teacher professional development interventions. In our analysis, controlled intervention studies (of which RCTs are a subset) represented less than 10 per cent of the output even in the journal *Scientific Studies of Reading* between 2006 and 2016. The growth in bespoke RCT funding mechanisms, such as the What Works Clearinghouse in the USA and the EEF in the UK, in the mid-2000s must be viewed in the wider picture of relatively low use of RCT designs in all educational research.

Fallacy 3: RCTs are uninformative in statistical terms

Deaton and Cartwright (2018) draw attention to prior knowledge wasted in RCTs because, unlike some other designs which build on established knowledge, often each RCT starts afresh in its analysis and assumptions. However, this is not always the case. The use of pre-randomisation sample stratification techniques as an aid to the precision of estimates is one common prior. However, prior knowledge can also be built into RCTs by adopting Bayesian conceptual and statistical models that, unlike classical Fisherian models, start with the notion that a statistical decision in any given case is affected by relevant preceding activity affecting the current context. Bayesian analyses thus establish a prior that is used to frame the scope of statistical analysis. One might use the metaphor of such a prior being to the statistical decision-making process like having a rough compass direction (for example, knowing to travel somewhere between northeast and north-northeast) over having no direction whatsoever, as in a Fisherian case.

Using a prior, especially an informative prior, allows sequential trials to accumulate evidence, rather than starting afresh each time (see also Wasserstein et al., 2019). There is, however, sometimes a tension between the greater precision provided by some sorts of adjustments for prior knowledge, and the process of randomisation (Deaton & Cartwright, 2018). In the present case, the Flexible Phonics trial was not a direct replication of a trial, so there was no exact posterior distribution to follow in quantifying a prior. However, there were adjacent (comparable) trials, so a weakly informative prior was constructed using the mean of the reported effect size and using the sampling uncertainty around that estimate as a proxy for the Bayesian uncertainty. This approach is in line with EEF policy. An astute reader may have noted that this work on priors is about how we design and analyse RCTs most appropriately, not whether RCTs are inherently uninformative.

Lortie-Forgues and Inglis (2019) point to the weak overall pattern of results of RCTs. They assessed the effects found in large-scale RCTs commissioned by the EEF (UK) and an equivalent US body, the National Center for Educational Evaluation and Regional Assistance. Both agencies evaluated interventions aimed at improving academic achievement in kindergarten (Reception, age four to five, in the UK) to Grade 12 (Year 11, age 15–16, in the UK) – some 141 RCTs with 1,222,024 students. The mean effect size was 0.06 standard deviations (SDs) within relatively large confidence intervals (mean width 0.30 SDs). This, they argue, based on the view that intervals which include zero are uninformative, means that most results were typically uninformative. The very few positive effects across all trials do not speak to an efflorescence of Hawthorne-type effects, but do show how hard it is to find positive effects in rigorous trials. We also note, on informativeness, that some leading voices in the American Statistical Association explicitly caution against using confidence intervals as 'a new $p < .05$'. Wasserstein et al. (2019) note that confidence intervals indicate only which value is and is not in a given variable range. Beyond purely statistical methods, small-scale preliminary qualitative studies may give a range of important insights into how candidate RCT interventions are played out in complex professional contexts where no prior study exists. We collected such pilot study data here.

Fallacy 4: RCTs are unethical

Unarguably, there have been some egregious examples of unethical scientific research in the past. In psychology, of the studies undergraduates learn about as examples of potentially unethical research, the case of

'Little Albert' is often highlighted as an egregious example of unethical behaviour. This young boy experienced a deliberately conditioned fear response via electric shocks, with (apparently) no ethical control or deconditioning in place (Watson, 1919). Zimbardo's prison simulation experiment (see Zimbardo et al., 1974), where students acted as brutal guards or prisoners in assigned role play, is another example. While some of this psychological work has the trappings of science (white-coated 'scientists', claims to 'science', use of scientific words like 'experiment' and so forth), none of this work is arguably scientific research per se. Certainly, none involved RCTs. Considering the harm that observational, role-simulation and case study methods such as those mentioned above have arguably had, this should signal a warning that *any* methodology can be grossly unethical if applied unchecked.

Today, ethics committee certification is mandatory before any educational or psychological research, including RCTs, can be undertaken by universities. The potential risks and harm, in comparison with the potential advantages of any proposed work, are considered by an expert independent ethics committee in relation to a formal code of research conduct, such as that of the British Psychological Society, and considerations of data protection and informed consent for all participants. (The study reported here was subject to close ethical scrutiny by a university ethics review committee prior to receiving formal approval.) The impacts of consent processes on the sample attainment process mean that results in any research, including RCTs, may not generalise to real-world contexts where such consent is not mandatory. Such issues may be particularly acute for researchers investigating reading, a field where written consent – from sometimes low-literate parents of low-literate children on necessarily long printed forms – may be particularly hard to obtain. Informed consent is also lost if a parent misunderstands a study but signs consent forms nevertheless.

There are many situations where RCTs would never receive ethics approval. One cannot ethically explore the effects of withholding language exposure or instruction of known efficacy, or of withholding early attachments or childcare, to name but a few. However, even in such contexts, methodologies that come as close as possible to RCTs – such as natural experiments, statistical modelling using causal path mediation approaches, and various post hoc matching techniques – are used in reading and other research. In the same vein, the UK Medical Research Council's Complex Interventions Framework (Skivington et al., 2021) can be used when RCTs are not appropriate. Computational simulation has also become a powerful tool in the reading researcher's armoury;

one can deliberately denude the 'environmental' experiences of a word-learning computational network to study its impact (for example Harm et al., 2003). Nevertheless, there also remain very many contexts where RCTs can be run in education with business-as-usual controls, particularly in domains relevant to improved policy and practice.

A more informative contemporary question regarding RCTs is whether withholding a treatment or intervention from a group that might benefit from it is still in some sense inherently unethical. One does not know confidently if an intervention really works better than existing practices (or indeed if it is safe) before running the relevant RCT. In the medical RCT literature, this is termed an uncertainty, or the clinical equipoise principle (Callréus, 2022). We do often know that business-as-usual activity does not solve issues of inequity and exclusion, so doing nothing may have its own ethical costs. For example, doing nothing was not an option in the evaluation of untested Covid-19 vaccines in 2020 and beyond. One might see the same imperative as equally relevant to interrupting the well-documented cascading negative effects of low literacy on life chances, health and even life expectancy. This (moral) imperative drives our attempts to improve on 'business as usual'. Arguably, one might usefully ask two questions here: would you take a key medical treatment, say a vaccine, that had been 'proven' by a non-experimental design where RCTs were ignored? I would not! If so, why then is education fundamentally different? It is thus troubling to hear, as we once did from eminent scholars at an ancient university, that RCTs are an 'unethical, discredited, and obsolete' methodology.

In the field of social policy, it is not possible to offer a treatment to everyone who might be eligible. Allocating individuals to the treatment at random may be considered more ethical than other ways of deciding who receives the treatment. If the RCT suggests the intervention is effective, decisions on whether resources should be redeployed from other activities can be based on evidence. Arguably, that is more ethical than continuing to fund some things indefinitely when other interventions might be more effective. It is in this spirit that the Flexible Phonics trial operated.

Fallacy 5: Homogenous interventions are suboptimal in the educational sciences

A more compelling limitation of educational RCTs, unlike many medical trials, is that the intervention is most often a novel classroom intervention. Unlike some medical trials, the educational context does not allow for the construction of equivalents of one universal intervention, represented

most visibly by a pill, vaccine and so forth. Changes in teaching practice after professional development activities given to a randomly selected intervention group of teachers, such as that favoured by EEF, can vary for all sorts of reasons: teacher willingness to adapt their practice, their understanding of a brief intervention and the perceived social validity of the intervention, where some teachers might not accept an intervention as appropriate even if it is evidenced.

More fundamentally, children are at different developmental points. In Reception classes, there are generally some children who have already mastered a given curricular ability (such as decoding words with phonics) on arrival at school in September, whereas others have not mastered such decoding. There are often up to five years of variance in reading ability in any typical Reception class. We know from meta-analytic work that strongly differentiated teaching, adjusting literacy learning practices to the entry-point abilities of children, is highly effective (Connor et al., 2009; Puzio et al., 2020). One might ask: where does such a picture leave an RCT with a 'pill-like' intervention – one which is discrete, homogenous and universal?

In many cases (especially in the UK), teachers are not supported in understanding why an intervention is better practice (or the theory and rationale on which it is based). Too often, teachers are simply told to do it, which can result in a battle of wills. Professionally, in our experience, teachers are often like magpies – picking up useful bits of practice from diverse sources and using them – instead of relying on formal training. Furthermore, there are often structural constraints, such as the need to fit into assessment structures (for example, the Office for Standards in Education, Children's Services and Skills, or Ofsted), and an approved approach or other influential expert voices in school, or government emphasis. We have come across all of these. One can take account of treatment integrity – the degree to which an intervention protocol was actually followed in practice – but this is often fundamentally not binary (Savage et al., 2013). It might be better to acknowledge all this treatment complexity. Teachers are also often strongly constrained by time. There is a need for simplicity, for getting practitioners on board in busy professional contexts. Here, the minimal change that makes a difference on the ground is the goal. We note that exactly these practical imperatives that motivate our intervention here also drove the Covid-19 RECOVERY trial (Samarasekera, 2022).

Conclusions

We have argued that RCTs have unique methodological strengths in the domains of hypothesis testing and programme evaluation that can withstand many criticisms levelled at them. Their main strength is in supporting causal inference through randomisation. However, the control that RCTs offer for extraneous factors may lead to the simplification of the sustained, complex, multiple systemic and distal causes that may exist in the real world – forces such as poverty and inequity. RCTs are at their strongest when they form part of a mosaic of methods for scientific understanding; when they are closely linked to strongly theorised interventions often based on non-RCT methods; and when they are augmented with statistical path analyses, including mediation and qualitative pilots and process analyses, and pre-specification of key processes and functions within an intervention. Strong RCT designs, including the Flexible Phonics trial discussed here, have much in common with realist RCTs (Bonell et al., 2012).

We have also argued, however, that educational RCTs may need to be reconstrued. While we can always obtain indices of average intervention effects from a given RCT, we are better off viewing the delivery of educational change as being inherently dimensional and not uniform or akin to a pill in pharmaceutical research. We also strongly suspect that impact will be maximised through complex, differentiated interventions with the breadth of content to accommodate a range of reading levels. Futhermore, we can predict that teachers will vary in their uptake of any intervention. Without variation, the intervention will likely not succeed (as well) at scale. Homogeneity is not, even in principle, desirable. Consider also, for example, a 10 per cent improvement in cancer survival rates. This may not add up to a significant overall effect in an RCT, but it will nevertheless be crucial for those individuals affected. Equally, a 10 per cent improvement in treatment non-responders in a novel reading programme potentially contributes to theory and practice, especially if linked, prior to the start of the data collection phase of a study, to strongly theorised aspects of individual variation. RCT intervention designs that take account of all human variation within an RCT are needed to achieve this.

Any teacher professional development delivery should, we argue, start with the assumptions, as above, of inherent and desirable treatment heterogeneity. It can (and should) be based on a clear and potentially falsifiable Theory of Change (a pre-specified model of exactly *how* an intervention impacts an outcome measure), a change model that can

accommodate treatment effect heterogeneity across multiple settings (see Athey & Imbens, 2017, for a detailed statistical treatment of some of these issues). This also fits with stronger mediation-based models of RCT analysis (Bonell et al., 2012). Another related, but distinct, reasonable assumption when construing professional development is that teachers have greater expertise in their classrooms than anybody else (Boldrini et al., 2023). Thus, here, we explore the theory and practice of multi-agency collaborative working with partner schools in our RCT, wherein teacher knowledge is embraced. We will have much more to say on specifics in future chapters.

Educational RCTs can be reconstrued to embrace the apparent methodological flaw of non-homogenous, non-blinded intervention as a strength. Mentoring and professional development is, after all, an inherently human activity. While such work is not double-blinded, many trials, including the one discussed here, fully separates the delivery team (here, Savage and Fox, n.d.), responsible for the intervention, from an entirely independent evaluation team (here, Dawson et al. 2024). Given this control, one is left with potentially confounding expectation effects in non-blinded RCTs. As noted earlier, active control interventions are helpful tools here and, again, mediation-based models also help even in their absence: expectation-based effects, including generic Hawthorne effects, do not predict a theorised mediation effect or longer-term effects of intervention in delayed post-intervention tests. (Hawthorne is construed as a short-term motivational advantage only.) Mediated RCTs are rare in EEF trials, but are used here, we think, for the first time. We believe this more dimensional conception of effective intervention and assessment potentially contributes to a still fledgling implementation science of reading – the study of the implementation of evidence-based educational trials.

Notes

1 For more detail, see Fisher (1935).

2 It is worth noting, however, that while Hume first described human understanding of causality in 1748 as the 'constant conjunction' of any events A and B, and wherein by way of counterfactual the absence of event A is strongly tied to the absence of event B, his greatest contribution to epistemology (the philosophy of theory of knowledge) was to also show that this notion of constant conjunction means that causes so often widely assumed to link events A and B are invisible and unknowable to human minds, which operate in a world only made up of event associations.

3 See https://educationendowmentfoundation.org.uk/

2
Introduction to Flexible Phonics

This chapter presents the scientific background to the content of the Flexible Phonics intervention, a large-scale randomised controlled trial (RCT) that took place in 118 schools in England in 2020. This sets the stage for a detailed assessment of relevant interventions worldwide that used a range of approaches consistent with the Flexible Phonics intervention to teach children to be better readers. The review considers the most recent scientific work available on early literacy education. Prominent gaps in the research literature are noted, and the need for policy-relevant school-based trial data on Flexible Phonics is highlighted. This chapter serves as a precursor to our extended consideration of the Flexible Phonics trial in Chapter 3 and beyond. We first consider some relevant concepts and principles before turning to the specific existing science base.

Principles behind effective early literacy pedagogy

How might we construe any optimal pedagogy for literacy? One pragmatic starting point is to consider the task children face in learning to read English (Savage, 2019). The learning task for a child is to internalise the spelling system of English, with all its myriad complexities and the myriad of links between written forms of words, their phonology and semantics. As soon as this neuro-cognitive representation is partly established, a child can start to use the English spelling system to understand and communicate meanings in print. Development is synergistic, reflecting better and more elaborate lexical representations of multiple aspects of word knowledge through productive engagement with print (Perfetti & Stafura, 2014).

Before turning to details about Flexible Phonics, a note on context. Our primary pedagogical focus in Flexible Phonics is on accurate and fluent word reading. However, while Flexible Phonics primarily involves

attention to words, it also involves close attention to the role of word meanings and thus contributes to wider comprehension goals. The strategies of resolving phonic and semantic/syntactic inconsistencies through active sentence monitoring and word pronunciation repair through self-correction may improve sentence comprehension. This sets the stage for the development of wider comprehension. Indeed, Flexible Phonics does this more so than current synthetic phonics approaches. Our focus on word reading does not in any way diminish the importance of wider language comprehension and language pedagogy. Both fluent word reading and linguistic comprehension are necessary but not sufficient features of reading comprehension (Gough & Tunmer, 1986; Kirby & Savage, 2008). Other interventions show that a joint pedagogical focus on decoding and comprehension in Year 1 (age five to six) improves reading to a greater degree than more narrowly focused pedagogies do (for example Vousden et al., 2022). Nevertheless, there are also asymmetries to child development; children typically start learning to read with some substantial language ability established and reading may at least in part co-cause the development of wider language skills (for example Duff et al., 2015). Reading is also complex, and we therefore focus on one part of the support for reading comprehension: word reading and fluency competence. Later work might usefully explore integrating a successful Flexible Phonics approach within effective wider language comprehension teaching.

A fundamental aspect of the child's task of internalising the spelling system of English concerns the structural features of the English spelling system (orthography). English is widely acknowledged to be an 'opaque' or 'quasi-regular' orthography, wherein grapheme (letter or letter cluster) to phoneme (speech sound) correspondences (GPCs) are often inconsistent, producing lots of exceptions to and contextual complexities within supposed phonic 'rules' (Seidenberg & McClelland, 1989; Daniels & Share, 2018). Partly as a result of this structure, grapheme–phoneme knowledge and word reading take longer to learn in English than in more 'transparent' alphabetic spelling systems such as Spanish, Welsh or Greek (Seymour et al., 2003; Hanley et al., 2004).

The challenge of learning the opaque English spelling system is quite widely acknowledged. A further computational problem for any child learning to read is the distributed patterns of words breaching common phonic rules across the opaque spelling system. Corpus analyses of texts analyse the entire body of words and/or constituent GPCs in multiple texts (for example Gontijo et al., 2003; Foorman et al., 2004; Masterson et al., 2010; Vousden et al., 2011). Such analyses provide strong insights

on the computational complexity of the reading task, which then inform pedagogy. Masterson et al.'s (2010) Children's Printed Word Database (CPWD) is an open-access (that is, publicly available) analysis of words in over 1,000 popular published English children's books. Masterson et al. analysed the CPWD corpus and confirmed the frequently reported finding that there are about 100 to 150 very frequent words (for example *the*, *at*, *he*) that occur in most texts. Importantly, and perhaps startlingly, they also found that most words in young children's texts occur once or twice in over 1,000 texts. Similarly, Foorman et al. (2004) report low-frequency counts even in basal prescribed reading scheme texts.

Given there are many thousands of words in English and that many of them are exception words that do not follow common phonic rules, certain computational and pedagogical conclusions follow this observation. First, children will meet many unfamiliar exception words when reading that they will have no obvious way to read, even if they have been taught and have learnt many high-frequency sight word lists. Second, these exception words are both too numerous and too infrequent for all to be learnt in advance by sight. Third, some children, through some mechanism as yet under-researched, manage to span this divide between the task complexity of reading and existing pedagogy (typically using basic phonic rules for blending and sight words) to read texts fluently. It is also important to note here that most polysyllabic words breach strict phonic rules due to the impact of syllabic stress on vowel pronunciations.

This observation about fluency is simultaneously both a key insight and the key scientific challenge addressed in this book. A much more sophisticated model of reading process is needed to understand how some young readers acquire this ability and why some struggle. Pedagogically, the analysis drives us to consider teaching additional generative approaches, particularly for multisyllabic and exception words. In both cases we build on existing research. In David Share's (1995) influential self-teaching hypothesis, children use phonic decoding to have productive engagements with written words in alphabetic spelling systems, not only to sound out words (independently) but also to use this phonic assembly to build detailed item-specific representations of letter strings in words encountered in texts. This growth in item-level orthographic knowledge is, for example, demonstrated by Share (1995) through children reading a meaningful text involving a pseudoword such as *yait*. They decode and are subsequently able to distinguish between *yait* and *yate*. Phonic decoding is an extremely useful generative mechanism that, once mastered, allows children to self-generate successful pronunciations and then detailed orthographic representations of words through wide

reading. This is very efficient learning. These pedagogical assumptions around self-teaching have been formally modelled in computational machine learning algorithms that 'learn' to read accurately (Pritchard et al., 2018; Ziegler et al., 2020).

Such models work well for monosyllabic words in alphabets that follow simple, common phonic rules. The pedagogical challenge, as we describe it above, is to elucidate generative self-teaching pedagogies for all words. Our Flexible Phonics intervention, described in detail in this chapter, is based on two research- and theory-based ideas that are, in this sense, generative, as they build squarely on the context-independent rules of standard phonics programmes to also consider context-dependent patterns (Compton et al., 2014). We aim to make opportunities for learning the spelling system as rich and informative as possible, to maximise the chances of generalisable learning from both supported and independent encounters with print. This chapter describes the scientific evidence that supports this approach and the evidence for the intervention to date.

Box 2.1: The importance of learning efficiency

There exist other pressing task demands necessitating efficiency. Consider available time to learn to read in Reception (age four to five) and Year 1 (age five to six). Classes consist of about two hundred school days in Reception and Year 1. Teachers often allocate a one-hour slot to English teaching each day and, within this, it is common to use 20 minutes or so for daily word-level teaching. This equals 60–70 hours each in Reception and Year 1. In Reception class, it is also important to consider young children's limited attention spans and the number of genuinely fully focused minutes per day a typical child devotes to learning abstract phonic rules, even if physically present in class!

Nevertheless, if a child is not a competent reader by the end of Year 1, the risk of falling behind is high. Longitudinal studies show that those at the lower end of literacy in the first year tend to remain so at the end of Year 6 (Verhoeven & Van Leeuwe, 2008). It can be an uphill battle, with low motivation and increasing socioemotional issues, to normalise reading once children show reading delay. Efficiency at the start of learning to read is therefore key in any year. In 2020–1, when we ran this Flexible Phonics intervention, the immediate effects of school closures due to the Covid-19 pandemic were an estimated learning loss of two to three months for literacy (Thorn & Vincent-Lancrin, 2021) and school reports of widening gaps between children based on features of family background (Achtaridou et al., 2022). Our emphasis on efficient learning pedagogies could not be better timed.

The components of Flexible Phonics

Flexible Phonics has two main constituents: Direct Mapping and Set-for-Variability.

Direct Mapping

A first, key idea is the teaching of Direct Mapping of letter to sound information to aid efficient word decoding. We present the evidence and models relevant to the approach before considering what is known about how to teach it (for example Savage et al., 2020).

Direct Mapping speaks to the way GPCs are taught. We follow the literature in defining a grapheme as a letter or group of letters that represents a phoneme, and where a phoneme is the smallest unit of sound in spoken language. Such correspondences are taught routinely in classrooms in standard phonics in most government-approved programmes in England and, indeed, worldwide. Much evidence suggests that one component of effective early reading programmes is expertly and systematically delivered phonics (see, for example, National Reading Panel, 2000; Torgerson et al., 2018; for nuanced and recent reviews of this quite extensive meta-analytic literature, see also Fletcher et al., 2021; Savage, 2022). Systematic synthetic phonics is the dominant approach in England, where this Flexible Phonics trial took place. Systematic synthetic phonics involves the assembly of pronunciations through the synthesis of phonemes representing graphemes in print (for example *c-a-t* to make *cat*). Therefore, children need to know GPCs to use phonics. However, until quite recently the specifics of which, how or even how many GPCs are taught has received surprisingly little scientific attention (though, see Shapiro & Solity, 2008; Vadasy & Sanders, 2021; Møller et al., 2022).

In Direct Mapping, for any given GPC taught in a phonics lesson, children should *on that same day* and ideally as soon as possible – that is, *within that same lesson* – also read text that richly embodies the taught GPC, to consolidate GPC learning and link it unambiguously to the task of text reading (Chen & Savage, 2014; Savage, 2019). The idea of temporal contiguity in learning GPCs is arguably classically behaviourist in spirit, but the Direct Mapping approach also ensures that the declarative GPC knowledge (that is, facts, such as *b* makes 'buh', which are standard in most phonics programmes) is immediately elaborated through procedural use (that is, use of the GPC knowledge in sounding out words and reading). Direct Mapping thus aids learning and retention. In practice, a child's early text reading happens through

shared reading of texts, carefully supported by a teacher. However, over time the child's cumulative use of GPCs learnt through Direct Mapping, both currently and previously, should lead to the child taking on more of the text reading task.

The broad idea behind Direct Mapping is not wholly new, but instead builds on cumulative science worldwide. In the US, Jenkins et al. (2004) have shown that, when children are learning phonics, reading development is not damaged by early exposure to real (such as, typical) children's books as opposed to the simpler, decodable texts mandated in many schemes. Turning to GPCs, the influential phonological linkage hypothesis (Hatcher et al., 1994) suggests that to become fluent readers, children must link graphemes to phonemes and then, with increasing automaticity, to text reading. Hatcher et al. (1994) report an intervention in England contrasting phonics and book reading with other conditions where either phonics alone or book reading alone were taught, or where neither were provided. The results showed an advantage for children taught both phonics and book reading. In the intervention reported on by Hatcher et al., the children experienced phonics and shared book reading over the course of the intervention, but these two components were not integrated at a fine level. In Direct Mapping we link each GPC taught to texts read on that day, consistently and intentionally.

Tse and Nicolson (2014) ran an intervention in school contexts of low socioeconomic status in New Zealand. A 'big books' approach was supplemented with phonics, and the graphemes taught were systematically linked to the text of the shared big book (a book big enough to share with a group). Significant advantages offered by this combined intervention, in contrast with interventions delivering phonics alone or big books alone, were evident on all reading outcomes. Results here are consistent with a Direct Mapping interpretation, though it is possible that the effects reported reflect the interaction of GPCs linked to books *and* a 'turtle talk' phonic activity in which children identified words with 'stretched' component graphemes (where sounds extended over time, for example 'bbbbb' for 'b').

The Simplicity Principle

We have already considered the complexity of English and the need for efficiency in the task of mapping out the spelling system, as well as the sparsity of evidence on which and how many GPCs to teach for optimal effects. One approach could be to extend the logic of corpus analysis for words to also consider component GPCs. In a rational analysis, one can select the most frequently occurring GPCs in children's texts and teach

them in rank order of frequency. These will give children the capacity to read the greatest amount of text for the initial effort of learning GPCs (Vousden et al., 2011). As Share (1995, p. 156) puts it, to facilitate self-teaching one needs the 'minimum number of rules, maximum generative power'. Corpus analyses by Vousden et al. (2011) have identified the optimal and most explanatory GPCs that avoid diminishing returns for lower-frequency GPCs – the Simplicity Principle.

Simplicity-based corpus analyses often reveal that complex, cluster GPCs (for example *sh*, *ch*, *ea* and *oo*) are more frequent in texts than some singletons such as *x* and *z*. It is, however, a common pedagogical practice to assume young children should be taught all singleton GPCs before progressing to the more complex cluster GPCs. Is this early exposure to graphemic complexity arising from the pedagogical use of the Simplicity Principle problematic for children? Evidence from an intervention by Vadasy and Sanders (2021) suggests not. Vadasy and Sanders showed that young children benefited more from early exposure to mixed-sized GPC instruction (singletons and more complex multi-letter GPCs) than from an exclusive early focus on singletons in their early reading (see also McGuiness et al., 1996).

Turning directly to simplicity-based interventions, Shapiro and Solity (2008) taught phonics to typical readers using selected GPCs based on the Simplicity Principle; they reported positive effects on reading when compared to business-as-usual teaching. However, it is not possible to attribute these effects as being due to Simplicity Principle GPCs directly, as the intervention also included taught sight words, distinct synthetic phonics routines and distributed learning, and thus differed in multiple other ways from the business-as-usual control. Chen and Savage (2014) ran an intervention with Grade 1 and 2 children (age six to eight) from one school in Quebec, Canada, who were at risk of reading difficulties. In the treatment condition, complex simplicity-based GPCs were first taught, and children were then supported in the shared reading of texts selected to have a high density of the GPC taught moments before. Control group children were taught vocabulary 'word usage' – how to use the words appropriately in sentences (for example the use of the word *says* when another person is speaking and in singular and plural form). The results showed large advantages at post-test on word reading and spelling for the intervention group compared to the control group. Additionally, one aspect of motivation – children's perception of task difficulty – was improved, possibly signalling a virtuous circle of raised reading attainment improving reading motivation.

To further specify the effects of the Direct Mapping intervention, it is important to establish whether Direct Mapping improves reading compared to a control intervention where the same printed graphemes are an explicit focus but where GPCs are not taught. Thus, we compared an intervention where the complex GPCs and their pronunciations were taught and explicitly reinforced in children's books on the same day with a near-identical taught control intervention (Savage et al., 2020). The control differed from the intervention only in the sense that the same grapheme letter cluster was linked to letter names (but not phonemes). Additionally, there exists strong evidence that decoding itself is based upon the dual foundation of both phoneme awareness and GPC knowledge (for example Byrne, 2000). Therefore, the moderating effects of pre-intervention phonological awareness on children's response to a GPC intervention were explored post-intervention. We predicted an interaction, but not a main effect of GPC intervention: improvements in reading and spelling for the intervention group when compared with the control group are expected to be bigger where children demonstrate stronger pre-intervention phonological awareness. We ran an intervention with 149 children at risk for poor reading in Grade 2 (age seven to eight) in Canada, delivering around 12–15 hours of small-group intervention in each of the two conditions. Post-intervention results showed the predicted significant interaction effect at post-test favouring the simplicity intervention on word reading, spelling and sentence comprehension (Savage et al., 2020). Finally, Yeung and Savage (2020) ran a very similar intervention in English (as an additional language) for Cantonese-speaking Year 1 and Year 2 children (age six to eight) in Hong Kong. Yeung and Savage reported very similar positive effects of Direct Mapping to those of Savage et al. (2020), with better outcomes on word reading, spelling and reading comprehension, moderated by pre-intervention phonological awareness.

In sum, there exists a modest body of evidence supporting the use of Direct Mapping in early reading teaching. Direct Mapping most often produces measurable effects for some struggling readers in grades 1 and 2 (aged five to seven years). Some children did not improve in reading. Improved interventions inspired by Direct Mapping will need to also include support for phonological awareness development. As yet, it has not been established that simplicity-based progression is superior to other types. There is, however, enough evidence already to situate it as one component strategy among others in a multicomponent intervention. Simplicity alone does not govern all aspects of GPC content; the Direct Mapping approach is the more overarching principle here. As we describe in the next section, other principles driven by other features of Flexible

Phonics apply to GPC selection. Most prominently, we wish to expose children to the common variability that they will encounter. A further distinct feature, particularly for struggling readers, concerns the ease with which GPCs can be blended.

We now turn to the second key idea in Flexible Phonics. This is Set-for-Variability, where these and other ideas are considered.

Set-for-Variability

Starting this subsection with definition(s) and the history of this term is particularly important, as there is not necessarily full agreement on the range of phenomena that Set-for-Variability encapsulates. The term 'set for variability' is often attributed to Eleanor Gibson (1965, p. 1070) and refers to how children must make sense of the inconsistency of letter-to-sound correspondences in English words as they learn to read (see also Gibson & Levin, 1975). Gibson also refers to this facility more formally as 'set for diversity'. Gibson describes a series of early studies, directed by Harry Levin and colleagues, contrasting the effects of exposure to an artificial spelling system that was initially consistent but then varied in vowel pronunciations, versus a spelling system that always embodied variable vowel pronunciations. Effects on the learning of a new transfer word set were investigated in a sample of 40 Grade 3 (age eight to nine) children in the US. Levin et al. (1963) found the group who had been exposed to inconsistency from the start to be advantaged in terms of speed and accuracy of transfer list learning. They theorised that learning a variable list 'created an expectation of learning set for variability of correspondence which was transferred to the second list and facilitated learning it' (Gibson 1965, p. 1070). Little detail was, however, provided on exactly what this expectation in Set-for-Variability consisted of, nor was the precise nature of the transfer observed elucidated.

The report cited by Gibson first uses the term 'set-for-variability' (Levin & Watson, 1963, p. 19). The authors also earlier note the need for an 'approach of variability'[1] in phonic decoding of the inconsistent English spelling system, wherein children 'resound the elements, giving those with variable possibilities other values to see whether the second combination will yield a word'. They further note that 'this process is akin to problem solving' and suggest this might be achieved by a child working down a hierarchy of candidate grapheme–phoneme patterns (Levin & Watson, 1963, p. 4). Levin et al. (1963, p.1), in the same volume, argue that a 'set-for-diversity' 'leads a child to expect new associations to old stimuli'. It is also worth noting that the empirical work reported here was

with simple, singleton letter–sound correspondences only and among capable, older readers already exposed to English (though subsequently, in Chapter 14 of the report, also to Italian and Spanish, where positive effects of inconsistency on transfer learning were less evident).

Venezky (1999) subsequently used the term 'Set-for-Variability' to describe the challenges a learner faces in navigating variable pronunciation of graphemes, and particularly vowels, in written English. Venezky addressed the specifics of graphemic variability in English as part of a wider conception of written English as being a morpho-phonemic spelling system involving both grapho-phonological and morphological (meaning) patterns and 'wherein phoneme and morpheme share leading roles' (Venezky, 1967, p.77). It is from this that the partly consistent patterns of English spelling often derive. Given this, Venezky (1999) asserted that one of the main functions of phonics instruction is to provide a form of problem-solving heuristic: the accurate application of phonics gives children a mechanism for generating approximate rule-based phonological representations of unknown words that will often get them close enough to the correct (such as, standard) phonological form. The result is that, with the provision and effective use of semantic and other word context features, children can derive the correct identification.

Venezky follows Levin and Watson's earlier analysis in considering Set-for-Variability a capacity for active problem-solving of printed-word pronunciations for unknown words, particularly those words containing irregular, orthographically complex or polyphonic spelling patterns, and most especially vowels. Venezky also shares with Levin and Watson the potentially key pedagogical idea that there may be a cyclical aspect to this problem-solving process in terms of matching candidate pronunciations derived from spelling rules to stored pronunciations in memory: 'If what is first produced does not sound like something already known from listening, a child has to change one or more of the sound associations (most probably a vowel) and try again' (Venezky, 1999, p. 232).

In contemporary psychological research, several subtly different definitions of the scope of Set-for-Variability have been offered. Several approaches closely follow and formalise Levin and Watson and Venezky's original conceptualisation. Tunmer and Chapman (2012), for example, outline a model of word decoding where the first step of applying synthetic phonic procedures produces only an approximation to the conventional pronunciation for many words. Tunmer and Chapman (2012) then argue that additional, flexible mental Set-for-Variability is required to achieve correct word pronunciations, particularly where pronunciations derived

from GPCs are quite different from entries in the mental lexicon. These include exception words such as *the*, *he*, *wasp* and *giraffe*, which do not follow cardinal taught GPC rules (see also Elbro et al., 2012). Within this broad view, some researchers have placed particular emphasis on vowel variability in conceptual modelling and intervention (for example Lovett et al., 2000; Steacy et al., 2016).

Alternatively, Set-for-Variability may be required for the synthesis of pronunciations for all words, irrespective of their status as regular or exception, and perhaps maximally so when children start learning to read (Elbro et al., 2012; Kearns et al., 2016; Elbro & de Jong, 2017). At its most basic auditory level, a 'spelling pronunciation' derived from standard phonic assembly procedures (for example 'cuh', 'ah', 'tuh' for *cat*) is a linear string of distinct phonemes. In a word, by contrast, each phoneme is co-articulated or merged such that acoustic analyses of the speech signal, for example for *cat*, shows one long sound rather than three distinct phoneme sounds (see, for example, Liberman & Shankweiler, 1991; Bishop, 1997). Distinct speech features in phoneme strings and spoken words, especially schwas and glottal stops, bring further complexity to speech processing. Schwas are mid-central, neutral vowel sounds that often occur naturally in spoken, unstressed syllables in English (for example in the first syllable of *because* and *today*). In the context of synthetic phonic blending, schwas occur as vowel attachments in articulated phoneme strings such as *c-a-t*, such that they are pronounced as something closer to 'cuh', 'ahh', 'tuh'. Schwas are unavoidable in the sense that they are a result of the mechanics of the production of a class of consonants, such as *c* and *t*, that are plosive stop consonants – phonemes that are produced by the controlled stopping and then sudden release of air under pressure in distinct ways when speaking.

A host of additional complications occur, particularly when we consider pronunciations with more than one syllable. We have already noted the tendency of vowels in unstressed syllables in words like *because* to become schwa pronunciations. In other regular bisyllables often assumed to be regular, such as *trumpet*, the unstressed second syllable sounds more like *pit* than *pet* in continuous speech. Another class of stop consonants – glottal stops, created in the glottis – results in the truncation of phonemes in spoken words such as *kitten* and *umpteen*, leading to distance between spelling pronunciations and conventional pronunciations. Perhaps most fundamentally of all, as Kearns et al. (2016) note, even highly frequent and consistent phonemes such as the /*t*/ phoneme are not invariant: they are subtly influenced by lexical context, as shown in the variants of /*t*/

in *cat* and *stop*. Elbro and deJong (2017) thus argue that, conceptually, a spelling pronunciation is stored as a distinct part of the representation of rich word knowledge in the mental lexicon.

We have only touched upon the range of relevant linguistic issues here, but we have demonstrated several reasons why the synthesis of pronunciations from phoneme strings is best viewed as a complex (and almost certainly abstract and conceptual) task, for all words. Furthermore, this task can be construed as a distinct and often untaught second step in phonic assembly (Elbro et al., 2012; Kearns et al., 2016). It is therefore possible that this is a cognitive bottleneck for at least some children. Kearns et al. (2016) note that while Gibson and many others construe Set-for-Variability as a conscious problem-solving ability, at least some aspects of it may be implicit.

Box 2.2: Set-for-Variability does not involve guessing words

Whatever specific model of Set-for-Variability one prefers, it should be clear that Set-for-Variability is not simply a 'psycho-linguistic guessing game' (Goodman, 1967). In the controversial whole-language approach, Goodman first notes some of the complexities of written English, but then concludes that phonology and GPC are too unreliable to use for reading. Goodman thus claims that reading must become a high-level 'guessing game'. In such whole-language approaches to reading construction, phonics is frequently de-emphasised or eschewed in favour of multiple wider semantic and syntactic language cues, and models of apprenticeship-style learning – by experience rather than direct instruction – are used (see, for example, Wyse & Bradbury, 2022). By contrast, Set-for-Variability is primarily a bottom-up or data-driven model of reading acquisition that is closely linked to the words on the printed page. As such, it builds firmly on systematic phonics use, albeit also employing semantic knowledge and a directly taught acquisition strategy, together with phonics, to render English orthography more fully learnable without any encouragement of guessing. English orthography is not so indeterminate as to require guessing, but it is sufficiently inconsistent to require problem-solving heuristics as well as phonic rules. For these reasons, Set-for-Variability can be incorporated in formally implemented triangle connectionist models as a semantic 'clean-up' mechanism to resolve pronunciation errors emerging from orthography–phonology mis-mappings (Rueckl et al., 2019; Edwards et al., 2021).

In sum, Set-for-Variability is, in principle, a powerful and efficient adjunct to standard phonics which meets our criteria of being a generative self-teaching mechanism for all words, and especially those we identified earlier as imperfectly supported by regular phonic assembly alone – irregular (exception), complex and multisyllabic words. Venezky reaches these conclusions based on linguistic theory and from a detailed, principled linguistic analysis of English (for example Venezky, 1967). He reaches somewhat similar conclusions to those we have from modern analyses of word corpora, pedagogical data and rational task analysis of learning English. However, while these conceptual analyses all support Set-for-Variability, we have not yet considered empirical work. It is to these issues we now turn.

A scientific case for Set-for-Variability

To establish a scientific case for Set-for-Variability, we need to establish its importance in child development empirically. We thus start with evidence from cognitive science, drawn from longitudinal correlational studies and experimental work, before later turning to evidence from intervention research.

There has been a number of studies exploring spelling pronunciations in the context of spelling development. Drake and Ehri (1984) report a study wherein typical Grade 4 children were encouraged to sound out 20 words that included inconsistent graphemes, silent letters and schwas (*president*, *anchor* and *regular*, respectively). This use of 'careful pronunciations' that reflected regularised sounding of standard pronunciations for all variable graphemes and silent letters (for example, more like '*press-i-dent*', '*an-chore*', '*reg-uh-lar*') was contrasted with children exposed to conventional spellings along with dictionary-style phonetic transcripts (for example /prɛz ə dənt/, /æŋ kɚ/, /rɛg jə lɚ/). The group exposed to regularised 'careful pronunciations' showed post-test advantages in spelling accuracy. Positive effects were most marked among weaker pre-test spellers. Drake and Ehri interpret these improvements within Ehri's influential theory that GPC pathways underpin precise representations of words in lexical memory. Subsequent replication work in Dutch and German has provided at least some support for spelling pronunciations as an aid for spelling accuracy, particularly among weaker spellers (for example Hilte & Reitsma, 2007; Thaler et al., 2008). However, the transparency of the German spelling system forced researchers to use imported words that are atypical in German more widely. The focus here was on accurate item learning, not on the creation of a generalisable strategy.

Turning to reading, Elbro et al. (2012) explore the the ability to identify the Dutch equivalent of exception words such as *was* from regularised spoken spelling pronunciations of those words – what they call word recognition from mispronunciation (WRMP). They report that WRMP was a significant unique predictor of both regular and exception word reading in 74 typical Grade 1 children, and that it remained significant even after controlling for known predictors of reading such as vocabulary, rapid automatic naming and phonological awareness. Exception words are quite rare in Dutch, which is a relatively transparent orthography. A second longitudinal study carried out by Elbro et al. (2012) in the opaque orthography of Danish, where spelling–sound inconsistencies are more common, explored the predictive power of Set-for-Variability among 187 Danish pre-schoolers. The children were tested before formal instruction in reading began and followed into Year 1 (age six to seven) to identify causal developmental links. Set-for-Variability was assessed using a new WRMP task wherein children corrected spoken pseudowords created by substituting a phoneme (for example *telefonen* presented as *delefonen*). The results again showed a prediction of later regular and exception reading from pre-school WRSP measures, even after controls for early word reading, vocabulary, rapid automatic naming and phonological awareness.

Tunmer and Chapman (2012) also present longitudinal data exploring the role of Set-for-Variability in conjunction with vocabulary and syntax, following 149 typical children in New Zealand for three years from Year 1 to Year 3 (age five to eight) of elementary school. Tunmer and Chapman employed two measures of Set-for-Variability: the ability to self-correct regularised spellings of exception words, such as *wasp*, in isolation from Set-for-Variability (Set-for-Variability isolation) was assessed against the additional benefit gained from reading the regularised spelling in a relevant sentence context (Set-for-Variability ratio). In Year 1 only Set-for-Variability isolation was a modest predictor of exception word reading in years 2 and 3, even after controlling for the effects of Year 1 measures of reading, phonology, vocabulary and syntax. The results of path analyses showed that Year 1 vocabulary impacted reading comprehension directly in Year 3 and word reading indirectly in Year 3 through Set-for-Variability isolation, with additional, distinct and direct paths for phonological awareness in Year 1 impacting on reading in Year 2 and 3.

Kearns et al. (2016) show that a measure of Set-for-Variability, created from a subset of 25 words drawn from Tunmer and Chapman (2012), had psychometric reliability (measurement consistency) and validity (the test measures what it claims to measure) in a somewhat

older sample of 206 children (aged between seven and 11 years) who were oversampled for reading difficulty. The Set-for-Variability task in Kearns et al. (2016) (an Set-for-Variability isolation measure) was again a significant predictor of word reading accuracy and reading fluency, even after the effects of vocabulary, phonology, morphology, attention difficulties and working memory were considered. Kearns et al. conclude that Set-for-Variability is an important measure of reading development.

Steacy, Compton et al. (2019) explored predictors of pseudoword reading in 96 children in grades 2 to 5 (ages seven to 11) in the US, oversampling atypically poor readers. Less frequent pronunciations of variable vowels (for example *chead*, pronounced as rhyming with *head*, over-pronounced as rhyming with *bead*) were predicted by Set-for-Variability and reading level, where reading level likely measures exposure to variant vowel pronunciations. A study by Steacy, Wade-Woolley et al. (2019) offered an analysis of Set-for-Variability at an even finer grain size. They explored performance on Set-for-Variability tasks and other child-level predictors (phonological awareness, rapid naming), alongside item-level predictors – word-specific transparency of phonology-to-pronunciation correspondences as rated by language experts, word concreteness (the degree to which the referent can be experienced by the senses), length and frequency – in predicting the use of less frequent vowel pronunciations of pseudowords in 103 children in the US in grades 2 to 5, oversampling atypically poor readers. Steacy, Wade-Woolley et al. (2019) found that both word-specific Set-for-Variability indexed by item-specific transparency and child-level Set-for-Variability ability predicted exception word reading, even after controlling for phonological awareness and overall reading ability.

Edwards et al. (2021) extended this analysis to explore the component structure of Set-for-Variability in 489 children in grades 2 to 5 in the US and to assess the role of spelling knowledge in Set-for-Variability task performance. They measured the degree to which a regularised pronunciation of a word via a spelling pronunciation could nevertheless yield the correct spelling of an exception word. For example, assuming in both cases that wider consistency of phoneme–grapheme rules is important, deriving and storing a spelling pronunciation for *lamb* would likely yield a correct subsequent word spelling, whereas deriving a spelling pronunciation for *rhythm* is unlikely to. This is because of the high phoneme–grapheme consistency for *b* in *lamb*, and the low phoneme–grapheme consistency for *rh* in *rhythm*. Edwards et al. (2021) found Set-for-Variability to be the strongest child-level predictor of word reading ($r = .79$) and that, at the item level, a role for wider orthographic (word

spelling) knowledge in Set-for-Variability was suggested, above and beyond that of phonology. Such patterns were evident only in the better readers, suggesting that more skilled readers bring wider knowledge of candidate GPC probabilities to bear in Set-for-Variability tasks.

There also exists experimental evidence of real-time processing of words, suggesting that children use Set-for-Variability to resolve word pronunciations. Murray et al. (2022) taught 78 typical Grade 5 children the oral pronunciations and invented definitions of half of a list of pseudowords, such as *vaik*, defined as a tool for cleaning fish tanks. Pseudoword pronunciations were taught to render the words either regular (for example *vaik – vake*) or irregular (for example *vaik – vike*) when children subsequently encountered them in print. Children then read the words in contextually supportive or neutral sentences (for example 'The fish in the dirty tank swam around the vaik as it worked' vs 'The boy in the blue shirt walked around the vaik as it worked'). Eye movements were monitored during sentence reading. Murray et al. (2022) found that fixations were shorter on regular pseudowords than on irregular ones, suggesting a 'processing cost' for resolving irregularity. There were larger differences between irregular and regular words when encountered in the contextually supportive sentences, suggesting that orally known irregular words undergo additional processing when subsequently met in print. Children were able to read irregular words more accurately after they had encountered them in a supportive sentence context. Arguably, this latter finding fits with the argument for the type of self-teaching mechanism for exception words we have discussed earlier as being essential for the internalisation of the English orthography. Murray et al. (2022) conclude these findings are consistent with the 'online' operation of a mispronunciation correction mechanism.

Evidence from an emerging intervention science of Set-for-Variability

The evidence strongly suggests that Set-for-Variability plays an important role in literacy development. However, for a science of Set-for-Variability intervention, we also need to show both that Set-for-Variability is teachable and that teaching Set-for-Variability measurably improves reading more generally. Ahead of such analysis, we note that Set-for-Variability has already been incorporated into several established intervention programmes in North America. Foorman et al. (2004, p. 173) analysed six basal reader and phonics schemes, widely used from the mid-1990s

onwards, and found that they varied in the degree to which they explicitly taught children to flex their decoding in light of silent letters, schwas, contractions and polysyllables when reading printed words. Children were encouraged to decode, with some programmes devoting significant time to Set-for-Variability. As Foorman et al. (2004) note, however, the earlier curricula that they reviewed, which included Set-for-Variability instruction, preceded any empirical evidence of its efficacy. There is also an established body of intervention work for weaker readers using some aspects of Set-for-Variability (see overview by Scanlon & Anderson, 2020). In Foorman et al.'s (2004) interactive strategies approach, struggling readers were encouraged to use broad semantic information alongside phonics knowledge when resolving word pronunciations of exception words derived from applications of phonics. Here, the use of both lexical and pictorial context was encouraged among struggling readers and were brought to bear in 'confirmatory ways' in word reading. The use of pictorial context is distinct here and is not encouraged in most models of Set-for-Variability instruction, including our own.

Notable also is Lovett's explicitly taught vowel alert strategy in the PHAST programme (Phonic and Strategy Training, later the EMPOWER programme; see Lovett et al., 2000). The What Works Clearing House (2016) criteria for evidenced practices in early word-reading instruction also has a subsection titled 'Teach students to self-monitor their understanding of the text and to self-correct word reading errors' among its best practices and includes a 'fix-it' approach that may, in some contexts, bear at least some broad family resemblance to Set-for-Variability as we describe it here. The PHAST programme has been the subject of several evaluations (for example Lovett et al., 2000). However, it is not possible to attribute the improved attainment shown by intervention over control groups at post-test in some PHAST interventions to Set-for-Variability instruction per se. The vowel alert strategy, which embodies aspects of Set-for-Variability, is only one among several taught strategies in a multicomponent intervention.

To advance understanding of Set-for-Variability, Steacy et al. (2016) explored the effects of teaching the PHAST programme in the US. The programme focused on graphemes, rimes, variable vowels and affixes, including strategy instruction in Set-for-Variability vowel pronunciation and morphological decomposition. In contrast, a taught comparison group received only graphemic-level phonics. The participants were 37 children with identified learning difficulties in grades 3 to 6, randomised to one of the two intervention conditions. The results showed that children in the PHAST programme were more likely to

produce pronunciations of experimental words that included variant vowels at post-test than children in the comparison intervention. This was interpreted as children having improved Set-for-Variability. There was no evidence, however, that the PHAST instruction improved overall reading compared to controls, or that it did so in a sustained way. It is therefore hard to interpret the significance of this short-term item-level biasing of the vowel pronunciations shown here on wider reading development, and some caution in interpreting the multicomponent PHAST programme as evidence of the importance of Set-for-Variability alone also remains necessary.

Zipke (2016) explored the effects of teaching Set-for-Variability to children in grades 1 and 2 in low socioeconomic contexts in the US. Fifteen children each received five one-to-one lessons, of 20–25 minutes each, encouraging them to use strategically variant pronunciations of graphemes to derive the pronunciation of high-frequency exception words. Control children received support in reading aloud without Set-for-Variability. The results showed the children in the Set-for-Variability strategy condition were more likely to apply variant vowels to novel exception words at post-test. While the two groups did not differ post-intervention in any reading accuracy outcomes, Zipke (2016) concludes that the strategic pattern of word problem-solving shown by the Set-for-Variability intervention children demonstrates that the concept of Set-for-Variability is 'teachable'. Notable also is Zipke's report that children found this game-based intervention of problem-solving word pronunciations interesting and engaged enthusiastically with it.

Dyson et al. (2017) also report a brief experimental intervention that trained children in the self-correction of exception words (what they term 'mispronunciation correction') using an Set-for-Variability-style problem-solving approach to derive accurate pronunciations from spelling pronunciations. The sample was 84 typical readers in years 1 and 2 in England, either exposed to an intervention – consisting of eight Set-for-Variability lessons in small groups of up to eight children for 160 minutes in total – or placed in a business-as-usual control. Here the mispronunciation correction lessons involved, first, the assembly of a phoneme string, followed by children deciding if they knew the word derived. If they did not, the children were taught to think of words that sound like the word. Finally, the children were taught to check whether the chosen word made sense in context. Teaching also included instruction in exception word meanings. The results showed a small but significant child-level effect of the intervention on one of two untaught experimental exception word lists (d =.23) compared to the control

condition at post-test, suggesting transfer of children's self-correction instruction to novel exception words. Item-level analyses showed that for untaught exception words, mispronunciation correction predicted transfer, but word definitions did not.

Gonzalez-Frey and Ehri (2021) present data from a US study with kindergarten (Reception-age) children. The results suggest that initially teaching blending by avoiding confusing schwa sounds (such as in 'cuh', 'ah', 'tuh' for *cat*) – by teaching certain GPCs (such as *f*, *m*, *s* or *sh*) ahead of others (such as *t*, *d*, *g* or *k*) and then using a process of continuous phonation (ongoing and seamless articulation of each phoneme) to blend them – is more effective than segmented phonation (distinct phoneme boundaries) in teaching children to learn to use phonics. While not framed by the authors as an Set-for-Variability study, these successful techniques reflect the abstract and complex nature of phonic assembly highlighted by Kearns et al. (2016) and Elbro et al. (2012).

Recently, Colenbrander et al. (2022) compared three methods of teaching exception words among 85 kindergarten children. The children were randomly assigned to either Look and Say (LSay), Look and Spell (LSpell), mispronunciation correction or wait-list control conditions. All intervention group children were taught 12 irregular words in 3 sessions of 10 minutes each. Subsequently, children in the LSpell and mispronunciation correction conditions showed evidence of superior learning of trained words and some improved orthographic-choice task performance. However, no evidence of generalisation to untrained items was found. In evaluating this study, it is worth noting that unlike LSpell, which the authors describe as 'laborious' but effective (Colenbrander et al., 2022, p. 2), Set-for-Variability does not include explicit spelling instruction. Arguably, the correct contrast is either between Set-for-Variability and LSay (where Set-for-Variability is clearly advantaged) or between Set-for-Variability plus spelling and LSpell. In this study Set-for-Variability was taught without using the words in the context of a sentence, which the authors argue was so that the two groups could be 'matched', but which may have further denuded the effects of Set-for-Variability. It is also not clear if the children actually could or did phonically decode items, thus creating spelling pronunciations, or whether this process was simply demonstrated by the trainer – or indeed whether the children then actively self-corrected using Set-for-Variability. It is quite possible this modest instruction does not generalise, because (as we argue) the aim of Set-for-Variability instruction is to teach children to use Set-for-Variability as a strategy, not just to expose children to a word. More sustained instruction is likely needed, as Colenbrander et al. (2022) also conclude.

In sum, the short interventions described here suggest that Set-for-Variability and related processes, such as mispronunciation correction, are teachable and in at least one case (Dyson et al., 2017) impacted performance on a reading transfer task. The literature could thus now benefit from a larger controlled and sustained intervention to show that the teaching of Set-for-Variability along with foundational phonics impacts reading outcomes more generally – and in a sustained manner – among young and poor readers. Finally, before we consider one such study, Kohnen et al. (2020) report a pilot study of a teaching intervention delivered using digital video-conferencing media during the Covid-19 pandemic in Australia. Eighteen struggling readers received 30 multicomponent lessons of 35–45 minutes each, which included text reading, phonics and self-correction techniques for inaccurately pronounced pseudowords (Set-for-Variability). The study had no control condition, but the results showed significantly greater gains in literacy in the intervention phase of the study for the children than in a no-intervention phase, which served as a comparison. The study suggests that interventions including Set-for-Variability, delivered via now widely used video-conferencing platforms, may aid literacy. As the study authors conclude, the promising preliminary work here sets the stage for larger randomised trials of Set-for-Variability.

To date there exist only three published randomised controlled trials (RCTs) or quasi-experimental design studies that have explored extended Set-for-Variability intervention. The first is reported by Savage et al. (2018). The researchers followed a group of 497 Grade 1 children in two Canadian provinces (Quebec and Alberta), selecting 201 at-risk children who all performed below the 30th percentile on a standardised word-reading test. The children were randomised to either a Direct Mapping and Set-for-Variability (Direct Mapping Set-for-Variability) intervention or a common and best practices control. The children then received an average of 11–12 hours of small-group intervention in one of the two conditions, delivered by trained university students.

In the Direct Mapping Set-for-Variability programme, the children received foundational synthetic phonics, including instruction in variable vowel pronunciations. Taught GPCs were closely mapped to real books, chosen for having a high density of the GPCs taught that day. When the children could reliably blend consonant-vowel-consonant words, they were taught Set-for-Variability as a five-step strategy encouraging reflection on word pronunciation problem-solving from spelling pronunciations. The children also played oral language games such as 'Simon says . . . ', where spelling pronunciations were given (for example

'Simon says . . . touch your ar-m or k-n-ee'). Differentiated delivery of this content was undertaken with a sense of fun and playfulness, typical of the approaches used in classroom teaching for this age group. Children in the common and best practices programme received the same content, except that daily instruction on common sight words replaced all Set-for-Variability instruction and that Direct Mapping of taught GPCs and shared reading of real books were eschewed. The results showed statistically significant advantages for the Direct Mapping Set-for-Variability programme on standardised measures of word reading and word spelling at post-test. There were also significant advantages for Direct Mapping Set-for-Variability on standardised word reading and sentence comprehension measures at delayed post-test in the autumn of Grade 2, some five months after the teaching of Direct Mapping Set-for-Variability had finished.

Finally, in one intervention site from the trial by Savage et al. (2018), children in English-language schools in Quebec received some 50 per cent or more of their entire curriculum in French. Côté et al. (2021) took advantage of this fact to explore possible crosslinguistic transfer of Set-for-Variability, reasoning that if it is a generalisable problem-solving strategy, children may also apply it to their French word decoding, even if not instructed to do so. The results showed that children in the Direct Mapping Set-for-Variability condition improved in French pseudoword, regular word, and irregular word reading measures at post-test compared to common and best practice controls. The Direct Mapping Set-for-Variability reading intervention was conducted solely in English, with absolutely no reference to French (not even reference to the idea of cross-language transfer), and solely with taught English GPCs that were not shared in any of the reading tests in French. This result can therefore be cautiously interpreted as suggesting that the Direct Mapping Set-for-Variability intervention facilitates the crosslinguistic transfer of a generative Set-for-Variability reading acquisition strategy.

A very recent study, by Dunn et al. (2024), examined the impacts of two theory-driven multicomponent reading interventions – phonics plus Set-for-Variability versus phonics plus morphology (a focus on shared word meanings) – on struggling readers' reading performance. The participants were 273 struggling readers in grades 2 and 3 (ages seven to nine), recruited in Alberta, Canada. These children were poor readers identified by class-wide screening after phonics-based whole-class teaching in Grade 1. Both groups received a total of 30 hours of small-group intervention over 15 weeks, delivered by university-trained school staff. The results of statistical modeling showed a consistent

significant effect of intervention from pre-test to post-test – and a delayed post-test given some months after the intervention had finished – on all reading outcomes in both intervention groups. Savage at al. (2024) further modelled the data in this sample to show that Set-for-Variability – and not morphology – was a strong predictor of regular and irregular word reading specifically at delayed post-test. This suggests that both interventions had aided the development of Set-for-Variability in many children, an ability that was then used to independently learn to read new words.

Basma et al. (2024) report an RCT of a Direct Mapping and Set-for-Variability reading intervention based on Savage et al. (2018) in 30 children aged nine years with sustained word-reading difficulties. The active control group were taught phonics and sight word reading. An online reading intervention was delivered for 10–12 hours over 10 weeks by trained university students. At post-test the Set-for-Variability group showed a significant advantage in exception word reading and in the oral Set-for-Variability task described by Tunmer et al., suggesting that Set-for-Variability may help children manage print–sound inconsistency in reading acquisition in English.

Together, the results from the studies by Savage et al. (2018; 2024), Basma et al. (2024), Côté et al. (2021) and Dunn et al. (2024), alongside those of some other shorter intervention studies, suggest that Set-for-Variability is teachable and provide promising evidence of the impact of Set-for-Variability instruction on wider reading and spelling outcomes in at-risk and struggling readers. However, while the children in some of these studies showed great improvement in word reading (and measurably improved on sentence comprehension too, in Savage et al. 2018), they remained weak readers on overall sentence comprehension measures after the interventions in the studies by Savage et al. (2018) and Basma at al. (2024). Direct Mapping Set-for-Variability was not an 'inoculation' for children against reading comprehension difficulties. Direct Mapping Set-for-Variability might be usefully paired with an effective language comprehension programme in the future. Secondly, as Petscher et al. (2020) note, this work needs further replication.

The data from the studies by Savage et al. (2018) and Basma et al. (2024) also speak to the impact of Direct Mapping and Set-for-Variability instruction as delivered by trained university students. Such delivery models are atypical in schools, of course. Before scaling up Direct Mapping Set-for-Variability, we need external validity trials of the impact of training regular school staff (teachers and teaching assistants) to deliver the intervention with typical children and, as Dunn et al. (2024)

reported, with at-risk children. Finally, none of this work has yet been undertaken in England, where synthetic phonics is well entrenched in policy. However, as we describe later, when this evidence was shared with the EEF it became the basis for the scaled, teacher-delivered intervention we discuss in this book.

Conclusions

We argue that a child's task in learning to read is to internalise an accurate model of the spelling system so that the child is able to understand and communicate using the printed word. For the complex spelling system of English, current approaches to instruction based on systematic synthetic phonics realise this goal only in part because teaching synthetic phonics provides only some of the generative self-teaching mechanisms needed to complete this task. Wider generative self-teaching tools are also needed to deal with the opaque spelling system of English. Direct Mapping and, most especially, Set-for-Variability are promising tools to this end.

The ideas behind Direct Mapping Set-for-Variability originated in the 1960s, in cognitive and experimental psychology and in linguistics. Theorising here has also drawn from computational cognitive science, hand in hand with rational pedagogical task analysis and corpus analysis, to first inform experimental and longitudinal scientific work across alphabetic languages and, most recently of all, a fledgling intervention science of Set-for-Variability. We also anticipate the development of a neuroscientific line of enquiry, exploring neurological processing measures associated with Set-for-Variability, alongside further eye-tracking studies of online Set-for-Variability processes, following Murray et al. (2022). It is notable here that Basma et al. (2024) took electro-encephalographic measures of brain electrical activity in all intervention children before and after the reading interventions described above. The results showed changes in wave amplitude and latency in one theorised area (the N400 pattern) at post-test, suggesting a neurological correlate of the behavioural changes evident with the Set-for-Variability literacy intervention.

More research is needed, but strong conceptual grounds and at least some empirical grounds for a scaled Set-for-Variability intervention exist. We thus conclude that Direct Mapping and Set-for-Variability are powerful generative tools that may form the basis of increasingly optimal reading interventions in future. In the next chapter, we explore the construction and delivery of one such intervention.

Note

1 On Chapter 12, page 5, and at several points beyond that, Levin and Watson (1963) also use the term ‘set toward variability’ in inverted commas, so nomenclature is inconsistent here.

3
Planning and delivery of Flexible Phonics

This chapter outlines the planning and delivery phases of the Flexible Phonics trial. It is divided into five subsections. The first section focuses on the initial planning and preparation of the trial, starting with the EEF grant application. It considers the complexities of the decision-making process and insights gained from that, which led to the agreement of key outcomes, including the novel piloting of aspects of Flexible Phonics. The second part outlines the processes involved in partnership working with the funding body and independent evaluation teams. Here, we consider the negotiations and drivers of this discussion, which include funding constraints, timing, political context, the remit of the EEF and ethical issues. Insights from the funding process for future research are considered. The need for independent review, government education, consideration of nuance of interventions and other themes are explored.

The third section discusses the design and implementation of the pilot study and how we collaborated with key beneficiaries – the teachers – to shape the design of the intervention so it would be more in line with the education context in England and workable in the classroom setting. We introduce the theory of partnership roles and explain how we used theory-driven mechanisms for co-delivery, including our novel use of ideas drawn from implementation sciences – a theme picked up in later chapters. In the fourth section, we consider the evaluation of the trial manual for teachers and the wider trial project delivery team. In the last part of the chapter, we discuss the implementation opportunities and challenges faced, which included flexible management and delivery, contingency planning and our response to the Covid-19 pandemic, and any project adaptions.

The initial planning and preparation of the Flexible Phonics trial

The history of the Flexible Phonics study starts in 2018, with the launch of the Education Endowment Foundation's (EEF) Early Years Professional Development Round 1 research call in England. This was a public call for research proposals across all domains, with relevant intervention work in the Reception (age four to five) and pre-school phases. A draft proposal for a teacher-led Direct Mapping and Set-for-Variability intervention, delivered alongside a phonics taught control condition not exposed to Set-for-Variability and scaled for a total sample of 100 schools, was duly created, costed and submitted, along with the required evidence of proof of the likely efficacy of the study. The paper by Savage et al. (2018) that showed the effects of sustained Set-for-Variability and Direct Mapping instruction on reading outcomes, as described in Chapter 2, was submitted with the proposal as evidence of efficacy. The proposed trial, however, differed from the study by Savage et al. (2018) in focusing on teacher and teaching assistant professional development to deliver both Direct Mapping and Set-for-Variability content to whole classes of typical children in Reception rather than to the at-risk Grade 1 (aged six to seven) samples in the original Canadian study.

This initial proposal was rejected by the EEF without feedback. It was not until a year later that the lead author (Robert Savage) was contacted, somewhat unexpectedly, by the EEF, who were willing to consider the proposal in a subsequent round of public proposal calls. A phone conversation with EEF programme managers took place in March 2019. It emerged from this discussion that this Early Years professional development call was partly funded by England's Department for Education (DfE), along with EEF funding. The DfE phonics team raised concerns regarding whether the intervention was appropriate for English school settings given its novelty and, related to that, whether it aligned with published government phonics policy in Reception classrooms (DfE, 2014; see Torgerson et al., 2018 for a historical review of policy). Simultaneously (and sitting rather uncomfortably with the first concerns), an additional concern was raised by the DfE about whether the intervention was sufficiently different from business-as-usual phonics teaching, which might already incorporate methods for managing grapheme–phoneme correspondence (GPC) inconsistency in printed words. In short, the DfE had concerns about our intervention being both too similar and too different to existing practice. A range of other specific and relatively technical questions around delivery of the intervention, design and supporting evidence were also raised.

A series of exchanges regarding these issues then ensued with EEF programme managers over several weeks. The concerns about participant sample, raised primarily by the EEF, were relatively quickly assuaged by considering the comparable absolute level of reading ability of the groups of children in Savage et al. (2018) and in the proposed Reception study – both sample groups were at the very outset of word reading. Specifics of the proposed training model were shared, and consensus was achieved on a revised proposal. It was proposed that a business-as-usual control condition would be most appropriate, and that a scaled randomised controlled trial (RCT) be constructed rather commissioning only a small pilot study.

More difficult to resolve was the view, originating from the DfE Phonics Advisory Team and conveyed to us via EEF trial commissioning colleagues, that existing systematic synthetic phonics programmes approved by the government were already managing the noted complexity of written English and that as such they appeared to be at odds with our proposal. In line with the analyses presented in Chapter 2, the EEF proposal outlined that 'English is a deep orthography containing many grapheme-to-phoneme rules, profuse pronunciation patterns (for example, 'ou' in 'mouth', 'shoulder'), and many high frequency exceptions ('pint', 'the', 'one', 'two'). Children may benefit from a *strategy* for managing this complexity' (Savage et al., 2018, p. 225, emphasis added). DfE literacy advisory (phonics team) committee members counselled that England's DfE-approved systematic synthetic phonics programmes do not accept 'exceptions', but rather only that some GPCs occur less frequently in English and are therefore learnt later in the systematic sequence. In addition, it was further claimed by the DfE phonics team (as reported by the EEF programme manager) that certain high-frequency words are initially taught directly, as 'tricky' words (that is, as sight words), and later decoded normally once the relevant GPCs have been reached in the teaching sequence. However, formal computational analyses of the English spelling system show that to fully map it probably requires the assumption of as many as 461 GPCs (Gontijo et al., 2003). Nowhere near that number of GPCs is (or could be) taught in schools. Sight words taught do not reflect the GPCs taught in phonics schemes. Neither the sight words (generally taught as such) nor the GPCs typically taught in English primary schools fully cover the acquisition of the orthographic complexity of English. Furthermore, even if children do know variant GPCs, current practice provides no clue on the selection and management of these variant GPC pronunciations to derive word pronunciations. Furthermore, and key to

our approach, the ubiquitous sight-word instruction techniques used in schools are not generative in the sense described in Chapter 2 – that of empowering wide word learning.

A second assertion by the DfE team was that validated systematic synthetic phonics programmes already teach a two-step process of mispronunciation correction (the synthesis of phoneme strings and then the linkage of the resultant string to an existing pronunciation, as described in Chapter 2). DfE advisory group members felt that validated systematic synthetic phonics programmes fully recognise these two distinct stages of decoding and rigorously teach each of them in sequence. The DfE team were perhaps correct in seeing some evidence of this two-step process in the decoding of regular words (though the claim that it was taught as a distinct sub-skill was probably incorrect) but they did not, however, consider how such an approach could be applied successfully to exception words, as demonstrated in the research discussed in Chapter 2 (for example Dyson et al., 2017; Savage et al., 2018).

A third concern was related to the use of real texts. According to DfE phonics team sources, systematic synthetic phonics programmes should map GPCs to decodable texts, whereas the EEF proposal suggests that real books – authentic, popular children's literature such as *The Gruffalo* (Donaldson, 1999) and *Dear Zoo* (Campbell, 1982) – will be used instead to map GPCs. While real books are a part of practice (used to share with or read to children), DfE phonics team sources felt that encouragement of such texts for early reading practice is at odds with, and undermines, the systematic synthetic phonics approach of current policy in England. Here the DfE team were not, it seems, aware of relevant evidence (discussed in Chapter 2) presented by Jenkins et al. (2004), showing that children's reading development is not damaged by early exposure to real children's books when learning phonics compared with the simpler, decodable texts mandated in many schemes. Indeed, it is here that the rich self-teaching opportunities for learning exception words through Set-for-Variability are evident.

What became clear from these interactions was that demonstrating flexibility and framing the proposal to respond to the concerns of the DfE was very useful in cementing funding. Finally, a phone call with the DfE phonics team in late March 2019 was illuminating. This provided the researchers with an opportunity to give concrete case examples of how Set-for-Variability can be used to resolve inconsistencies between phoneme strings and word pronunciations in English. That same morning, the child of an acquaintance of a senior policy advisor had struggled to spell the word *kitten*, which is spelt as if pronounced 'kit'-'ten'. This single

case (the effects of a diphthong) seemed to trigger an understanding of the broader Set-for-Variability approach among the DfE phonics team for the first time. Throughout this period, responses to the DfE phonics team sought to clarify that Direct Mapping and Set-for-Variability are methods to optimise and extend, not supplant, systematic synthetic phonics.

As noted above, the need for alignment with systematic synthetic phonics was strongly emphasised throughout this process by the DfE and a strong focus on delivery team flexibility in order to satisfy DfE demands was implicit throughout the research commissioning phase. It should be noted that such an approach to interventions is quite inconsistent with the approach of scientific innovation we champion here. In the latter, existing practices are improved upon with new scientific insights to develop and evaluate better teacher-delivered interventions for the benefit of young people and, where necessary, demonstrably superior processes supplant existing practices. Better practices and public services emerge from such pragmatic scientific approaches (Haynes et al., 2012). It seems that greater awareness among governmental phonics advisors of this broad scientific principle, as well as of the specifics of the reading process, is warranted.

As it transpired, some of the potential for failure to align was reduced in light of a favourable interview with EEF trustees, chaired by the then chief executive, Sir Kevan Collins, on 10 April at Millbank Tower. Here, the EEF made an in-principle decision to sole fund the project, without DfE funds. The EEF also eschewed a further DfE request that any testing of Flexible Phonics be in the context of and measured only against what is formally recognised to be the best of current effective systematic synthetic phonics teaching (schools externally assessed by Ofsted as 'outstanding' on systematic synthetic phonics as opposed to schools with a more representative range of assessed systematic synthetic phonics practices).

Eventually, the EEF fully funded the agreed proposal, with at least 50 intervention schools in London and at least 50 business-as-usual control schools, also in London, with a representative range of systematic synthetic phonics practices and Ofsted reports. In the agreed proposal, training of teachers and teaching assistants was to consist of an initial day of training, followed by one to two weeks where teachers and teaching assistants trial the strategies, before returning for another day of training. Following this second training day, teachers and teaching assistants would return to school and continue to use the strategies as part of their everyday phonics teaching, both to the whole class and in small-group sessions. Following the second day of training, Professor

Savage and his team planned to visit each school twice in two weeks to support delivery and answer questions, alongside ongoing online and telephone support. A request for budget to allow for the assessment of treatment integrity across all 50 intervention schools was not approved, though a more specific qualitative process evaluation to be undertaken by the evaluation team in eight or so schools was approved, in principle. Importantly, a budget was approved for a team of support partners to support schools with embedding the intervention, by providing mentoring and coaching to teaching staff as and when required and through planned classroom visits. Our prior research has strongly suggested that such sustained support is both essential and effective in enabling teachers and teaching assistants to take on the new practices they have encountered in professional development training (Savage et al., 2013). We have also used these same individualised approaches to sustain post-professional-development school support in a successful EEF multi-site scale-up trial of another reading intervention called ABRACADABRA in England (Bell et al., 2022).

A novel requirement that was added at this stage by the EEF was that the Flexible Phonics approach to be used in the main trial be preceded by a formal pilot study. As far as we are aware, pilot studies have rarely been used in EEF trials. The pilot, intended in part as a way of overcoming ongoing DfE concerns about the project, also provided an opportunity to test the approach's claims to novelty in the English school context directly with teachers. An initial requirement that the pilot demonstrably improve reading outcomes was subsequently relaxed, as rigid timelines for recruiting over a hundred main trial schools precluded it. It was agreed that the project was to be piloted in spring 2020 in around five to 10 London schools. The aim was to adapt the content to align with best practice phonics teaching in England and pilot the training model before running a one-year trial in the academic year of 2020–1. Ongoing development work on the content of the intervention was envisaged to take place from October 2019 to June 2020. Within that time, a pilot intervention lasting around three months was envisaged, with a short evaluation report that would detail how the pilot was delivered, how schools have responded and what changes might be made for the trial. A more detailed plan for development work and piloting was thus constructed in April 2019. With the expert support of an EEF programme manager through these early phases, the budgeted project (June 2019) preceded formal project sign-off and approval in September 2019.

The roles of the delivery and independent evaluation teams

This second part of the chapter outlines the processes involved in partnership working with EEF funding partners and the independent evaluation teams. One of the methodological strengths of the EEF approach to RCTs is that for any given funded trial, there is a delivery team charged with all aspects of preparation, delivery and support for the intervention content in schools. This team is wholly separate from the independent evaluation team, which is charged with providing expertise in impact evaluation, implementation and process evaluation, and the formal analysis report to the EEF. In this trial, the Institute for Employment Studies (IES) in Brighton, England, was appointed by a panel directly as the result of a competitive evaluation bid by the EEF to their panel of approved evaluators. As part of best practice for impact evaluation (Moher et al., 2001; Schulz et al., 2010), evaluators complete a protocol describing the whole evaluation and a statistical analysis plan before data collection, describing in detail how impact evaluation statistical analysis is to be conducted and reported. This pre-specification prevents a host of statistical and interpretative problems, such as post hoc 'fishing trips' and '*p*-hacking' techniques, which can both lead to repeated unprincipled significance testing that undermine rigorous hypothesis testing. (See Chapter 6 for more detail.)

While the delivery team and the independent evaluation team are typically quite separate in role and action, there is always a limited degree of EEF-managed interaction between them. We will see in Chapter 6, for example, that evaluators observe professional development training and interview delivery team members as part of the formal process of evaluation at the early stages of the research process (also see Box 3.1).

Box 3.1: The centrality of partnership between delivery and evaluation teams

One of the key features of the work reported on in this book is that, at every stage, working more closely than is typical with the evaluation team was fundamental to the successful completion of the trial, particularly in the context of Covid-19 and its effects on schools. This partnership working was particularly important when it came to collecting data to ensure stronger outcomes in terms of minimising attrition (loss of schools and children out of the trial, a potential threat to the validity of RCTs), keeping schools informed about the evaluation process, supporting evaluators with relevant aspects of data collection and keeping schools on

board. We will have more to say on this in later chapters. We note also, however, that in this judiciously delivered collaboration the fundamental independence of the evaluation team was still maintained.

In all EEF trials, after an initial set-up meeting to introduce the two teams to each other and to outline key aspects of the evaluation and delivery, the first formal meeting is the first intervention delivery evaluation analysis (IDEA) workshop meeting. In the IDEA workshops, teams from the IES and UCL set out to explore the intervention, with the aim of best evaluating it. The first IDEA session took place shortly after set-up (October 2019) and another followed after pre-trial development (in June/July 2020) to finalise relevant documents and decisions. Atypically for EEF practices, but reflecting the rapidly changing pandemic context, these decisions were revisited in April and June, with input from the assessment partner (Qa assessment) in June. Building on these meetings, the IES and UCL partners worked to co-develop several important protocol documents that surround this trial (and, generally, all recent EEF trials).

The first document was a tool called the Template for Intervention Description and Replication (TIDieR) framework (Hoffman et al., 2014). It is a checklist and guideline developed to help to improve completeness in the reporting of interventions in research studies, originally developed in the context of medical trials. It describes 12 features of any given intervention and is described in Box 3.2.

Box 3.2: The TIDieR framework for interventions

1. Name (of intervention)
2. Why (underlying theory of intervention)
3. What (materials used)
4. Who (intervention provider)
5. How (delivery mode)
6. Where (infrastructure used)
7. When
8. How much (duration, intensity)
9. Tailoring (personalisation or adaptation)
10. Modifications (changes during the study)
11. How well planned (intervention fidelity assessed)
12. How well actual (intervention fidelity assessed as planned)

Details of the TIDieR checklist analysis for Flexible Phonics are presented below. This is the final TIDieR framework document agreed between the delivery and evaluation teams. Following EEF advice, and again reflecting the unique pandemic context, many more major modifications were undertaken in this document than is usual for an RCT. The most notable changes – to the delivery of teacher professional development and support – were necessitated by school closures at that time. The full specification of pandemic-induced changes to protocol are outlined below. We did not include a 'How well (modifications were carried out)' section here, as this is considered in detail in later chapters given the additional complexities of running the trial during the Covid-19 pandemic. Some of these changes, while reflecting exigencies at the time, may be helpful to RCT trials run beyond the pandemic context. We also added in the treatment of the control group, as it was considered particularly important in an RCT.

1. Name (of intervention): Flexible Phonics

2. Why (underlying theory of intervention): There is now a lot of evidence to support systemic phonics, but there is still value in exploring whether it can be made more effective in supporting children as they learn to read. Recent evidence suggests that combining Direct Mapping and Set-for-Variability strategies can help to do this.

3. What (materials used): Teachers and teching assistants who were allocated to the intervention condition received three half-days of professional development training. During the Covid-19 pandemic, remote training using video-conferencing software, such as Zoom, was used in this version of the intervention. Intervention participants also received a hard copy of a teacher manual and access to the UCL Extend online platform. The platform contained a discussion forum and videos of training activities, audio files for teaching activities, an online version of the training manual for professional development, frequently asked questions, training slides, Mentimeter feedback responses from the training sessions, and teaching materials developed and shared by other schools in the trial. Following the training, there were three follow-up visits with research assistants (known as support partners). These were also delivered using video-conferencing software during the 2020–1 academic year, but would normally have been in-person so that support partners could observe the classroom context and provide further feedback and guidance around delivering the intervention. Participating schools also received free children's books to the value of £400 per school,

which could be used to implement the strategies. The delivery team selected books from existing commercially available children's literature that they felt were of high quality – that is, age appropriate, well written, with engaging stories and appealing illustrations. This included popular classics such as *The Gruffalo*. Ongoing telephone and email support was provided as requested by schools.

4. Who (recipients): All pupils in Reception year (age four to five), but there may be added benefits for low-achieving pupils. All Reception class teachers and teaching assistants are direct recipients of the training and then deliver the intervention to their Reception class pupils in lesson time.

5. What (procedures): The training introduced two strategies for the teachers and teaching assistants to implement in their teaching.

The first strategy, Direct Mapping, requires children to read texts that include several examples of the GPCs that they have just learnt. In the first instance, these will be carefully selected decodable texts or specifically crafted controlled texts. After this, real books will be introduced slowly and strategically. While many models of phonics teaching link phonics and texts, Direct Mapping aims to do so more thoroughly and consistently, and on the same day as children learn the specific GPCs, aiming to ensure that children understand phonics in context.

The second strategy, Set-for-Variability, explicitly teaches pupils to add in another step after they have blended phonemes to graphemes – using Set-for-Variability. This is a metacognitive step where pupils recognise that they have not been able to successfully identify a word by blending phonemes and that they will need to use alternative strategies to identify the word. In Set-for-Variability, pupils consider what the word may be by thinking about the distance between these blended sounds and known words and about potential spelling-to-sound inconsistencies. For example, when they sound out the phonemes 'c'-'a'-'t', the sounds they make bear little resemblance to the actual word *cat*. Set-for-Variability encourages pupils to take a moment to consider what the word may be, based on the words that they know. This enables children to better recognise all words but can also be especially useful when learning to recognise exception words (for example *wasp*). In comparison with other phonics programmes, Set-for-Variability makes this metacognitive step following the blending of phonemes much more explicit, which can enable children to be more flexible when selecting strategies to decode difficult words.

The three online support appointments for Reception teachers and teaching assistants ran in February–March, March–April and April–June

2021 for this project, and would be expected to run in a similar way in future. In this approach, the first appointment is offered to each class in a school (so four appointments for a four-form entry school) and staff are offered group appointments if preferred (which some schools requested in this project). The appointments are approximately 30 minutes long and sessions are held at twilight between 3.30 and 6pm. In the sessions, staff can ask questions and get advice on best-practice implementation of the programme, and the Flexible Phonics support partners can deal with any misconceptions about the programme and provide further clarification. For example, some schools needed reassurance that Flexible Phonics was compatible with government requirements that they use a systematic synthetic phonics programme. It was further clarified that Flexible Phonics was meant to be used alongside their existing phonics programme and was not intended as a replacement for their main programme.

The online platform UCL Extend was also available for the duration of the intervention, with a variety of resources, as noted above. The platform also included a discussion board for all trained teachers and teaching assistants, through which they could ask for additional support as needed. Best practice and resources provided by partner schools were shared on schools' behalf by the Flexible Phonics team through this medium, or schools could upload and share resources directly.

Teachers and teaching assistants could also choose to share videos of their own practice through video calls with UCL staff for specific further feedback. A monthly email bulletin provided updates from UCL Extend, including resources shared by other schools, and highlighted any relevant articles on topics of concern for schools (identified in the support appointments and training), such as working with children with English as an additional language (EAL) or children with special educational needs or disability (SEND). The bulletin also shared answers to frequently asked questions raised during the training and in online appointments more widely.

Between online support appointments, the Flexible Phonics support team provided proactive support for schools via email. Schools could also contact their allocated Flexible Phonics support team by phone or email as needed.

6. Who (provider): Professor Savage and his team at UCL's Institute of Education[1] delivered the training and follow-up sessions to the teachers and teaching assistants, who delivered the strategies within their normal phonics practice (in both whole-class and small-group delivery) after the children had learnt GPCs. A phoneme is the smallest spoken unit of sound

(for example the word *rain* has three phonemes; 'r'-'ai'-'n'). A grapheme is the written symbol that represents that sound, which can be a single letter or a sequence of letters.

7. How (format): The strategies were delivered in normal phonics lessons.

8. Where (location): The schools in this project were recruited from Greater London.

9. When and how much: The original intention was for the intervention to be delivered over five months, from January 2021 to the end of May 2021. However, in this project, delivery was approximately three and a half months due to the Covid-19 pandemic. The expectation of the intervention is that, after the training, the strategies will be incorporated in all phonics lessons until the end of the school year, which is normally three to four times a week depending on the school.

10. Adaptation: Teachers tailored and differentiated the content to suit the children in their classes. There was freedom for teachers to adapt and modify as they go (although there was a defined core that they were required to follow).

11. Control condition: The control condition was business-as-usual phonics practice. Schools allocated to the control condition received £1,000 at the end of the academic year when post-testing was complete.

12. Modifications: Training and support visits were delivered online instead of in-person due to the pandemic, and the intervention was reduced from five months to three and a half months due to lockdown-related school restrictions. There were also some changes to the evaluation due to the pandemic, including remote (instead of in-person) case studies and some remote child testing. Theory of Change workshops were moved online and to later dates, and prior school phonics scores were not collected, due to the concern about placing an extra burden on schools.

A second key element discussed in IDEA meetings was the Theory of Change model. This is a formal logic model of the features of the intervention used and it specifies expected paths to outcomes. The generic template of a Theory of Change we started the IDEA meetings with appears in Figure 3.1. A Theory of Change can be understood to be 'a visual representation of a programme's inputs, activities, outputs, outcomes, and underlying causal mechanisms' (EEF, 2019). The Theory of Change model is helpful in all eventualities, but particularly when adjudicating the causes of non-significant interventions such as theory

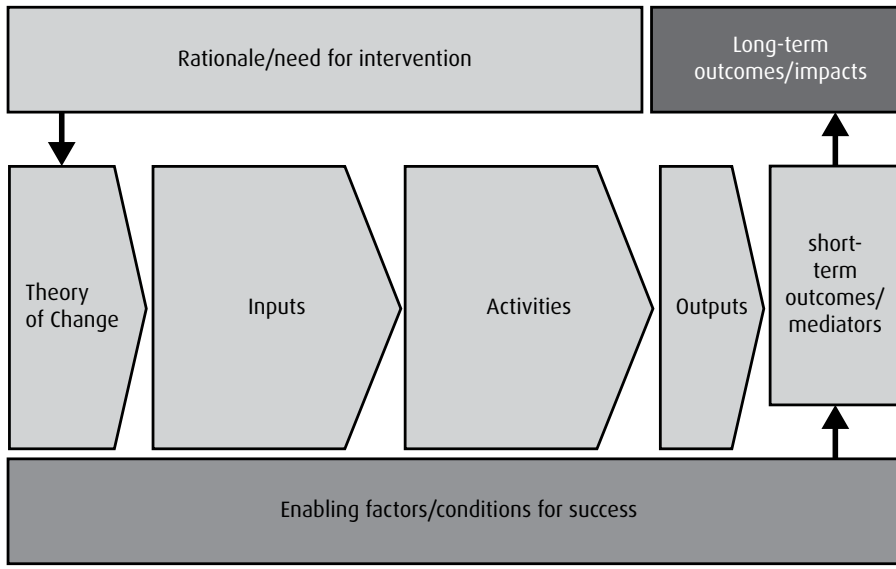


Figure 3.1: The generic template of a Theory of Change model used for the 2020 Flexible Phonics trial in England. © Authors.

failure (the intervention does not work as predicted), implementation failure (it was not implemented as intended) or methodology failure (inadequate evaluation methods were selected or suitable evaluation methods were used inadequately) (EEF, 2019). The Theory of Change model went through several iterations in IDEA discussions after the pilot and also at two points during the trial, given pandemic-related adaptions. The final, co-developed version appears in the pre-registered trial evaluation protocol report.[2] It describes the theory-driven levers of change (Direct Mapping and Set-for-Variability) as putative causal mechanisms and explores the potentially mediating effect of oral Set-for-Variability on outcomes. As such it reflects the theory-driven logic of specifics of a 'realist RCT' (Bonell et al., 2012), as described in Chapter 1.

Pilot design and implementation and the expert role of teachers

A relatively distinct feature of the Flexible Phonics intervention was the inclusion of a pilot study. Construed initially in part as a protection for the EEF and the DfE, as described above, this trial provided a rich learning opportunity for us prior to the main study. Furthermore, as

noted in Chapter 1, it has been argued that the results of EEF RCTs are often quite modest in size (Lortie-Forgues & Inglis, 2019). One reason for this may be that insufficient grounded rich picture work in authentic contexts (schools, regular teachers) is undertaken prior to trials to inform content and professional development delivery. Piloting may thus be key to improving RCTs (Dawson et al., 2018).

In overview, the qualitative pilot evaluation involved 13 Reception teachers and teaching assistants, working in primary schools serving deprived areas of London, who contributed to focus groups and a survey. These both explored perceptions of the quality, novelty and social validity of the content and delivery of the draft Flexible Phonics professional development training, delivered over one and a half days. Content analysis of the results of focus groups, and the quantitative and qualitative survey analysis, both carried out seven days after initial training, suggested broad acceptance of the value and usefulness of the intervention content. One teacher commented, for example, that through Flexible Phonics:

> [children] are not just approaching it [reading teaching] with one strategy; it's that they can approach it from various ways, and that they know that the language they are learning is complex, but they've got various ways, and it's like I'm not going to be put off because it's like, if that doesn't work then I can try something else. And also know that phonics is rooted in real life, it could be real texts, it could be conversations.

The results showed the social validity of the Flexible Phonics approach and also attested to its current novelty within the context of existing literacy policy and practice in England for Reception classes.

Teachers also broadly supported the model of co-expert intervention delivery that we presented and used in professional development. Here we saw teachers and teaching assistants as equal partners, coming together in discussions to consider effective early literacy. Teachers are experts when it comes to their pupils, their classrooms and the wider context, as well as in differentiation and creative implementation, while the delivery team brought expertise in models of reading and the particulars of evidenced intervention strategies. In our professional development approach we also focused less on teachers complying with specific approved content than has been the case in some professional development; instead, the focus was more on understanding the principles of Set-for-Variability and Direct Mapping so that educators

could apply them with clarity in Reception classrooms in ways relevant to their classes. Notably, we explained – referring to relevant research and exemplification – *why* these strategies might be useful for learning to read in English, emphasising principles of efficiency and generative learning for independence (Boldrini et al., 2023). Following the work outlined in Chapter 1, we expected the differentiation of Flexible Phonics and principled variation in the use of Flexible Phonics content in different school contexts. One teacher noted:

> The five areas kind of overlap slightly and that it's about being flexible within those approaches, and that the children knowing that they can be flexible, and I actually like the idea of them finding something that works with them, so they might actually start to merge their own ones, like the things they've learnt as a way to be successful. . . . and for the children they can be flexible with how they are going to learn to read, and that will be different for all of them, and I think that that's really good.

Pilot data also showed clearly that teachers were willing to modify and potentially improve on their existing reading practices, even in the mid-year, within this form of co-expert approach (for further details, see Savage and Fox, n.d.). We will also pick up the specifics of teacher and teaching assistant responses to our co-expert model of working in Chapter 5.

As a result of the educator feedback in the pilot work, we created videos of teachers demonstrating activities from the Flexible Phonics programme so that teachers could understand how the training fits into existing classroom practice. We also created videos of the training so all staff could access it. We attempted to reduce the technical vocabulary in the training and the manual where possible and we included a glossary of key vocabulary in the manual. We included more practical activities in the training, allowing for role play, and we explored resources such as the key texts in training. We added more time for reflection on learning and for team planning of lesson delivery after each set of weekly lessons. Finally, we improved the layout of the manual to make it more accessible and we improved the resources (for example the child strategy sheet resource) in line with the suggestions received.

Additional activities of the project delivery team

This section of the chapter considers the wider work of the broader project delivery team. Before exploring this, an important element of support for and validation of Flexible Phonics was expert involvement in the refining of the draft teachers' manual. Two British university academics, with expertise both in intervention and in Set-for-Variability, provided feedback on the draft Flexible Phonics manual. Both were supportive of the content and provided helpful, detailed and broad comments. Indirectly, they also provided some expert-based validation of the concept of Flexible Phonics for the EEF. It was initially hoped that highly experienced English primary teachers might also provide input on the manual, so that we could have practical perspectives alongside the academic perspectives. The novelty of the Direct Mapping and Set-for-Variability approach in the English context meant that it was quite hard to find expertise in these specific elements. In May 2020, the EEF placed an advertisement in their newsletter requesting feedback from teachers about the manual; however, they received no expressions of interest. It is quite possible that the pandemic context, with many in education being highly stretched at this time, affected the consultation process. In the end, the EEF programme manager, a former Early Years teacher, provided detailed feedback about the manual's content.

In January 2021, we employed a small team of support partners to support intervention schools with implementing and embedding the Flexible Phonics intervention. From both research and practice backgrounds, all team members had recent experience of phonics teaching; most had worked in primary teaching roles and so had a good understanding of the practicalities of classroom teaching. Each school was allocated a dedicated support partner, who provided tailored support via email and telephone and through a series of pre-arranged online support appointments. Queries, misconceptions and concerns about the programme were dealt with during the online appointments and any follow-up actions were resolved swiftly to help maintain schools' momentum with the intervention. Between appointments, the partners proactively contacted their schools to share any good practice and resources from other schools to support intervention implementation and delivery.

A record of the content of each online support appointment was made for later analysis, including the queries raised by schools and the advice given by partners. As a quality check, these records were reviewed on a regular basis by the project manager and academic lead to ensure that

the support and advice provided to schools was consistent and of good quality. Common queries and school challenges were shared amongst the team to ensure consistency in support (see Chapter 5 for more details).

Implementation opportunities and challenges

This final section explores the (of necessity) adaptable management and delivery of Flexible Phonics, contingency planning and our response to the pandemic and any project adaptions. Large RCTs are complex to run at the best of times. Many additional challenges were faced due to Covid-19 and the extended school closures in 2020 and 2021 (for all except vulnerable children and children of key workers, who still received in-school education).

Most school-based evaluation trials commissioned but not yet delivered were paused by the EEF in 2020. On the EEF's advice, contact with schools was paused from March to May 2020. Fortunately, school recruitment targets were met early, with 122 schools signing up for the trial before the first lockdown in March 2020. Given the increasing pressures on teacher workload brought about by the pandemic, concerns were raised by the EEF and the evaluator that many schools would drop out of the trial if it were to continue in its current form. Specific concerns were raised about conducting baseline assessments face to face, as some schools had put in place a 'no visitor' policy to reduce infection. Other concerns were around whether the training should be conducted face to face or online. At that time, many London venues were not taking bookings given the safety issues of managing large groups of people. This pertinent issue was raised by all teams.

A survey was sent to schools who had signed up to collect further information about the feasibility of the trial from teachers' perspectives and to make a final decision about whether to delay the trial or not. A total of 98 schools out of 122 responded, with 94 per cent indicating they were happy to continue to take part in the trial with baseline assessments being conducted face to face, as originally planned. These assessments were to take place in October to December 2020 and would be carried out by the evaluation team's partner, Qa Research. Schools were also asked about their preferred method of receiving the intervention training, with 63 per cent being in favour of online training. In light of this information, and the above concerns, it was deemed acceptable for the trial to continue, but with online adaptions, intervention training and follow-up support to be delivered remotely. More details about these adaptions can be found in Chapter 6.

One potentially positive change that we were able to bring to bear in this trial was to increase sample size. The evaluators tested more schools than originally planned, as budget was no longer needed to cover the costs of face-to-face training days at a London venue. We thus reopened school recruitment in July and ran it until September 2020. During this time we recruited additional schools, with a total of 123 schools entering the study pre-test, to mitigate against potential attrition. We suspected that attrition might become higher than usual due to the impacts of the pandemic on school functioning. Ensuring that as many of these schools stayed in the trial as possible was crucial to keeping power high for the analysis of intervention effects.

The delivery team also strongly emphasised to the EEF the need to support young learners through the anticipated pandemic learning loss and the consequent need for highly efficient, evidence-based, language-rich interventions to support young learners and teachers. Emerging evidence of the impact of online reading interventions (Furlong et al., 2021), and some remote interventions that involved Set-for-Variability content (Kohnen et al., 2020), was considered in discussions. Communications from the EEF at various points during this time indicated that there were no guarantees the trial would go ahead. While constantly at risk of closure by the EEF, particularly after the second UK lockdown (January to March 2021), the Flexible Phonics trial was kept open with the adaptations mentioned above.

To respond to any concerns and questions schools might have about the trial, such as safety when conducting baseline and follow-up assessments, training and delivery of the intervention, the delivery and evaluation teams ran a series of webinars in July 2020. Over 95 per cent of schools who signed up for the trial attended. These webinars gave schools a forum to put forward questions anonymously to the teams, using audience participation software. The majority of questions raised during the webinars related to practical issues of how the intervention might fit with their current phonics practice and existing programmes, the impact on their daily timetable, how schools might be selected to take part in the intervention (schools were randomised in December 2020) and how the reading assessments would be conducted (detailed in Chapter 6). A detailed information pack was shared with schools, outlining the project timeline and evaluation processes, to ensure schools were fully informed about the trial and the adaptations that had to be made related to the Covid-19 pandemic.

Other major changes in light of the pandemic included the abandonment of the proposed wider classroom quality evaluation by the

support partners. These were intended to be face-to-face observations in classrooms which were carrying out the intervention. Instead, details were collected during pre-arranged online support calls to determine the extent to which schools had implemented the intervention. All communication between the delivery team and intervention schools was conducted remotely via video-conferencing calls and emails, as opposed to face-to-face visits and support. All of these elements undoubtedly impacted the quality of the trial – compared to traditional delivery and evaluation methods – but this was deemed preferable to trial abandonment.

Conclusions

This chapter explored the set-up and delivery of the Flexible Phonics intervention. Partnership working with EEF and, most particularly, the English government in setting up the study was in many senses the art of realpolitik as much as the pursuit of a science of reading. Insights from the early funding process suggest the importance of expert independent review of proposals beyond government advisors and the need for educating the government (DfE official and advisory team) on how improvement of public services is undertaken through novel RCTs. The basic scientific approach – theorising evidencing and then undertaking a genuinely new approach to teaching that differs from existing policy, which is then assessed in a carefully controlled way to improve public policy – seemed not to be clearly grasped. Partnership working with the EEF and independent evaluation teams showed that there were many funding and timing constraints, in part reflecting the EEF's role and evaluator remit, but sensible compromise was largely possible.

There are several potential implications of the pilot study approach used here for subsequent EEF-funded RCTs. Most broadly, such situated pre-RCT approaches aided the RCT and suggest the importance of social validity of co-expert interventions in intervention uptake. The modifications needed in light of the Covid-19 pandemic were many. One key adaption was to an online model of training and associated follow-up support. We consider some of these issues further in the next chapter.

Notes

1 Professor Savage moved to York University in Canada in summer 2021, after delivery was completed.

2 The evaluation protocol for the Flexible Phonics trial is available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/pages/projects/EEF_trial_protocol_flexible_phonics_updated_final.pdf?v=1701427207

4

Intervention training and support for schools

This chapter focuses on the training and support model developed to help practitioners deliver the Flexible Phonics intervention in schools. We first outline the Flexible Phonics intervention training delivered to practitioners during early 2021. We then explore the model and the content of the Flexible Phonics intervention and the associated delivery challenges and adaptations given the context of the Covid-19 pandemic. We then detail the model of support for schools, explaining how the delivery team worked with schools to support practitioners to implement Flexible Phonics alongside their existing phonics practice. The team's experiences of supporting schools to implement Flexible Phonics and schools' perceived challenges and barriers are highlighted. Finally, we outline the tools that were developed to assess implementation fidelity and we explore the extent to which the intervention was implemented and delivered in schools as intended.

The Flexible Phonics training and community of practice

The Flexible Phonics intervention was designed to dovetail with existing phonics teaching, as a complementary approach rather than a standalone phonics scheme or programme. It can be best conceived of as a set of tools or strategies for optimising phonics teaching efficiency that teachers can share with children so they can utilise them to learn to read. The educator intervention training was focused on two core concepts considered in previous chapters: Direct Mapping and Set-for-Variability, or mental flexibility (oral flexibility in mapping phoneme strings to word pronunciations and mispronunciation correction – the ability to correct mispronunciations to read exception words). As described later, teachers

were afforded some freedom to differentiate the content of Flexible Phonics and adapt it to suit the needs of their pupils, provided the core elements – Direct Mapping and Set-for-Variability teaching – were included. The scientific background to the Flexible Phonics intervention and the concepts of Direct Mapping and Set-for-Variability has been explained in detail in Chapter 2. Later in this chapter, we outline exactly how these concepts were translated into practice in the Flexible Phonics intervention.

Knowledge sharing through partnership working between the schools in the intervention and the delivery team was a fundamental component of the Flexible Phonics approach and was strongly emphasised during training. To this end, we considered teachers to be the experts in their classrooms, knowing what will work best with their classes, and throughout the professional development sessions and subsequent support sessions practitioners were encouraged to be creative and to draw from their professional experiences. An element of the Flexible Phonics approach was also co-constructed at the start of the intervention and tailored to meet school needs and differences. For example, if a school felt their children were already secure in a part of the content, such as oral vocabulary, or if there were no less able readers who could not blend, then following our agreement, these elements of Flexible Phonics were not taught in those schools.

We initially envisaged (and shared) a model of communities of practice and content sharing across teachers. Given the number of schools involved in the intervention arm of the trial (61 out of 123 schools), the delivery team felt there would be much value in practitioners contributing their experiences and practices while problem-solving and generating ideas to build a repository of knowledge and shared expertise. In this way, a community of practice – a collective of people who share their interests and problems around a specific topic – was operating. Through regular interaction, members of a community of practice learn together and gain a greater degree of knowledge and expertise about a particular subject, allowing for further professional development (Wenger, 1998; Wenger et al., 2002). This often happens face to face and informally in a school setting, such as in the staff room, where teachers might collectively discuss specific teaching situations or strategies during lunch. A community of practice can also occur successfully online, through discussion boards and on social media, where educators can discuss a particular topic to seek advice and guidance and share resources. This has become commonplace as a way for less experienced teachers to connect with more experienced peers and to engage in professional conversations as an informal type of continuous professional development (Rosell-Aguilar, 2018).

Originally, the model for the Flexible Phonics training was to deliver it face to face over one and a half days in the early part of 2021, using all insights from the pilot study, which had been delivered face to face. However, during this period the UK entered its third national lockdown in response to the Covid-19 pandemic, with most children taught remotely through online lessons and only a handful of vulnerable children and children of key workers being taught in each school. Therefore, the professional training sessions were adapted to run as online sessions delivered via video-conferencing software, to accommodate both teachers working remotely and those in face-to-face school settings.

Teaching staff were invited to join one of seven training cohorts, with each cohort consisting of three synchronous half-day online sessions run over two weeks. Multiple training dates were provided over two weeks, so schools could stagger staff cover and register for their preferred cohort. Around 40 staff attended per cohort, with just under three hundred Reception (age four to five) teachers and teaching assistants from 58 (out of 61) participating schools attending all three training sessions. The other three schools only partly attended the training sessions due to difficulties covering staff delivering online and face-to-face teaching. These schools viewed recorded sessions of the training and received additional support through an individualised online tutorial with the academic lead. Alongside the online training, all practitioners received a hard copy of the training manual, which contained the core content of the intervention covered in the training sessions along with example lesson plans and activities. As part of the community of practice, practitioners also had access to an online portal containing additional resources developed by schools during the intervention delivery phase. All intervention schools received books worth £400, carefully chosen to represent the taught graphemes in Flexible Phonics and thereby allowing the Direct Mapping process to be taught.

The first and second half-day online training sessions for each cohort were run in the same week and the third session was held two weeks later. The second and third sessions were spaced apart to give practitioners time to consolidate learning from the training and to practise carrying out the intervention in their schools, either through online teaching or with any children being taught face to face in school. Feedback from the pilot study (Savage & Fox, n.d.) indicated that teachers wanted the opportunity to try out the intervention between training sessions so they could familiarise themselves with the intervention and ask questions at the final training session. One challenge, as with any online professional training, was to keep the

sessions engaging and to encourage active participation. Accordingly, the training was designed to involve multiple interactive elements, including breakout rooms and question-and-answer sessions through the videoconferencing chat function. Anonymous polls were created using the collaborative audience participation software Mentimeter (https://www.mentimeter.com/). These contained open questions assessing practitioners' understanding of the intervention and reflection activities. Using this software, the delivery team could gather immediate feedback to gauge understanding of the materials taught and pinpoint where further clarification might be needed. The polls also gathered anonymous feedback, including from those who might have been reluctant to offer it in a face-to-face training situation.

All training sessions were led by Professor Robert Savage. In the first two sessions, the teaching staff received an introduction to the theory underlying the Flexible Phonics approach and learnt how to deliver the first four aspects of the intervention (vocabulary, grapheme–phoneme learning through Direct Mapping, oral Set-for-Variability and printed-word mispronunciation). Authentic video recording of Flexible Phonics activities in schools was impossible at this time. The training sessions included short mock-up videos of activities from each of the core aspects of the intervention, designed to enable practitioners to better understand how to implement activities in their settings. The videos were discussed in small groups in virtual breakout rooms; through these peer-to-peer collaborative activities, attendees reflected on how these activities might be implemented in their classrooms while sharing ideas about delivery and evidenced good practices.

In the third and final half-day session, teaching staff received training related to the fifth aspect of Flexible Phonics – activities to aid phonic blending aimed at children yet to achieve this prerequisite ability for Set-for-Variability. Small breakout groups shared their experiences of either integrating Flexible Phonics into their literacy planning or delivering an aspect of Flexible Phonics as a face-to-face activity with key worker and vulnerable children or through online teaching. Outcomes of the breakout room discussions were shared with the training group. Here planning and delivery challenges were solved as a group, and any teaching ideas or good practices were shared. Such teacher-led discussions can be a particularly valuable way of collective learning amongst practitioners who might be experiencing similar challenges and difficulties when implementing and delivering the intervention. These practices were informed by observer and participant feedback from a detailed pilot training study, discussed in Chapter 3 (Savage & Fox,

n.d.), which indicated that they were a valuable way of learning. School support partners attended these sessions and facilitated activities in the breakout rooms.

To facilitate a community of practice for delivering Flexible Phonics, knowledge sharing was supported virtually through an online platform where practitioners were able to upload example lesson plans, ideas and classroom activities and post questions and share experiences through discussion forums. The online platform also included a repository of training resources, such as videos of the online training sessions and mock-up activities. Active engagement from practitioners in this virtual community of practice was limited, however. A few practitioners posted questions in the online forums during the early stages of delivery (the few weeks following the training), but this subsided in the subsequent weeks.

To facilitate and encourage interaction throughout the intervention phase (January–July 2021), the delivery team proactively shared good practice in the online forum on behalf of some schools. Moderators posted questions, encouraged open dialogue and shared responses to common queries, such as how to adapt the intervention for children with special educational needs and disabilities (SEND) and those with English as an additional language. Despite best efforts to encourage regular and active engagement, active use of the forum by practitioners was limited (two posts were made in total).

Various factors may have contributed to this lack of engagement. Remoteness may have made it difficult to create a sense of community and foster meaningful connections between members. Given the challenges of the pandemic, teachers struggled with a lack of time and competing work priorities. Teaching staff may have addressed queries with their dedicated support partner (a member of the delivery team) during their prearranged support meeting rather than sharing these online with the wider group. This can be a common occurrence in online platforms, with those who never or rarely share knowledge accounting for 50–90 per cent of participants (Marett & Joshi, 2009; Walker et al., 2013). Importantly, even though participation in discussion forums was limited, a steady stream of visits to the online portal throughout the intervention period demonstrated that the platform was regularly accessed by practitioners to download resources, read best-practice messages posted by the support partners and revisit training videos. The online portal was therefore still a valuable addition to the intervention.

Most queries about the implementation and delivery of the intervention were instead dealt with during the regular, prearranged online support meetings. Each school received three 30-minute support

sessions between February and June 2021, the period schools were running the intervention. (Later in this chapter, we outline the model of support provided for schools in more detail.) To ensure useful information was shared widely, other methods of dissemination were utilised. This included a regular e-newsletter sent to all school staff, with links to resources, best-practice advice from other schools (including information that practitioners volunteered during online support meetings and had agreed to share), articles on common queries and questions that had arisen from support appointments or during the training. The e-newsletters were viewed between 130 and 175 times.

The model and content of Flexible Phonics

We next outline the model and content of the Flexible Phonics intervention. This content is directly motivated by the research, as reported in Chapter 2, that suggests the need for children to be afforded more efficient and generative strategies to read all words they encounter in books in order to become fluent readers. Alongside synthetic phonics and necessary item vocabulary knowledge, Set-for-Variability potentially provides such an efficient reading acquisition mechanism for the

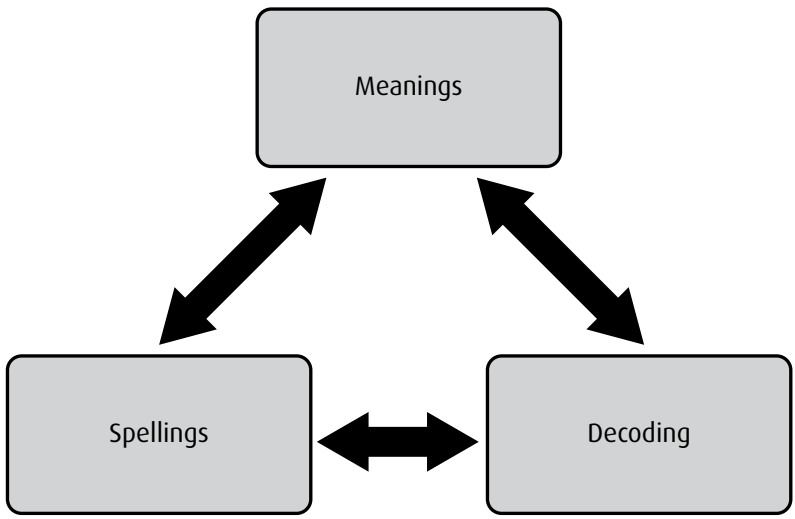


Figure 4.1 A modified triangle model for successful reading development in children (Source: Adapted by the authors from Seidenberg & McClelland, 1989. © United Kingdom Literacy Association).

complex and inconsistent spelling system of English. Teachers were shown a version of a triangle model to illustrate where Set-for-Variability might be situated for reading. The model, adapted by the authors from Seidenberg and McClelland (1989), is presented in Figure 4.1 and has also been published in Boldrini et al. (2023, p. 74). The connectionist triangle model suggests that successful reading development involves a high level of interplay between decoding (speech sounds and grapheme–phoneme correspondences, or GPCs), orthography (the spelling system, how words appear in print) and semantics (the meanings of words) (Seidenberg & McClelland, 1989; Plaut et al., 1996; Perfetti & Stafura, 2014). Both learning to pronounce a written word and then subsequently fluently recognising it can be achieved through the combination of such print-to-sound and print-to-meaning-to-sound mappings.

Teachers were taught that, according to this model, there is a division of labour for exception words. Orthographic–phonological mappings require the operation of mappings from print to meaning, while regular words depend on spelling-to-sound mappings. For exception words, semantic knowledge can be useful when words can only be partially decoded using regular spelling-to-sound decoding rules. This partial decoding can be combined with semantic knowledge and context of the sentence in which the word appears to enable the reader to carefully select the correct word from their mental dictionary (Share, 1995).

As mentioned earlier, the Flexible Phonics intervention consists of five aspects:

1. Direct Mapping
2. Set-for-Variability
3. Teaching vocabulary: exception words
4. Mispronunciation correction: a strategy for reading key exception words
5. Support for struggling readers: specific help for children who have struggled to learn how to blend simple consonant-vowel-consonant patterns.

Figure 4.2 illustrates an overview of the intervention. Each of the components are discussed in the following subsections.

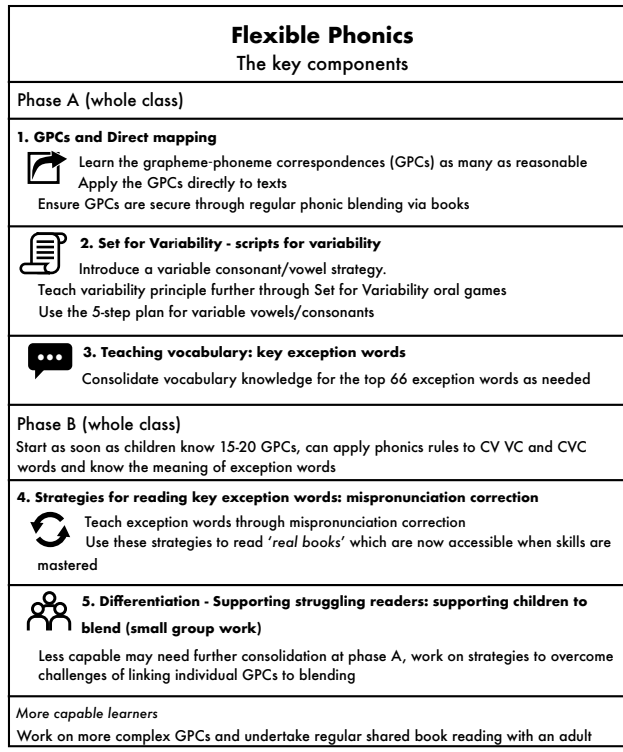


Figure 4.2 Overview of the components of the Flexible Phonics intervention © Authors.

Direct Mapping

In Direct Mapping, teachers were told that children should be explicitly taught a recommended sequence of GPCs and the following procedure. First, children pronounce the GPC of the day and write and decode it in the context of the word of the day. Then, on the same day, instances of the GPC(s) taught are explored through a meaningful context, such as shared reading in the classroom of a well-known children's book containing a high density of those GPCs. In this way the teacher supports the children to read the words.

In the context of reading texts, children are asked by teachers to identify the grapheme and decode – or, with support, partly decode – words containing the grapheme, thereby applying and consolidating knowledge they have just learnt in a structured way. This daily linking of GPCs to children's books is a key idea in the intervention. The aim is for children to understand the relevance of GPCs in texts, using them

immediately in context and linking these to reading success rather than learning them as a set of abstract facts. In this first aspect, teachers can choose how they present the GPCs and text – as a whole-class activity or in small groups, depending on the abilities of the children in the class.

To facilitate this, sets of high-quality, well-known children's books were supplied to all the schools taking part in the intervention. The books contained a high frequency of GPCs so that children would have many opportunities to spot, read and reinforce the GPCs taught. These books were chosen from a list of award-winning children's books generated from the charity BookTrust (https://www.booktrust.org.uk/) and other texts reflecting England's cultural diversity. This gives children the opportunity to link GPCs to books, helping them to understand explicitly how GPCs are used to read authentic texts. The use of children's books was not intended to replace all scheme texts in Flexible Phonics, and we did not encourage schools to abandon their current scheme books. The aim was for children to be able to read books as soon as possible, as high-quality books introduce children to language forms and structures and wider and more complex vocabulary. This, in turn, supports wider language development and potentially facilitates their progress in learning to read (Mol & Bus, 2011). As noted in Chapter 2, there is no consistent evidence that the use of real texts impedes the teaching of phonics (Jenkins et al., 2004; Cheatham & Allor, 2012).

Teachers were made aware that there is evidence of this approach of teaching phonics in the context of book reading being more successful than teaching phonics alone, particularly when GPCs are taught, applied directly to books and read in the same session (for example Hatcher et al., 1994). Teachers were similarly made aware of the evidence for the Direct Mapping approach of explicitly linking the learning of GPCs to a relevant reading task (a text containing a high frequency of GPCs) (for example Chen & Savage, 2014; Savage at al., 2020) and of its effects in improving reading motivation (Chen & Savage, 2014).

The GPCs taught in Flexible Phonics were carefully selected based on three principles: frequency, ease of blending and variability. Influenced by the evidence for the Simplicity Principle, as described in Chapter 2 (Vousden et al., 2011; Savage et al., 2020), the order for teaching the GPCs in the Flexible Phonics intervention was based on the frequency with which they occur in the selected children's books. For the intervention, 35 high-quality children's books were analysed to identify the most useful or most highly occurring GPCs. Based on this analysis, it was determined which books had the highest occurrence of given GPCs to aid the teaching of Direct Mapping. Ease of blending was another

important factor in the order of GPCs taught, based on evidence from Gonzalez-Frey and Ehri (2021) which suggests that certain GPCs should be taught ahead of others, such as *f*, *m* or *s* ahead of *t*, *d*, *g* or *k*.

Finally, it is important that children gradually encounter the variability that occurs in GPC pronunciation in different word contexts. This will help them understand the realities of the English spelling system in order to develop these skills and generalise these principles to other words. For example, in the Flexible Phonics intervention, *c* (in *cat* and *face*) is taught with a strategy for working out which pronunciation is appropriate (see Set-for-Variability, the second aspect of Flexible Phonics, below). This will help when children meet the words *can* or *nice*, giving them a strategy to become independent readers. In line with a wider analysis of texts (Gontijo et al., 2003; Vousden et al., 2011), voiced and unvoiced *s* (c.f. *us* and *uses*), *g* (c.f. *giraffe* and *gorilla*) and *y* (c.f. *you* and *happy*), and variants of *ed* (c.f. *stopped* versus *wanted*), are taught as part of the intervention, as these are some of the most common variable GPCs.

Set-for-Variability

The second aspect of the Flexible Phonics training was Set-for-Variability, which focuses on teaching children ways to tackle the reading of exception or irregular words. For example, in *was*, the middle grapheme is pronounced differently from its regular decoded form, such as if the graphemes *w*, *a* and *s* were read in the regularised forms 'w', 'æ' (as in *cat*) and 's'. If read in the regularised form, this would result in the pronunciation 'wæs', which is not a word, thus resulting in a mismatch between the decoded form (regularised form) and the word pronunciation. Teachers were taught that Set-for-Variability is a process that remedies this mismatch through the child comparing the phoneme string produced via phonic blending to words stored in their mental lexicon (Savage et al., 2018).

Children are first introduced to and practise Set-for-Variability through oral tasks, where they orally correct an exception word pronounced in a regularised way by substituting alternative vowel or consonant pronunciations. This is so they can practise mental flexibility orally, in an accessible game format, before they attempt reading exception words. Activities include a range of oral tasks, such as 'Simon says touch your . . . ar-m/k-n-ee' to teach the component skill of word recognition from given spelling pronunciations for both regular and exception words, within a clear semantic category (here, body parts; elsewhere, animals or numbers). Teachers were encouraged to first use the 'Simon says' game to

establish vocabulary knowledge for body parts, and then to regularise the pronunciation of words such as *shoulder* (where *ou* is pronounced as in *out*). Finally, teachers were encouraged to present regularised phoneme strings, such as 'sh'-'ou'-'l'-'d'-'er', with the aim that children link that string to their shoulder.

Vocabulary

The third aspect of the intervention training was focused on teaching children the *meaning* of exception words. We told teachers that research suggests that a child is likely to employ semantic information to determine the correct pronunciation of novel exception words when reading aloud. We shared the Flexible Phonics approach, which suggests that children should be taught the meanings of the top 66 most frequently occurring exception words in children's books. These are words that occur at least once in every 1,000 words in real texts in Masterson et al.'s (2010) children's book corpus analysis.

The teaching of these words is clustered, so they are linked around a shared theme. This is particularly valuable for function words, which include pronouns, prepositions, articles and conjunctions that occur with a high frequency in the English language. For example, activities such as a scavenger hunt are used to teach the meanings of prepositions in context. Conceptually related words are taught as clusters in order to help children learn conceptually linked categories of words as part of semantic networks (for supporting research, see Hadley et al., 2019; Manyak & Latka, 2020). A central idea for learning and retaining words is that children need to have multiple encounters with the new words, seeing the printed words to consolidate them into memory (Coyne et al., 2007; Ricketts et al., 2021). In line with this research, and with associated research by Biemiller and Boote (2006), consolidating by reviewing and using words that have been taught previously – alongside teaching new sets of words – will help build lasting, deep word knowledge. For each word taught, the age of acquisition for when children start to use these words accurately is given as a way to guide teachers on which words to focus on first (Brysbaert & Biemiller, 2017).

Mispronunciation correction

Mispronunciation correction in the Flexible Phonics intervention is the print-focused deployment of Set-for-Variability. Here, children are taught an explicit five-step strategy to deal with vowels and consonants with variable pronunciations in printed words. Children apply standard

synthetic phonic decoding as a first step, then substitute and apply different phoneme pronunciations, and then check these attempts against known words in their lexicon. This teaches children to be flexible in their application of phonics and to potentially read regular and exception words by correcting mispronunciations.

A five-step plan for variable vowels and consonants

In this aspect of Flexible Phonics, children are taught the following procedure:

1. Blend phonemes of a letter string, looking for and applying well-taught phonic rules. Children need to be able to apply these well-taught rules accurately (as taught in aspect 1 of the intervention).
2. Evaluate their first attempt to synthesise a pronunciation: 'Is this a word I know?'
3. If not, replace the vowel or consonant with an alternative they know – 'swap in another sound'. Teachers support this process as appropriate. Children will have been exposed to some common variants as part of the first section of the intervention.
4. Synthesise this revised phoneme string.
5. Re-evaluate this blended string using the same reflective lexical decision process as above: 'Does this make a word I know?'

Children are also supported by reading the word in a context that can help make the meaning clear. The sentence context thus acts as an aid to help children work out the word. During the training, teachers were taught and frequently reminded that Set-for-Variability is rooted in a synthetic phonics process of applying GPCs to sound out words; as such, this process is firmly not 'guessing' word pronunciations.

Continuous phonation

Continuous phonation was a strategy shared with teachers as being appropriate specifically to help only those children who had yet to master phonic blending within a monosyllable. It was thus presented as a precursor to the use of Set-for-Variability in such individual cases as warranted. The content draws in part on the work by Gonzalez-Frey and Ehri (2021), reviewed in Chapter 2. To aid blending, children were initially taught to blend only with continuant phonemes – those that can be continuously sounded (for example 'aaaaaaaaammmmmm' as letters *a* and *m* are pushed closer together in space). Gradually, as they child shows emerging expertise, non-continuant phonemes are carefully added.

Support for schools post-training

Due to the Covid-19 pandemic and schools being partially closed from 5 January to 8 March 2021, most teaching was delivered remotely, with only a few key workers' children or vulnerable children taught face to face in school. The original intention was for schools to deliver the Flexible Phonics intervention straightaway, from mid-January until the end of May, but given the challenges of delivering online teaching, many schools preferred to wait and deliver the intervention once face-to-face teaching had resumed on 8 March. A few schools attempted to deliver aspects of Flexible Phonics online, although there was no expectation to remotely deliver given the sizable pressures schools were already under. Most schools used this period to integrate Flexible Phonics into their planning, so they were ready to start delivery once fully opened post-lockdown.

The intervention was thus extended until July to give schools sufficient time to incorporate the strategies into their teaching. The expectation was that schools would deliver core aspects of the intervention at least three to four times a week. To support integration, suggestions were made by the delivery team on how this could be accomplished in a busy curriculum. For example, classes could undertake a sound hunt to look for GPCs in books in the book corner, do a Direct Mapping activity or use transition times to develop oral flexibility activities.

Originally it was planned that, following Flexible Phonics training, schools would receive two in-class follow-up visits where members of the delivery team (with the teacher's consent) would observe the intervention in the classroom context and provide further feedback and guidance around delivering Flexible Phonics. However, due to partial school closures during the pandemic, these were instead run as three prearranged online support appointments, of 30 minutes each, via video-conferencing software, led by a Flexible Phonics support partner. The support partners were members of the UCL delivery team, with extensive experience of working as practitioners or in education-related roles. Each school was allocated their dedicated support partner, who was their first point of contact for queries or additional support outside of the prearranged appointments.

The support partners' role was to work in partnership with schools and to provide ongoing support with implementing and delivering the Flexible Phonics intervention following the training. Through this collaborative way of working, both partners had equal status in the relationship: the delivery team brought their expertise in designing and

running reading interventions, and the school brought their experiences of teaching and knowledge of their pupils and classrooms. In education research, the valuable knowledge and expertise of practitioners have often been ignored in favour of a one-directional relationship of research influencing practice. This can sustain an unhelpful dichotomy between research knowledge (implied to be more important) and practice-based knowledge (implied to be less important) (Cochran-Smith & Lytle, 1999; Niemi, 2008). Closing this research–practice gap can be achieved by working to involve practitioners in research in a collaborative manner rather than conducting research 'on' them (Parsons, 2021). Considering both partners as equal-status collaborators who each bring useful experience to collectively solve problems and share knowledge can work better for all involved.

Depending on schools' preferences, appointments were conducted either individually with the practitioners delivering the intervention or with the year group. In group sessions, the Reception team were encouraged to facilitate information sharing through a collaborative community of practice approach. Support partners could offer greater flexibility around meeting times, accommodating schools who preferred to meet during the school day rather than after school. The majority took up the non-obligatory offer of support, with 95 per cent of schools accessing all three support appointments. Online meetings offered greater time efficiency for both parties, although for some schools these could be more challenging to schedule. In such situations, emails were sent sensitively to highlight the usefulness of the support and to assist the research team to find out how the intervention had been implemented. This way of working was well received by some schools, as evidenced from informal feedback from schools via email.

This collaborative model was in contrast with other types of training that some schools had previously received, where there was no opportunity for discussion or feedback about whether they were implementing the new strategies in the intended manner. Several support partners reported that, at the start of delivering the training, schools were apprehensive about how it would work in practice and appreciated the tailored support and the reassurance that they were on the 'right track'. This positive experience was shared from written accounts of the support partners' experiences, as illustrated below:

> Certainly, I feel that ongoing support allowed for more successful implementation of Flexible Phonics. Support appointments offered an opportunity for teachers to ask questions following the online

> training or as a result of trying out aspects. In addition, we had insight into how Flexible Phonics was being delivered and could offer support and advice where appropriate. Misunderstandings of any aspect of the intervention could be swiftly identified and remediated.

The allocation of a dedicated support partner was mutually beneficial for both parties, as it helped to build strong working relationships, establishing trust and rapport. This consistency allowed the support partners to obtain a full picture of the school context, including class characteristics, teaching and streaming arrangements, and the schools' approaches and the challenges they experienced. Having such knowledge allowed for higher-quality support, as the support partners were able to track and monitor progress between meetings, follow up on previous queries or areas for further development, and anticipate what could help support and facilitate schools' delivery of the intervention. Outside of these online appointments, the delivery team was available to schools on an as-and-when-needed basis. The uptake of this service was limited – only three schools contacted the support partners for as-and-when support – and the research team thus proactively made contact fortnightly to ensure staff were well supported and that useful information was shared.

To ensure high quality and consistency in the support given to schools, weekly team meetings between support partners and the project manager were held to review support and collectively share good practice. Outside of these meetings, the support partners formed their own community of practice. This extract from a written account by a support partner describes their experiences of online support:

> 'We tended to share the request from the school with one another, elaborating our initial thoughts with the other support partners to discuss various ideas as a group and enhance the quality of advice given to teachers.'

As support was provided remotely, schools were offered the option to share videos of their Flexible Phonics practice for supportive feedback from the delivery team, although no schools chose to do this. With the schools' permission, written records of online meetings were made to enable the research team to understand which aspects of the intervention were being implemented and how these were being delivered in practice in the classroom context.

Structure of online support

Consistent with the construction approach, a four-stage flexible framework based on the solution-focused coaching model (Grant, 2012) was applied to structure the online appointments. Solution-focused coaching is an approach where the emphasis is on solution construction in preference to problem analysis. Here, the coachee is considered capable of solving their own problems as they, rather than the coach, are the expert in their own life. As part of this approach, the coach helps the coachee to recognise and utilise their existing resources and, through collaboration, develop a solution to their problems.

After receiving extensive training in applying these principles, the support partners assisted practitioners to use their own professional knowledge and expertise to build solutions to issues around the implementation of the intervention and to take ownership of these solutions using a variety of techniques to promote positive change. In the first stage, using active, non-judgemental and reflective listening, partners collected information about how and when practitioners had been implementing and delivering Flexible Phonics, asking about lesson planning, differentiation and challenges. The second stage involved identifying either a problem or an area for further development by asking questions to elicit more information and paraphrasing the concern to check for accuracy and clarity.

The third stage involved co-constructing possible solutions. A key focus was working in partnership, using targeted questions to support practitioners in realising a solution to their concern or issue using their own knowledge and expertise. The final stage, resolution, included a summary of the solution and agreement on a plan of action so the practitioner would be confident to take this forward. The support partners considered the solution-focused model to be essential in enabling partnership working. An extract from a written account of a support partner's experience illustrates this further:

> The solution-focused approach shaped the way I engaged in discussions with schools, using open questions to facilitate the conversation, paraphrasing what was discussed to make sure they felt heard and understood. In case of issues, the solution-focused approach helped in offering useful and tailored support towards a solution, constructing a plan for the next steps working together, rather than imparting instructions.

Themes from online support

From a content analysis of the written accounts of these support appointments, it was apparent that the schools' implementation of the intervention moved through distinct stages. The first appointment, conducted between mid-February and mid-March, highlighted that the majority of schools were at a planning and practising stage, considering how to integrate Flexible Phonics with their current phonics provision teaching.

The pandemic posed a considerable challenge for school staff, as partial school closures until 8 March meant teachers had to co-ordinate teaching children of key workers and vulnerable children face to face in school whilst also offering remote provision. These challenges placed substantial limitations on schools' delivery of the intervention, with many schools delaying the implementation until face-to-face teaching resumed. There were reports from some staff that they did not see their pupils online consistently, partly due to the sharing of devices amongst families and parents being unable to supervise and monitor children's activities whilst working from home.

Despite these challenges, partners reported that a few schools tried out aspects of the intervention, such as the oral Set-for-Variability games, online, either live or in pre-recorded sessions to ease children into Flexible Phonics once they returned to face-to-face teaching. These schools reported that differentiating learning via remote teaching was particularly challenging. It was also not possible during the initial support meetings to gather much data about how schools were delivering the intervention. However, these meetings did provide an opportunity to answer queries or address misconceptions about the intervention and to gather baseline information from the class. The support partners reported that these meetings provided a good foundation, fostering positive relationships and establishing rapport, which likely led to more detailed qualitative data being gathered during subsequent appointments.

During this period, most queries raised by practitioners during appointments were related to planning and how to integrate Flexible Phonics alongside their current phonics programme. Unsurprisingly, the support partners mentioned that Flexible Phonics appeared to work well beside programmes where there was significant overlap between the intervention strategies and the strategies teachers were already employing, allowing them to move forward with their practice. This contrasted with more prescriptive schemes, as described by one support partner: 'Schools using other schemes quite rigidly . . . I found that it

initially took time for schools to see how to tie the strategies into what they were already doing.'

Once schools had fully reopened between March and May, they reported the added challenge of managing infection control along with intervention delivery. In some instances, local Covid-19 cases led to 'bubble closures', where children and staff were sent home to self-isolate. Some schools had no such closures, whereas one school experienced four bubble closures, which inevitably impacted the delivery of Flexible Phonics, although teachers did their very best to continue delivering this remotely.

Staff absence due to Covid-19-related illness and staff changes were other substantial challenges for schools and the delivery team. Some teaching staff had missed the online training, and in these cases recorded training sessions were shared so staff could catch up. Understandably, there were a few occasions where staff could not fully catch up on the training, and here the support partners did their best to offer additional guidance and advice. Additional support, including one-to-one tutorials run by the academic lead and their support partner, were offered to get staff up to speed.

Due to the pandemic, schools in certain London boroughs reported an increase in the number of families migrating out of London, either to work remotely or to return to their country of origin. Pandemic school closures and the resultant learning loss meant that new starters sometimes came with a range of challenging behaviour, language needs or socioemotional needs. In short-staffed classrooms without teaching assistants, attention was paid primarily to settling in pupils, as described by one support partner:

> I often heard about children arriving having challenging needs relating to their behaviour, language development or emotions. One of my schools began implementing Flexible Phonics quite late due to a pupil that joined with a range of needs – she felt that all her focus had to be placed on supporting him. Unfortunately, consistent Flexible Phonics implementation took a back seat.

There were other consequences of school closures that impacted children's development more broadly, which in turn affected teachers' ability to implement Flexible Phonics. Several support partners reported teachers commenting that children showed delays in their language development and their emotional and social development. This might be attributable to reduced uptake of pre-school and nursery provision before joining

the school, as well as a lack of opportunity to socialise with peers and family members as a consequence of national lockdowns. One teacher commented: 'Children are behind with communication and language, personal, social, and emotional milestones. We are spending much more time on this than in an average year.'

By the time of the second appointment (March to early May), the support partners reported that most schools had started to implement the Direct Mapping and oral flexibility aspects, with many implementing more extensively following the Easter holidays. Following the return to face-to-face teaching, where this occurred, the gradual integration of Flexible Phonics strategies seemed to be working well. One such example, provided by one of the support partners, is described in Box 4.1.

Box 4.1: School-wide uptake of Flexible Phonics

'One school taught phonics in small ability groups and focused upon Direct Mapping with each group, choosing the GPCs which were appropriate for the children. The teacher worked with each of her five groups, using real, rich picture books supplied as part of the Flexible Phonics programme, to engage the children in identifying the next GPCs relevant to their learning. She also used the text to work on helping the children learn key high-frequency words from the Flexible Phonics list and steadily introduced the mispronunciation correction strategy to some groups through modelling and having fun with reading words phonetically.

'After trialling these strategies with small, differentiated groups online, it was clear to the teacher that knowing which book contained which sounds would be helpful, and so in the spirit of co-construction, the school support partners created a list which enabled teachers to identify which texts contained the highest number of each high-frequency word. The school went on to use this resource and extend its use, with books for students in Year 4 and Year 5, where training in these strategies was provided in-house to teaching assistants. The school felt strongly that the Flexible Phonics strategies should be shared throughout the school so that all children could benefit.'

During this period, several misconceptions were commonly picked up by the support partners. These are highlighted below:

> For example, [school name] thought that they were teaching vocabulary in a Flexible Phonics-approved way just by having a Word Wall.

> I felt as though some staff had not gone through the manual and were relying on what they remembered from the online training sessions in January. [School name] mistook Flexible Phonics as a standalone 'scheme' rather than a toolkit of strategies to complement their existing phonics scheme. This meant at the time I expected Flexible Phonics implementation to rocket, I was re-introducing and clarifying aspects instead.

Another query raised by several schools with a high number of children with English as an additional language (EAL) was the concern that some children had spoken or read very little English at home during lockdown. This may have led to this group of children having a lower retention of GPCs that had previously been taught.

Other concerns from teachers were surrounding the suitability of Flexible Phonics for children with EAL. Although this was clarified during the online training sessions by the academic lead, some schools still needed further clarification and support in understanding how aspects such as mispronunciation correction, vocabulary and Set-for-Variability would be appropriate: 'Something that often came up during the initial appointments was the concern for students with little to no English, as the children do not have the foundation to grasp some of the graphemes.' The issue of suitability of Flexible Phonics for children with EAL was addressed with suggestions surrounding its implementation in the newsletter sent out to schools and as resources to the community of practice. However, even with this advice, some of the teachers judged that the aspects were too difficult for their students with EAL; therefore, Flexible Phonics was not carried out as it was 'not applicable for them'. Nevertheless, where schools appeared fully engaged with the strategies and were proactive in implementing these once the children returned to face-to-face teaching, teaching staff reported that these strategies helped to close the gap between EAL and English first language (L1) learners.

By the final appointment (mid-May to June 2021), the support partners reported that schools were well underway with delivering Flexible Phonics, sharing examples of their successes with the delivery team:

> The Reception Team in [school name] felt strongly that Flexible Phonics provided excellent strategies in helping children become effective, confident readers . . . To maximise the impact of the strategies, she introduced the strategies at teaching and learning sessions and staff meetings. Secondly, she planned training for

> staff in other year groups to ensure staff were upskilled to support children who were behind 'expected levels' in reading. Staff from Year 2 had commented that knowing these strategies would be especially useful for teaching. The teachers felt so passionately that Flexible Phonics had made a significant difference to all children across the year group that not only did they share their perspectives within the school, but they also shared the principles of Flexible Phonics at a local cluster meeting, resulting in several schools being keen to learn more and come on board, should further opportunities become available.

During support meetings, some teachers shared how Flexible Phonics had led them to reflect more extensively on their existing phonics teaching, as described by one support partner:

> One teacher mentioned how Set-for-Variability and mispronunciation correction made her change her mind about children making mistakes, and that it is OK to make mistakes as long as this is turned into a learning opportunity. Another teacher realised that more application to real books and activities throughout the day can help children make sense of GPCs and their function.

A recurring theme from support meetings was that many schools expressed an interest in carrying on Flexible Phonics into the next academic year, requesting further training for their Year 1 (age five to six) staff. To meet the need, a two-part online training session was run in July at the end of the trial, which was attended by over 120 staff from 29 of the participating schools. This additional training covered the same five components of Flexible Phonics as described above for Reception teachers. The aim was to explain the approach, explain what Reception teachers and teaching assistants had taught and to provide ways to use Flexible Phonics in age-appropriate ways in Year 1. Here, for example, the potential of Set-for-Variability in managing adjustment of phonics in the context of bi- and polysyllabic words was explained, as were diphthongs and other language features not strongly emphasised in the more introductory Reception phase of educator professional development. Feedback from the educators present indicated their interest in pursuing Flexible Phonics in Year 1.

Overall, the support partners reported that using open questions, engaging in active listening and taking a solution-focused approach to encourage self-reflection were effective at uncovering the extent to which

schools were implementing Flexible Phonics. However, as this was run remotely there were some challenges, particularly if the practitioner had turned their camera off during a support meeting. For example, in these cases it was hard to pick up on attitudes and feelings during the meeting, as non-verbal cues and body language were not visible. There are limitations with using self-report methods to determine intervention adherence, and validity is threatened by social desirability and selective reporting effects.

Treatment fidelity

Next, we focus on the global treatment fidelity rating (GTFR) tool and how this was used to capture the extent of implementation of the Flexible Phonics intervention in schools. Understanding how teachers implement interventions is an important part of understanding why interventions are successful or not (Varghese et al., 2021). Originally, the support partners intended to visit schools in person and observe practitioners delivering Flexible Phonics in order to rate compliance using an observation tool. Covid-19-related visitor restrictions in schools meant these were all moved to online meetings. Based on discussions with schools during the three online follow-up support sessions, treatment fidelity was rated using the global teacher fidelity rating tool. This was a substantially adapted version of an implementation fidelity measure rubric developed by Savage et al. (2013). The support partners used evidence collected from discussions with their schools during the three follow-up support sessions, email conversations, posts on UCL Extend, an online portal and any resources shared by schools to determine the extent to which each of the five aspects of Flexible Phonics had been implemented, the levels of adherence and quality of implementation.

The global teacher fidelity rating consisted of a four-point scale, ranging from zero to 3, where 0 = no implementation of Flexible Phonics; 1 = entry level: some (but likely poor-quality) implementation; 2 = adoption: clear and competent regular delivery of the intervention; and 3 = adaptive delivery: expert and extended delivery of the intervention. To be compliant, a score of 2 or 3 was required on the core aspects of the intervention: Direct Mapping and oral flexibility and/or print-based Set-for-Variability flexibility. For the Direct Mapping aspect, N/A could be a possible option if Direct Mapping was not required; for example, children might have moved past this element if they already knew all the GPCs. To establish reliability, global teacher fidelity ratings were double rated by

two different support partners, and inter-rater agreement was analysed using Fleiss's Kappa, which was 0.77 across the three core areas: Direct Mapping, oral flexibility and print-based flexibility.

Levels of implementation: key themes

Where ratings from the global teacher fidelity rating tool were low, several other themes could be identified besides the pandemic-related challenges mentioned above. One-form entry schools which were teaching whole-class phonics reported that they found the pressures of planning the curriculum – along with the additional challenges of the Covid-19-related workload – particularly demanding. One other school described introducing Flexible Phonics strategies through whole-class teaching as challenging. Some teachers felt that certain children were being exposed to teaching through Flexible Phonics that they were not quite ready for.

Notably, the most enthusiastic and confident implementers of Flexible Phonics were based in schools that reported being fully invested in the intervention as a team during support meetings and who were open-minded to trying a different approach to teaching phonics. These teachers often reported feeling well supported by their senior leadership teams, who had taken an active interest in Flexible Phonics. This extract from a written account of a support partner's experience demonstrates that the Flexible Phonics intervention:

> . . . resulted in a wider school shift . . . a culture change which sought to encourage children to read more storybooks in school and at home rather than just [phonetically decodable] scheme books. Teachers felt well supported to deliver Flexible Phonics and the enthusiasm from the senior leadership team was infectious.

Conversely, where teaching staff were more reluctant to implement Flexible Phonics – either due to an overwhelming workload created by the pandemic or through a lack of staff buy-in to the approach – the delivery of the intervention was less extensive. This was also associated with confusion about the Flexible Phonics approach, which persisted despite educators being taught that Flexible Phonics emphasises the acquisition of printed exception words and not their fluent subsequent reading: 'A recurring obstacle that I [support partner] met was engaging with teachers who did not necessarily want to be flexible in their teaching

methods. Most complained that mispronunciation correction would not work as the children had been 'programmed' to recognise tricky words by sight.

Working in partnership with parents to reinforce the Flexible Phonics strategies at home was apparent in some of the schools that implemented the approach consistently. For example, two schools worked to involve parents, maximising the use of children's books and modelling Flexible Phonics strategies to parents so they could reinforce these at home. In another school, the reception classes had a core book (children's books linked to a GPC of focus) for the week, accompanied by two others. The core book was used for shared reading at the end of the day, while the others were for children to use during continuous provision and were sent home to read with parents. Parents were involved in the implementation of Flexible Phonics by teachers sharing the strategies, explaining Direct Mapping and mispronunciation correction so these strategies could be reinforced in the home environment.

Another example of how working in partnership with parents was successful is illustrated here:

> While teaching remotely, teachers shared the strategies in Section 5 of the manual with parents, receiving positive feedback. Once [the children had] returned to school, the parents of children who had particularly struggled had a one-to-one with the teacher where she explained Flexible Phonics strategies more in-depth, also providing the logic behind the continuous phonation and starting with vowel-consonant (VC) words. She then gave a list of VC words to the parents so they could work on them with the children at home.

Strong implementors were also characterised by their usage of children's books within the wider curriculum. This meant that children were further exposed to Flexible Phonics strategies and the practical use of GPC knowledge outside of phonics sessions in the home-learning environment.

Finally, it is important to note, in spite of the challenges described above, that the attrition rate during the delivery phase was minimal. As noted, only one school delayed carrying out the intervention until the following academic year due to pandemic-related work pressures and the challenges of simultaneously carrying out another EEF early language intervention in the same classroom. This school had signed up for a second EEF-funded intervention after signing up for Flexible Phonics.

Conclusions

This chapter has focused on the Flexible Phonics training and post-training practitioner support model. Issues associated with the intervention training in the context of the wider Covid-19 pandemic – and how these issues were managed – have been discussed, including the support provided for schools and the support partners' experience of supporting schools with the Flexible Phonics intervention.

Through solution-focused support, we drew out the key themes associated with schools' experiences of implementing and delivering the invention. We argue that a key aspect for successful professional development in education is equal-status support processes for school partners. In this model, practitioners were placed at the centre; their professional expertise was valued, supported and actively facilitated through partnership working with the delivery team. We have also discussed the tools that were developed to assess implementation fidelity. Finally, we explored the different ways in which the intervention was delivered by schools and noted the significant limitations in our certainty on these data given the remote support necessitated.

5

Involving schools in the conversation: practitioners as partners

In this chapter, we first explain the importance of involving practitioners in the research process from the outset and why listening to practitioners can provide valuable insights and knowledge about what will work in a school context. Next, we explain how this occurred in the Flexible Phonics trial, highlighting aspects of the trial where active collaboration with schools happened and describing how this informed decision-making at key points in the trial. We highlight practitioners' perspectives about their involvement throughout the trial – from the pilot study to the main trial and during the support phases. Throughout we discuss the many challenges schools faced during the Covid-19 pandemic and highlight the factors that affect educational collaboration in a real-life context, reflecting on the complexities and realities of the unprecedented impacts of the pandemic during the trial delivery phase.

Involving practitioners in research

A large and well-developed body of literature frequently highlights the gap between educational research and practice (Hargreaves, 1996; Chi, 2021). There is often a disconnect between those working in practice and those undertaking university-led research (Snow, 2015). On the one side, researchers claim that the results from their work are highly relevant for teachers and can inform their practice. On the other, practitioners have argued that educational research is too far removed from the classroom, too theoretical and inaccessible (Silver & Lunsford, 2017), addressing questions and issues that are not important to them or their practice (Gutiérrez & Penuel, 2014), and that the knowledge generated from research is overly broad and too abstract to implement easily in the classroom (McIntyre, 2005).

A number of writers have suggested solutions to bridge the research–practice gap. These include providing practitioners with knowledge (Cain, 2015) and training to help them understand research reports and how to critically evaluate such reports. Other suggestions include providing more practical, context-specific guidance, written in less academic and more practitioner-friendly language (Hirschkorn & Geelan, 2008), and creating guidance for how to implement research findings in classroom settings (Walker et al., 2018), particularly the pragmatics of what to teach and when (Seidenberg et al., 2020).

For practitioners not affiliated with a university, accessing research ideas can prove challenging, as many articles are located behind expensive paywalls (Rycroft-Smith, 2022). The open-access movement, by contrast, makes research publications freely available to all through sites such as ResearchGate, making research findings more accessible and encouraging information sharing. However, given the demands of practitioners' roles and workloads, they are generally time-poor. They often lack the time to read up and reflect on research findings (See et al., 2016) and determine whether an intervention could work in their classroom context (Joyce & Cartwright, 2020), even when specific guidance is given and research is more accessible.

One approach to encourage knowledge exchange is to create research translation roles to bridge the gap between practitioners and researchers (Hirschkorn & Geelan, 2008). Research translation is typically defined as a mediatory or boundary-spanning role that facilitates the flow of knowledge between research, practice and policy, using a variety of methods (Malin & Brown, 2019). These are professionals who understand research, practitioner perspectives and practitioner communities and who have deep contextual knowledge and experience of both research and practice (Hirschkorn & Geelan, 2008; Rycroft-Smith, 2022). When translating research findings into accessible forms for practitioners, these professionals can convey the interests and concerns of practitioners to researchers. However, quality evidence about the effectiveness of these research translation roles is limited (Rycroft-Smith, 2022). Gorard et al. (2020) report there is little evidence to suggest this type of role in itself, when undertaken by someone in a school, is effective in raising pupil attainment. There is some evidence to suggest that these approaches may positively influence teachers' beliefs and attitudes towards and engagement with research.

To bridge the gap from research to practice, the Education Endowment Foundation's (EEF) Research School Network has been developed to support schools to access and use research evidence to

improve teaching practice. This is done by encouraging schools and senior leaders to make the best use of evidence through training and school-to-school support. Buy-in and support from senior school leadership was found to be crucial in changing practice in schools and, where access to training and evidence was aligned with schools' improvement priorities, sustained uptake was more likely (Gu et al., 2020).

The area of translational research emerged recently from public health and biomedical science disciplines (Woolf, 2008). First introduced in the medical literature about 30 years ago and sometimes referred to as 'bench to bedside', it aims to translate research findings into research-informed programmes, interventions or classroom practices (Chawla, 2018). More of an iterative process than a linear one, this might include activities such as translating a scientific article for a practitioner or public audience, testing out reading theories in a classroom environment (such as in the Flexible Phonics intervention) or disseminating research findings to improve literacy outcomes as part of teacher training sessions. In practice it may be tricky to find one individual to undertake this position, given the specialist nature of such a role; instead, it might be best suited to an interdisciplinary research–practice team with a varied skill set (Terry et al., 2021).

It is evident that the strengths of both researchers and practitioners are needed. Roles in such a partnership may include researchers harnessing practitioners' experience as programme implementers and their knowledge about the practical implications, barriers and facilitators in order to better understand how to successfully translate theory into school-based settings (Race, 2010). Similarly, Solari et al. (2020) and Terry et al. (2021) suggest that a team-based approach is a critically important factor for the translation of the science of reading into authentic school-based settings. For these reasons, in the Flexible Phonics trial, we actively recruited research staff (the project manager and the support partners) with extensive, recent Early Years teaching experience who understood the priorities of teachers, the realities of teaching and the implications of translating research into classroom contexts. The team could thus provide practical suggestions and ideas around planning, implementing and embedding key aspects of the intervention into the school day.

Research collaboration is one way to bridge the research–practice gap, to contribute to professional development for teachers and to improve pupils' attainment (Schenke et al., 2017). In such collaborations, researchers work with practitioners as equal partners, drawing upon collective knowledge, expertise and experience (Carlgren, 2012; Kieran

et al., 2013) throughout the research life cycle, from formulating the research questions to designing the intervention and disseminating the research findings (Olin et al., 2023).

When this is not the case, and reading interventions are developed with minimal input, practitioners may feel that their knowledge and experiences are overlooked. The result may be a reduced motivation for implementation. This may particularly be true if the research does not align with the needs, interests and priorities of those working in practice and within the wider school system (Sullivan et al., 2008; Snow, 2015; Moir, 2018). Similarly, where teachers' beliefs and knowledge contradict the research ideas, or where research ideas are difficult to integrate into typical classroom practice, the implementation of an intervention tends to have a limited shelf-life (Scanlon et al., 1994). Where these beliefs and knowledge do align with the intervention content, there tends to be a greater degree of teacher acceptance for an intervention (Scanlon et al., 1994). To maximise this partnership process, McGeown et al. (2023) suggest that researchers and practitioners might benefit from mentoring, training and further support, as successful partnership working requires skills for which specific training may be needed.

Collaborative working in Flexible Phonics

In this section, we explain when collaborative working occurred in the Flexible Phonics trial and highlight aspects of the trial where such collaboration with schools occurred, including piloting, intervention design and support for schools. We detail how this informed key decisions about the design, management and implementation of all aspects of the trial.

In order to pilot the intervention before scaling up with 123 Greater London primary schools, a group of ten Early Years teachers and three teaching assistants from seven primary schools in London took part in intervention training in early February 2020. The training was delivered in two parts: one full day and a half-day face-to-face follow-up session one week later. The key purposes of the pilot were to determine teachers' acceptance of the intervention as well as their willingness to modify their existing phonics practice and to determine whether the key aspects could fit alongside existing phonics teaching and be implemented successfully in the classroom. Teachers' views about the accessibility and clarity of the training materials and training manual were sought to make modifications to the design of the intervention training and materials.

The intervention training was delivered with a partnership approach at the forefront, whereby teachers' practice and classroom expertise was acknowledged and valued. Practitioners were viewed as co-experts working collaboratively with the delivery team, with the shared purpose of improving phonics teaching and early literacy outcomes. A key premise of the training was a degree of principled and flexible implementation of the intervention. Overly strong demands to implement rigidly were thus avoided, in favour of emphasising wide, permitted variation based on practitioner judgement about the suitability of activities for particular children in their classrooms. This variation was around the delivery of the core aspects of the intervention – Direct Mapping and Set-for-Variability aspects.

When interventions are delivered by teachers in their classrooms, a non-flexible approach can risk undermining teachers' autonomy and their freedom to adapt an intervention for individual pupils or different settings (Green et al., 2019). However, careful guidance needs to be given, as offering too much flexibility in terms of implementation can threaten intervention fidelity (Wheatley et al., 2020). For this reason, in the Flexible Phonics trial teachers were given guidance about how far they could adapt the core aspects of the intervention. Variation was permitted, for example, in the order of grapheme–phoneme correspondences (GPCs) for the Direct Mapping aspect but not in the teaching of variable vowel and consonant strategies.

Findings from the focus group conducted after the pilot intervention training indicated that teachers accepted the idea of the Flexible Phonics intervention and its perceived value. Teachers broadly supported the co-expert model of intervention delivery; this shared professionalism and collaborative approach to the training was particularly appealing to some practitioners. Another pertinent finding mentioned by several teachers was the desire to form best-practice networks or communities of practice with colleagues from other schools to share ideas about implementing activities in the classroom when delivering the intervention. In response to these suggestions, for the main trial we created an online forum (discussed in Chapter 4) to facilitate information sharing between practitioners in different settings outside of the training. Also, the final segment of the training was adapted to give practitioners plenty of time to discuss their experiences of trying out aspects of the intervention.

On the whole, in the pilot, practitioners found the intervention training clear and appealing, and felt the programme was complementary to current teaching. They indicated professional interest and willingness to use this as part of their existing phonics practice. Some respondents

mentioned that they liked the use of key texts to consolidate GPCs, scripts for variability and the explicit teaching of new vocabulary and mispronunciation correction. Most practitioners felt the training was well pitched and that the elements of the programme were adaptable to their settings. The shared professionalism approach to the training appealed to the respondents, as demonstrated in these quotes:

> We are like experts for our classroom . . . I know this won't work for some, but it would work for others, it can be adapted . . . need to take it on board and adapt it where you see fit. I think it's been really good. Rather than just saying here's this, this will work, because then we are just consistently thinking that for these children it wouldn't work, but you've made it clear that you've understood that and made it clear that it's not one size fits all.
>
> I really like the idea of using the key texts and linking that with the sounds that you're learning. I think that's pretty good to immediately apply the sounds to real life contexts. I liked that one a lot.
>
> I think that it's good that it's an add-on rather than … changing your phonics approach. I think that works the most because it would be very difficult to change the approach partly through the year.

Feedback from the focus group about aspects of the training and resources that could be adapted and improved led to the production of instructional videos of other professionals demonstrating the intervention activities in context. Videos were developed for the main trial to demonstrate key activities; however, due to restrictions during the national lockdown, which made it impossible to enter classrooms to film videos, mock-up videos were instead created by the delivery team. Other feedback about the training highlighted the desire among some practitioners to have planning time built into the training to reflect on how to incorporate activities into the following week's lessons.

Other comments from practitioners in the pilot were related to reducing the amount of technical and academic vocabulary to explain key aspects of the intervention. The manual and training presentation were edited to take this into account and to include more 'plain English' explanations of the core aspects of the intervention, including a glossary to cover any technical information. More information about the pilot study can be found in Savage and Fox (n.d.) and in Chapter 3 of this book.

We had intended to gather the views of practitioners who had implemented the intervention for three months following the training. However, due to the national lockdown from March until June 2020, implemented in response to the Covid-19 pandemic, it was highly probable that staff were only able to implement about three to four weeks of the intervention prior to school closures. Questionnaire data from seven practitioners who had carried out the intervention for that short period indicated that they found the intervention clear, flexible and easy to implement in the classroom alongside their existing teaching. The delivery team would have liked to have visited schools to see elements of the intervention in practice, but restrictions meant this was not possible.

The content of the intervention training and manual was modified in light of the pilot feedback. Following that round of modifications, expressions of interest were also sought from experienced Reception (age four to five) teachers with knowledge of teaching Early Years reading, who were asked to review the amended manual and training materials. An advert was placed in the Education Endowment Foundation's research schools email newsletter. The intention was to gather expert advice about perceived barriers to implementing the intervention, considering the wider school system, and to gather additional feedback on the content of the manual. This was partly driven by findings from Bumbarger and Perkins (2008) and Forman et al. (2009), who suggested that interventions are less likely to be implemented when teachers perceive them to be in conflict with their school's current policies and existing initiatives, and when they are viewed as incompatible with available resources, such as staff availability and school budget.

Prior to the pandemic, the main trial plan was for face-to-face delivery of the intervention training (in January 2021), with follow up support via in-school support visits. However, due to the partial closures of schools from March to June 2020, added pandemic-related pressure on teaching staff of delivering online and face-to-face teaching to pupils, and increasing concerns about a second lockdown during the training phase, we chose to consult with schools to gauge their views about continuing to participate in the trial.

In June 2020, a questionnaire was sent to all schools (n = 123) who had signed up for the main trial, asking if they were still happy to continue to participate in the key aspects of the trial: autumn baseline reading assessments and winter intervention training, as originally planned (face to face). Out of 98 responses, schools were largely in favour of undertaking the baseline assessments as originally planned (93 per cent). Considering this, the baseline assessments, conducted by

the assessment team, Qa Research, largely continued as planned. For a small proportion of schools (17) who were not allowing external visitors to come onto the premises, the assessment team delivered assessments virtually over video-conferencing software.

In terms of the training format, an equal amount of respondents preferred online as compared with face to face, so given the uncertainty around another potential lockdown during the winter months, it was decided that training would be delivered remotely. From the responses to the questionnaire, we became aware of several queries about different aspects of the trial, particularly regarding uncertainty related to Covid-19. In July 2020, we invited all the schools who had signed up for the trial to attend an online webinar and Q&A session, where schools could pose questions anonymously to the delivery and evaluation teams about aspects of the training, reading assessments, intervention content and more. All schools were provided with an information pack, providing further details about the conduct of the trial and changes that had been made in response to the pandemic. The webinar was well attended, with over two hundred teaching staff viewing the hour-long session. Many queries were submitted, with common questions related to how the trial would run in light of the pandemic, how the reading assessments would work, how the intervention could fit alongside existing phonics teaching schemes and what adaptions needed to be made. These key queries were answered in full by the delivery and evaluation teams.

As mentioned in Chapter 4, the online training for the main trial was designed to be a collaborative process, with teachers bringing their expertise and ideas and sharing this knowledge through discussions in breakout rooms throughout the training. The collaborative learning benefits of using breakout rooms have been demonstrated in many studies and include deeper learning, empowering trainees, greater peer-to-peer learning support, greater retention of material and greater interactivity between trainees (Oakley, et al., 2004; Chandler, 2016). To allow schools more options to attend online training whilst juggling virtual teaching of their own classes, we offered multiple meeting dates.

The training also sparked reflection on existing practice within school teams. The quotes below highlight comments shared by teachers during support appointments.

> I feel very positive about Flexible Phonics and l like the ideas. I think some things, such as Set-for-Variability, I already do in my practice but in a more casual way. The training has made me place more emphasis on being explicit about alternative pronunciations and strategies.

> Flexible Phonics training has empowered all the staff, including support staff, to talk about the phonemes/GPCs and how they can be pronounced differently (for example 'th'), and how words are used in different contexts (for example 'to', 'two', 'too').

In response to requests from teachers, post-training online forums were set up to carry on these conversations, to support the exchange of ideas and support between schools and to provide best-practice case studies where aspects of the intervention had worked well in other settings. However, use of the forums remained limited despite a number of strategies by the delivery team (for example using facilitators and regular reminders when key information was added) to get these going. In addition, a monthly newsletter facilitated information sharing between schools. The aim of the monthly update was to showcase any new resources contributed by partner schools, share good practice, highlight any relevant articles and answer frequently occurring questions raised by schools in support meetings. Each edition was tailored to schools' needs and key themes from support partner visits. For example, the first edition contained articles to address some of the teaching-related concerns raised by schools at the training. The second edition focused on sharing good practice examples of Flexible Phonics in action and the Flexible Phonics philosophy – partnership working and the co-constructed approach. The third edition shared further examples of good practice and new resources.

A collaborative approach was taken to provide follow-up support to schools. Each school had a dedicated Flexible Phonics support partner, who provided tailored support to their group of schools via email and through prearranged online support appointments. Building trusted mentoring relationships between the support partner and practitioner, and ongoing feedback and coaching from partners, was intended to help practitioners apply the training to their settings. This type of approach has been found to increase teachers' self-efficacy, improving the implementation of the intervention (Klingner et al., 2013).

As mentioned in Chapter 4, in these online appointments support partners worked collaboratively with teaching staff to resolve implementation and delivery challenges using a solution-focused coaching model. This support was originally going to be offered from January until April 2021, but this was extended to provide ongoing support to those delivering the intervention later than initially intended. Due to staff sickness because of Covid-19, school disruption and partial closures, some schools only started delivering the intervention in March, two months after the training.

In a solution-focused approach, the coach (the support partner) directs the coachee (teaching staff) by asking questions to focus the coachees' attention on building solutions to problems and taking ownership of these solutions to move them forward. Common challenges which surfaced during these appointments included how to use the intervention to support children with English as an additional language, how to integrate the approach with existing phonics schemes taught in the school and how to fit the intervention into a busy school timetable. This support was received favourably by practitioners for its helpfulness in implementation and in sharing ideas and reflection, moral support and impetus, as highlighted below in a few comments by practitioners:

> The support has given us the chance to reflect upon our practices and our input of the programme within our individual classes. It has also been a time to share ideas together and get feedback on any new activities created. Having a mentor throughout this process has been fantastic, and more than anything, hearing from our mentor that we seem to be on the right track, and ideas for what more we can do, is reassuring.
>
> The training was done online, but it wasn't enough to understand the whole process. But the support team have been bridging that gap.
>
> It is really useful to know that there is an actual person to talk things through with if needed and that there are regular opportunities to do this. This also helps keep the momentum going.

The following are some representative comments reflecting the wider perceptions of educators about Flexible Phonics:

> I particularly like the versatility of the Flexible Phonics programme and how the teachers have agency in choosing what to implement and when.
>
> The real books have been great and to use a rich text over a decodable one – a text that the children already know, so that we can use the repeated language – has helped them identify themselves as readers. Flexible Phonics fits within the wider context of reading and helps with the development of a love of reading.
>
> Flexible Phonics has changed my practice in the teaching of exception words. I have now shifted the focus to tricky sounds rather than tricky words. This change has helped children become

more confident and this is reflected in their reading; they are now more motivated, as they feel like it is only one part of the word that is tricky, but they can read the rest.

We are using the real texts as an opportunity to do mispronunciation correction. We model sounding out an exception/tricky word and invite the children to correct us – and children love doing this. Parents have thanked us for using real texts in school and noted that children are more keen to join in with reading story books at home. We feel that by modelling mispronunciation correction, we have given children permission to 'have a go', make mistakes, and it doesn't matter if you sound a word out and it sounds funny – they can use [these] skills to work out what they mean.

Flexible Phonics fits really well with the new curriculum, which encourages the use of real and engaging texts with rich vocabulary across the curriculum in different subjects."

Overall, as illustrated by the comments above, the feedback collected during support appointments regarding schools' experiences with the intervention has been positive, particularly around implementing the intervention in the classroom and alongside existing schemes of work. The use of real texts in encouraging children to be more motivated readers has been particularly useful.

However, some challenges surfaced around integration with existing planning part way through the year, as the intervention's less prescriptive approach did not appeal to all. In some schools the implementation of a new intervention was challenging due to the ongoing impact of Covid-19 on staffing and workload, as evidenced by the following quotes by teachers, collected from support appointments:

I have struggled with planning, as Flexible Phonics is a huge change from a structured and prescriptive phonics programme, which I personally find easier to follow. That said, there are elements of Flexible Phonics that I totally agree on and understand are beneficial, but I struggle to know how to plan and where to start.

The Covid-19 pandemic has had a severe impact on our school, especially in Reception. Us teachers and our teaching assistants had Covid-19 and needed hospital treatment, and both teaching assistants are still absent from school due to ill health. For this reason, we have not been able to do nearly as much Flexible Phonics as we would have liked.

> We like the programme and the resources and support received throughout, but it was a bit difficult to implement it halfway through the year, especially as we have a high number of children with English as an additional language. We are willing to implement it more thoroughly next year.

Despite these considerable pandemic pressures on the schools, the trial attrition rate was minimal, with only three schools dropping out of the trial because of workload issues and long-term staff absences due to Covid-19. We believe this was partly due to the teams working flexibly in partnership with schools and adapting the trial delivery to school needs, keeping them abreast of the trial changes through clear and regular communication from the delivery and evaluation teams. Similarly, Dawson et al. (2018) found keeping in touch with schools regularly over the course of a trial is an effective way of minimising attrition.

Conclusions

In this chapter, we have outlined the key points in the Flexible Phonics trial where active collaboration with schools occurred. Key to this success has been flexible partnership working between the schools, the delivery team and the evaluation team. Listening to practitioners about what works in context, considering their needs and factoring these into the design of the intervention, management, training and follow-up support has helped the theoretical aspects of the intervention to 'translate into practice'.

The challenges of the pandemic highlighted the need for the trial to remain adaptable to the changing educational landscape, which included repeated school closures. These were met by pivoting face-to-face elements of the trial to virtual offers: training and post-training support. The pandemic brought about increasing demands on schools, in terms of increased workload, illnesses and long-term staff absences, alongside implementing and embedding the intervention. However, this chapter has highlighted how it was possible, despite these challenges and through strong partnership working, to successfully run an intervention with minimal attrition.

6
The formal process of evaluation

The overall aims of the evaluation were to measure the effect of the Flexible Phonics intervention on children's reading, to understand how Flexible Phonics was implemented in practice and to identify any enablers and barriers to successful delivery. In additional to this, a core aim of the EEF, who funded the research, was to understand how well the intervention worked for disadvantaged children (defined as those who receive free school meals). The Flexible Phonics evaluation was funded by the EEF as an efficacy trial – that is, to explore whether the intervention 'worked' to improve children's reading in a randomised controlled trial (RCT). The EEF requires that efficacy trials include the following three elements:

- An **impact evaluation** to measure whether the intervention makes a difference to children. For the Flexible Phonics trial, this was an evaluation of whether children's reading improved. This is usually reported in terms of number of months progress for the child
- An **implementation and process evaluation** to understand how the intervention is delivered in practice, whether schools find it easy to deliver and what factors enable or hinder successful delivery
- A **cost evaluation** to explore what the cost was to schools of delivering the intervention, reported as a cost per pupil.

This chapter will explain the main stages of the evaluation. This involved, firstly, the evaluation team applying to the EEF to evaluate the project and the design of the impact evaluation and the implementation and process evaluation in the proposal. Secondly, it involved the delivery of the evaluation, including:

- considerations when designing the initial evaluation protocol
- setting out the evaluation methodology in the protocol and statistical analysis plan documents
- the practicalities of preparing for the trial, such as ethics and data protection systems
- publishing the findings in a publicly accessible report on the EEF website.

This chapter will also discuss the role of partnership working from the evaluation perspective, including working with the delivery and funder teams, collaborating with our test administration partners to conduct remote testing in schools and, finally, working with the schools.

Evaluation stages

The application process

Once the EEF decides to fund a trial for an evaluation (discussed in Chapter 3), the process begins by matching the project with an evaluation partner. The EEF has a panel of evaluators who have been approved as having suitable expertise in methodology and subject matter to carry out EEF projects. Periodically, a list of potential projects together with a short summary of each project is sent round to this panel, so that they can decide which projects they are interested in evaluating.

There is then a two-stage application process. First, the evaluator sends an expression of interest to let the EEF know that they would like to evaluate a particular project on the list. This involves submitting a short summary of the proposed evaluation team and their experience of working on relevant research or evaluations. If the expression of interest is approved by the EEF, then the evaluator sends the EEF a full proposal for how they would carry out the evaluation. This full proposal typically includes a detailed description of the trial design, suggested outcomes and measures to be used, a description of the implementation and process evaluation methodology, a timeline, a list of risks to the project, ethical considerations and a budget. The Institute for Employment Studies (IES) was appointed as evaluator for the Flexible Phonics project in June 2019.

Proposal development

Impact evaluation

Detailed decisions around the design of the impact evaluation analysis and how to measure any outcomes are made at the very beginning of

the project, when the proposal is designed. This decision-making process considers a combination of factors, including drawing upon previous evidence and research for this intervention and other similar interventions. Looking at the research literature on the intervention in question and how it has previously been evaluated can be extremely helpful in making decisions around the design of the impact analysis and how to measure any outcomes. For example, one could look at past measures used to capture the primary and secondary outcomes, whether the measures were appropriate for the target group (whether the measures were easy or difficult to administer with the target group of 4–5-year-olds) and whether any impact was found in that study. It is also useful to know the effect size found for previous interventions, as this can help predict the likely effect size expected for the focal intervention in the current evaluation.

An effect size is the quantitative difference found between treatment and counterfactual control condition at post-intervention and is routinely used as a measure of the practical importance of any kind of intervention. The predicted effect size can be incorporated into calculations of sample size needed to fairly test for the existence of the expected effect size and directly informs recruitment targets. For example, if an effect size is likely to be very small (0.1 SDs), then a larger sample of participants would be needed to give the design enough power to detect the presence or absence of this effect. Box 6.1 illustrates how existing evidence was used to conduct the sample size calculations for the Flexible Phonics trial and explains the key terms and concepts which determine the likelihood that a trial will provide conclusive evidence of the effectiveness of the intervention.

Box 6.1: Conducting sample size calculations for the Flexible Phonics trial

At the time that the Flexible Phonics trial was designed, the expectation was that 100 schools would participate in the trial, with 50 schools randomly assigned to deliver the intervention and 50 schools allocated to the control group. Some schools have multiple classes for each year group, so it was decided to select one class at random from each school to participate in the trial.

The most recent figures on infant class sizes available at the time showed that there were an average of 27 pupils per class (DfE, 2018). It was assumed that around 15 per cent of pupils might either withdraw from the trial before it started or not participate in data collection for the full duration of the study. This was based on the rate of attrition observed in a

previous trial for an online literacy programme for Year 1 pupils (age five to six), called ABRACADABRA (McNally et al., 2016). With a similar rate of attrition in the Flexible Phonics trial, around 23 pupils on average from each school might participate. Therefore, a total of around 2,300 pupils were expected to take part in the trial, with approximately 1,150 assigned to the intervention group and 1,150 assigned to the control group.

The sample size calculations considered the likely minimum effect size – known as the minimum detectable effect size (MDES) – that the trial would be able to detect, given the expected number of trial participants. The MDES is expressed as a proportion of a standard deviation – a measure of the variation in the primary outcome between pupils. Accounting for the correlation between pre- and post-test performance for individual pupils and the extent to which schools explain the variation in pupil attainment – known as the intra-class correlation (ICC) – reduces the MDES. In the case of Flexible Phonics, there was little prior evidence on the likely correlation between pre- and post-test scores on the primary outcome measure – the Early Word Recognition subscale, one of the four subscales from the full York Assessment of Reading for Comprehension (YARC). However, another study which used the full YARC found that 55 per cent of the variation in post-test scores was explained by pupil characteristics and the pre-test score (Sibieta et al., 2016). Adjusting for the fact that the earlier study included pupil characteristics, it was estimated that the correlation between pre- and post-test scores in the Flexible Phonics trial might be around 0.4.

Evidence from previous evaluations focused on Early Years pupils aged three to five has shown that schools explain about 15 per cent of the variation in pupil attainment. For example, the efficacy trial of EasyPeasy (including pupils aged three to five years) found an ICC of 0.18 (Robinson-Smith et al., 2019), whilst the efficacy trial of Family Skills reported an ICC of 0.15 at class level (Husain et al., 2018). The ICC was therefore assumed to be 0.15 in the Flexible Phonics trial.

Based on the assumptions set out above, the MDES for the Flexible Phonics trial was estimated to be 0.23 standard deviations.[1] In practice, more schools participated in the trial than initially anticipated (118 completing, rather than 100), with 58 allocated to the intervention group and 60 to the control group. However, the number of pupils per class averaged 22, rather than 23. This meant that a total of 2,539 pupils participated in the trial, with 1,256 in the intervention group and 1,283 in the control group. The correlation between results in the pre- and post-intervention tests was also higher than expected at 0.54, rather than 0.4. However, the schools explained a slightly lower percentage of the

variation in pupil attainment than was anticipated based on previous studies (13 per cent rather than 15 per cent). Overall, this meant that the MDES for the final analysis sample in the Flexible Phonics trial was smaller than expected at the time the trial was designed, at 0.17 standard deviations, rather than 0.23. This increased the likelihood of detecting any impact from Flexible Phonics compared with initial expectations.

Past evidence can help inform the choices for randomisation design. For example, if contamination was a problem in a previous study because teachers in the same school shared materials from the intervention, a clustered approach – where schools, rather than teachers or classes, are randomised to the treatment or control groups – would be likely to reduce the risk of contamination. Past studies can also be a guide as to whether certain groups may benefit from an intervention more than others. For example, if access to resources for an intervention varies between areas, it may be advisable to use stratified randomisation to make sure that equal numbers of schools from a given area are assigned to the intervention and control groups, so the sample is not biased. However, existing evidence may not be available for all interventions.

Implementation and process evaluation

The aims of the implementation and process evaluation were fourfold:

1. To understand how schools delivered the Flexible Phonics programme in a real-world context
2. To explore if schools were implementing Flexible Phonics as the delivery team intended
3. To identify the factors that facilitated or enabled delivery and/or impact at the child level (including wider national/policy contexts)
4. To inform and contextualise the quantitative findings from the impact analysis.

The implementation and process evaluation used the Theory of Change developed during the intervention delivery evaluation analysis (IDEA) workshops to identify stakeholders (teachers, teaching assistants, pupils and other school staff), key delivery elements, and expected outputs and short-term outcomes to explore (discussed in Chapter 3).

The evaluation team drew on guidance from the EEF about using implementation and process evaluation as part of an RCT evaluation (Humphrey et al., 2016; EEF, 2019, 2022) to create a mixed-method design for the implementation and process evaluation. The implementation

and process evaluation is prespecified in the protocol. The EEF's (2022) implementation and process evaluation guidance specifies that, where relevant, the implementation and process evaluation should explore certain aspects or 'dimensions' as part of the implementation and process evaluation. For the Flexible Phonics implementation and process evaluation, we explored:

- fidelity
- dosage – do schools receive all intended training? How often and to what extent do participating teachers and teaching assistants deliver Flexible Phonics?
- quality – how well is the intervention delivered?
- reach
- responsiveness – do teachers/teaching assistants engage well?
- programme differentiation – how does the intervention enhance/differ from existing phonics teaching? And for whom?
- monitoring of control group
- adaptation.

One key point to note is that between June 2019 (when the IES was selected by the EEF to evaluate the Flexible Phonics efficacy trial) and the publication of the evaluation report in October 2023, the EEF revised its implementation and process evaluation handbook and guidance and its overall approach to the implementation and process evaluation element in RCTs that it funds. In particular, the EEF implementation and process evaluation quality pilot (Maxwell at al., 2021) was a study where an implementation and process evaluation quality measure was piloted. As part of this work, it was identified that there has been great variation in the quality of implementation and process evaluation research in EEF trials. The implementation and process evaluation quality measure looked at five aspects of implementation and process evaluation research – sufficiency of data sources, data collection methods, sampling, analysis and conduct – and rated these as high, medium or low quality.

In their review of 79 published EEF trials, they found that sampling methods were most likely to be classed as medium quality and analysis methods were most likely to be classed as low quality. A low rating could be awarded where no information was given about the analysis approach used, so it is possible that a high-quality approach was used in practice but this was not communicated clearly in reporting. Key points addressed in the subsequent EEF (2022) implementation and process evaluation guidance include:

- adding perceived impact, cost, context/moderators and mediators to implementation and process evaluation dimensions to be explored
- guidance on specifying analysis and sampling approaches
- guidance on reporting and interpreting qualitative findings
- a checklist specifying information to include in the implementation and process evaluation section of the trial protocol.

It is the evaluation team's experience that in recent trials, requirements for the implementation and process evaluation section of the trial protocols have become much more rigorous, including an external peer review. Previously, only the impact evaluation methodology was peer-reviewed, and the information required for the implementation and process evaluation section was much less detailed. However, the implementation and process evaluation section would be thoroughly reviewed by the team at EEF, and feedback would be given to improve or clarify the methodology. In their pilot, Maxwell and colleagues (2021) did find that the overall ratings for implementation and process evaluation quality increased over time from 2014 to 2019, which suggests that changes to EEF guidance were leading to improved quality in this area.

As described previously, the approach used by the evaluation team for the Flexible Phonics trial followed the EEF implementation and process evaluation Handbook (EEF, 2019) and already covered many of the elements added to the 2022 EEF implementation and process evaluation guidance, although the implementation and process evaluation section of the protocol was not peer-reviewed and included less explicit detail than would be required for current EEF protocols. The evaluation team believes that the implementation and process evaluation design and methodology was high quality and would meet the criteria for the current guidance; however, readers of the protocol and evaluation report could have benefited from more detail on the sampling and analysis approaches so that they could understand the full methodology used.

The discrepancy in rigour between impact evaluation and implementation and process evaluation elements is not unique to EEF trials and reflects an overall tendency in education RCTs. The CONSORT statement (Schulz et al., 2010), which provides recommendations for the reporting of RCTs, sets out very clear expectations around the reporting of quantitative research, but does not include recommendations for reporting implementation and process evaluations where this is relevant. Researchers across different fields (such as health and education) have conducted reviews and created recommendations around qualitative methodology and reporting, including for interviews and focus groups

(Tong et al., 2007), overall qualitative research (Johnson et al., 2020) and how qualitative methods must be appropriate to the context/purpose of the research (Yadav, 2022). The EEF implementation and process evaluation quality measure pilot (Maxwell et al., 2021) drew upon a conceptual framework for assessing the quality of mixed-methods studies set out by Tashakkori and Teddlie (2003). In future, we hope to see more of a standardised, rigorous approach to the implementation and process evaluation element of RCTs, consistent with that of the quantitative impact evaluation. Further to this, future developments in the use of artificial intelligence in qualitative research may make it more feasible to collect and analyse larger amounts of qualitative data to facilitate larger qualitative samples and richer insights.

As discussed earlier in this section, we developed research questions for the implementation and process evaluation in line with the relevant dimensions from the EEF's (2022) implementation and process evaluation guidance, and these are discussed, alongside the quantitative results of the study and the implementation and process evaluation, in Chapter 7. As part of the delivery pilot, the evaluation team observed a pilot of the intervention training – a day's staff training run by the delivery team. The delivery team also provided the evaluation team with two pilot reports to learn more about the intervention and how it had developed through the pilot. The evaluation team originally intended to observe two pilot training days and two pilot support sessions, but the outbreak of the Covid-19 pandemic in the UK meant that this was not possible. Information from the observation and the report informed further development of the Theory of Change model, as well as the process of designing research materials for the main trial, such as observation frameworks, interview discussion guides and survey questionnaires.

As first mentioned in Chapter 3, the planned implementation and process evaluation for the main delivery included two IDEA meetings to examine training and delivery materials and discuss the Template for Intervention Description and Replication (TIDieR) protocol, Theory of Change outcome measures and candidate measures of treatment integrity. Secondly, two in-person observations of training were carried out to assess engagement and for the team to learn more about the intervention. Case study visits to eight intervention schools towards the end of the programme were planned to observe teaching directly and to interview Reception (age four to five) teachers and teaching assistants, literacy or Early Years (up to age five) leads and a senior school leader. Interviews with the delivery team and analysis of other data potentially collected by them, such as attendance and cost data, were mooted. Finally,

a survey of teachers and other key school staff in intervention and control schools was proposed. The use of resources – such as the Ofsted reports on features of a good or outstanding Reception curriculum – and EEF guidance on improving literacy at Key Stage 1 (EEF, 2017) was envisaged to situate and formulate questions.

For the implementation and process evaluation analysis, qualitative and quantitative data from the case studies, interviews, surveys and delivery information would be combined in order to investigate the mechanisms of change set out in the Theory of Change model and inform impact analysis. The evaluation team planned to use the framework method, a spreadsheet-based qualitative analysis tool. Information from each interview or observation is entered as a separate row in the framework and represents one perspective. Before analysis begins, the framework is set up with themes drawing upon the implementation dimensions and the implementation and process evaluation research questions – that is, using a deductive approach. However, during analysis or data collection, evaluators may identify additional issues or questions that could contribute useful insights, so these are added to the framework using an inductive approach. The framework method allows for within-case analysis (looking in detail at each individual row) and between case analysis (comparing individual rows and groups of rows), as well as thematic analysis. This facilitates analysis at a setting level, to understand broader impacts and experiences within an individual setting context, and the comparison of individual experiences across a specific characteristic, such as role. It is designed to ensure a systematic and consistent treatment of all units of data (such as interview transcripts and observations).

This implementation and process evaluation work aimed to support the identification and understanding of the value-added impact of the intervention training, in combination with data on reading attainment, as an embedded mixed-methods approach. Such implementation and process evaluation analyses also potentially help identify the extent to which results may be explained by control schools improving phonics teaching (due to compensatory rivalry or other drivers) or by the displacement of other literacy activities in treatment schools (these are some of the key threats to RCT validity, as discussed in Chapter 1). Substantial changes were needed to the implementation and process evaluation plan due to the Covid-19 pandemic; these are summarised below, alongside the description of what took place within each part of the implementation and process evaluation (see Chapter 3 for details of changes to the intervention delivery).

The IDEA workshops

An initial IDEA workshop was delivered at the start of the project. An important part of the evaluation was determined at the initial IDEA workshop through discussion on appropriate outcome measures to assess the effects of interventions. Initial discussions shared a concern that a planned primary school baseline assessment by the Department for Education (DfE) might clash with, preclude or even replace the planned pre-testing of children. The need for sensitive, specific and psychometrically reliable and valid measures of attainment was emphasised by the evaluation team. It was ultimately agreed that a standardised psychometric measure of word reading (combining both regular and exception word reading subscales that could later be evaluated separately) was appropriate to pick up the anticipated effects of Set-for-Variability in Flexible Phonics on regular and exception words, as described in Chapter 2, as the primary outcome measure. The test was to be administered at pre- and post-test to assess change. The abandonment of government plans to undertake baseline assessment aided this decision significantly.

The YARC (Hulme et al., 2009) was selected to assess learning in the Flexible Phonics trial. This test has four sub-tests also measuring phonological abilities and letter–sound knowledge, along with isolated word reading. Following the content of Flexible Phonics, the second sub-test of the four in the YARC measure included in the statistical analysis plan, and which was tested at baseline, was letter–sound knowledge. This was one of the secondary outcomes of the trial. Reflecting the discussion in Chapter 1 about ways to establish causal links in RCTs, formal models of mediation (for example Kline, 2015) were planned by the evaluation team. Ultimately, a logical and statistical model of mediation of effects based upon using the oral Set-for-Variability measures developed by Kearns et al. (2016) was used in the statistical analysis plan to assess our theories of the sources of growth in word reading at post-test. The inclusion of this mediation analysis was something of a coup, as it was not commonly used. For budgetary reasons this analysis was, however, limited to a small subset of children in each school, given the Set-for-Variability measure ($n = 15$) in Kearns et al. (2016), and there was also not sufficient financial resources to assess Set-for-Variability at pre- and post-test, as a full mediation analysis should.

To evaluate the existence of more permanent causal developmental links post-intervention, formal delayed post-testing a year after the intervention finished was considered. There was, however, insufficient

budget to use the YARC in a formal delayed post-test. Instead, it was decided that secondary data from the phonics screener test, a mandated test already used and reported nationally in all English schools at the end of Year 1 (age five to six), would be selected. It was noted by the delivery team at the time that, as this test includes only regular and pseudowords, it is unlikely to pick up the most important distinct impacts of the Flexible Phonics intervention (that is, on exception words). Concerns about the construction and formal properties of this test (for example Darnell et al., 2017) were also shared by the delivery team with the EEF and the evaluation team. However, due to disruption caused by the Covid-19 pandemic to subsequent statutory testing, a decision was made, in consultation with the EEF, not to go ahead with this longitudinal analysis.

Secondly, a discussion of treatment integrity (fidelity) took place at the first IDEA workshop. The treatment fidelity measures were co-constructed based closely upon the Theory of Change model (shown in Chapter 3, Figure 3.1). There were five key components in the Flexible Phonics manual that were developed from autumn 2019 by the delivery team: Direct Mapping, oral vocabulary for the 66 most frequent exception words in children's books, oral Set-for-Variability, print-based mispronunciation correction, and continuous phonation – additional theory-driven and evidenced catch-up strategies for poor decoders. These five elements of Flexible Phonics shaped the broader content of the treatment integrity tool. The scaling used in the treatment integrity tool was based on successful models for assessing the fidelity of Reception-phase teachers delivering an intervention, as reported in Savage et al. (2013). All aspects of this work aimed to lay a solid foundation for the evaluation and to enable the evaluation team to tackle specific questions (such as an appropriate outcome, compliance measures etc.) informing the process evaluation. A review of the intervention materials was undertaken at this preliminary IDEA workshop stage to help us design the research tools and develop a good understanding of the intervention, which is especially important for the observations and interviews with teaching staff.

After the initial IDEA workshop at the start of the project, a further in-person IDEA workshop was planned for the end of the project to review the Theory of Change model and to capture any changes made over the course of the intervention to the design or expected outputs and outcomes. This second workshop was transformed into two shorter online workshops: the first was with Professor Robert Savage and the project manager alongside the IES team; the second also included some

of the support partners who had been delivering the support sessions with the intervention schools. This change meant that we got wider input on the model than would previously have been the case. It also allowed for further reflection time on key issues between the last two sessions, and enabled us to break up the session for accessibility reasons, as the workshops are quite long.

Observations

The aim of the observations was to gain a practical understanding of how the training and support sessions were delivered and how teachers and teaching assistants engaged with them. This included exploring variations across training cohorts, as well as providing an opportunity for the evaluation team to learn about the intervention itself. This information could be used to develop well-tailored research instruments (case study topic guides, post-intervention survey of teachers and teaching assistants) and provided contextual information to help when interviewing school staff, analysing qualitative and/or survey responses, or interpreting the findings of the impact analysis.

Observations of training and support partner visits were originally planned to take place in person. The training was moved online due to the pandemic, and this allowed us to access three training sessions remotely to get a thorough view of the different experiences of different groups of teachers and teaching assistants being trained. We were also able to join two support partner visits that were carried out remotely. While schools knew in advance of visits, it was slightly challenging to observe the support partner visits remotely, as participating teachers and teaching assistants had not met the observer/researcher in person to build rapport as they would have during an in-person observation. Teachers and teaching assistants were asked whether they preferred the observer to leave their camera on or off, which meant that for some observations the observer's face was on screen and they were a silent partner in the conversation, whereas during an in-person observation, the observer would usually sit a little out of the way or at the back, as appropriate, after they had introduced themself and explained the aim of the research.

Teacher and teacher assistant surveys

The implementation and process evaluation included surveys with teachers and teaching assistants who were delivering Flexible Phonics at the start of the trial (baseline) and towards the end of delivery (endline). The aim of the surveys was to understand what current phonics teaching practice looked like before Flexible Phonics was introduced (business-as-usual) and to explore evidence of any change in teacher and teaching assistant practice or behaviour and self-reported knowledge or confidence around phonics over the trial period. This information helps to identify the extent to which results may be explained by control schools improving phonics teaching (due to compensatory rivalry or other drivers) or by the Flexible Phonics intervention displacing some literacy activities in treatment schools. For teachers and teaching assistants who delivered the intervention, the endline survey also explored their experience and views on delivering Flexible Phonics in their school, their views on the training sessions and support sessions, and their perceptions of any impact on their pupils.

The surveys were completed online, and teachers and teaching assistants received a personalised link to the survey via email. The baseline survey took place in October, before schools were randomised to the treatment and control groups, and aimed to capture information about their normal approach to teaching phonics and their broader approaches to teaching reading, phonics and spelling. The evaluation team used resources such as the Ofsted criteria for the features of a 'good' and 'outstanding' Reception curriculum, which include characteristics of strong phonics teaching (Ofsted, 2017), and EEF guidance on improving literacy at Key Stage 1 (ages five to seven) (EEF, 2017) to formulate questions.

The endline survey repeated all questions about phonics teaching and, for treatment schools, covered experiences of taking part and the staff time and resources required to undertake the intervention (to inform the cost-per-pupil estimate). It included questions on adaptations made to the programme (beyond expected differentiation to meet the needs of individual pupils) and about schools' participation in another project which was rolled out in about half of both the treatment and control schools at the same time as Flexible Phonics. This was the Nuffield Early Language Intervention (NELI), a small-group, targeted intervention for improving spoken language, listening skills and vocabulary. The NELI project had previously been evaluated by the EEF and, after being shown to be effective, was rolled out by the DfE as part of the Covid-19 recovery programme from the 2020–1 academic year

onwards. Therefore, it was important to understand how this may have affected how staff supported children's language development (if they were running both projects at the same time). The surveys were sent to all teachers and teaching assistants of participating classes (not just the classes tested as part of the impact evaluation) and approximately 613 teaching staff were invited to take part in both the baseline and the endline surveys.

Case study visits

The aims of the case studies were to explore in greater depth the teaching staff's experiences of the intervention in terms of the feasibility and acceptability of the different elements (such as training, delivery in schools, support partner visits and using the online platform), the enablers and challenges in delivering the intervention, and the perceived impacts on all these different stakeholders. The original evaluation plan was to conduct eight case study visits in person, which included an observation of teaching practice and interviews with teachers and teaching assistants and senior leaders. The pandemic context of the trial delivery meant that case studies took place via virtual or telephone interview instead, and the observations of teaching practice had to be dropped entirely. This meant that we were reliant on information from the support partners about their virtual support meetings and any challenges or enablers discussed in these. As teachers and teaching assistants may not have been comfortable discussing the full reality of delivering the intervention with the support partners, this means that the information collected may only reveal a partial representation of their experiences. The opportunity to have informal discussions with setting staff between interviews, and to observe the context of delivery in person and the setting environment, was also missing from the case studies in this instance. However, the response to interviews was positive and we managed to achieve good coverage of interviewees (a wide range of perspectives from across a robust number of schools) by being flexible with interview timings and working to rearrange interviews where necessary. Interviewees were also dealing with the challenges of working in restricted groups (Covid-19 'bubbles') once schools were fully open again, which sometimes meant last-minute changes were needed to planned meeting structures.

Interviews with schools that withdrew

The implementation and process evaluation also included provision for up to nine interviews with schools that withdrew from the study once the trial had started. These schools were still free to keep delivering Flexible Phonics if they wanted, but children from these schools would not take part in any further reading assessments and staff would not be invited to take part in interviews or surveys. When a school withdrew from the intervention delivery and/or collection of child assessment data, UCL would inform the evaluation team, so that the school could be invited to participate in either a short telephone interview or answer questions on an email form. The aim of these interviews was to explore the reasons why the school withdrew and whether this was related to any aspects of the Flexible Phonics intervention. At the start of the trial, the evaluation team anticipated that quite a few schools might drop out of the study because of the challenges of delivering education during the Covid-19 pandemic. However, after the baseline testing had taken place with children, only three schools dropped out. The evaluation team was able to conduct interviews with staff at two of these schools and their feedback was incorporated into the qualitative analysis with the feedback from case study schools.

Interviews with the delivery team

Interviews with the team delivering Flexible Phonics training and support were planned to take place towards the end of the delivery period. All training and support visits were expected to be completed by the end of March 2021. However, as schools in England were partially closed from 5 January to 8 March of that year, delivering in-person teaching only to the children of key workers and vulnerable children, contracts for some of the support partners were extended to offer support to schools for a longer period, once all children were back to in-person delivery. For this reason, the support partners were interviewed at three points to capture their experiences before their contract ended: late March, late May and late June. The evaluation team spoke to four support partners in total. In late June, the evaluation team conducted interviews with Professor Savage, the project director and the project manager. These interviews with the delivery team were intended to explore their experiences and reflections around delivering the Flexible Phonics training, school engagement and participation, and enablers and/or barriers to successful implementation of the programme.

Cost evaluation

The primary aim of the cost evaluation is to provide an estimation of the cost per pupil to schools of delivering the Flexible Phonics intervention, as per the EEF's (2023) cost evaluation guidance. This is intended to allow schools to evaluate whether a particular intervention would be good value for money for their needs. In order to collect information about the cost of delivering the Flexible Phonics programme, school staff were asked about costs and time needed to attend training and support sessions and to plan and deliver the programme.

Questions regarding these aspects were included in the teacher and teaching assistant surveys and the case study interviews. Questions around costs were also included in the interviews with the delivery team, so that the evaluation team could estimate what price might be charged for the Flexible Phonics intervention if schools needed to buy it as a product. During the trial, the Flexible Phonics intervention was free to schools, as the delivery team were being funded by the EEF; furthermore, as the delivery team were based at a university, they did not have separate administration costs for the online conferencing software, online learning platform or other resources.

Collecting these data was difficult, as the response to the cost and time questions in the surveys was low. Only a few of the staff interviewed had reported any costs and there was wide variation in the amount of time individual staff had spent preparing for and delivering the intervention. This was partly because the intervention appeared to be very low cost and generally involved using existing resources or the books that were provided free of charge to the intervention schools. However, the evaluation team identified measures that could help improve data collection in future, such as redesigning questions to be more specific and including an option to indicate that there were no costs. One adaptation that the evaluation team was able to make during this project was to design a form for the delivery team to complete which collected information about their costs and time spent delivering the training and providing support to schools.

Protocol development

After the evaluation team was appointed, two in-person set-up meetings were arranged with the delivery team, the evaluation team and the EEF to discuss the project in close detail. These meetings included discussions

on the primary outcome to be explored by the project and how to measure it. Also discussed were the plans for the impact evaluation, measuring the impact of Flexible Phonics on children's reading; for the implementation and process evaluation, exploring how schools delivered the intervention in practice; and for the cost evaluation, determining the cost effectiveness of the intervention based on the cost per child over three years using specific calculations in the EEF's guidance (which gives the project a cost rating score). These meetings also provided an opportunity to start building the relationship between the evaluation and delivery teams.

The set-up meetings were followed by the IDEA workshop, described in detail in Chapter 3, which is where the TIDieR framework and the Theory of Change model were developed. Once more details of the proposed evaluation had been discussed in detail through these meetings, the evaluation team put together a comprehensive protocol which thoroughly described the decisions that were agreed on by the teams on a variety of areas, including:

- rationale and background for the intervention
- recruitment – inclusion and exclusion criteria
- research questions
- randomisation – unit and trial design
- sample size
- outcome measures
- a brief description of the analysis planned
- the implementation and process evaluation methods and research questions
- cost evaluation plans
- ethics and data protection
- the personnel forming part of the delivery and evaluation teams and their roles
- risks
- the timeline.

The EEF has a trial protocol template they require all evaluation teams to use to ensure that the process is standardised across projects.[2] The protocol was reviewed by the EEF and the delivery team, and then published on the EEF website ahead of the trial for full transparency.[3] If any areas of the protocol need to be changed, then a new version is developed by the evaluation team, including a table at the front with the changes briefly described and an appendix with all the details of the changes clearly outlined. The new version is reviewed again and then

published on the EEF website alongside the original. Due to changes to the intervention and evaluation caused by the Covid-19 pandemic, a second version of the protocol was published in June 2021. In September 2020, before the baseline testing began, the protocol was also registered with an independent trial depository called the International Standard Randomised Controlled Trial Number (ISRCTN) registry, which contains the details of international trials in all different subject areas. The Flexible Phonics evaluation is registered with the study ID ISRCTN18428598.[4] This record could also be updated as needed during the trial and these updates are summarised in a log of changes. After the project is completed the results of the trial are also added to this record.

Data management and ethics

After the protocol has been developed, a process of data management is necessary in all trials. The Data Protection Act (2018) is the UK's implementation of the EU General Data Protection Regulation (GDPR) and outlines key principles and responsibilities around processing personal data. As part of the Flexible Phonics trial, the evaluation team needed to collect personal information about both children and teachers and teaching assistants. This included personal characteristics, such as children's date of birth or gender (to analyse differences by age and gender), and personal identifying or contact data, such as teachers' and teaching assistants' names and email addresses (so we could invite them to participate in the surveys). In the case of children's data, the DfE considers gender to be special-category data and so there are further requirements for processing this. A core principle of the Data Protection Act and the GDPR is to collect and/or process only the personal data that are needed for the required task. Care was thus taken to identify the personal data necessary for the evaluation versus data that were not needed. In addition to this, the children's data, such as reading assessment scores and some personal characteristics, needed to be linked to the National Pupil Database (NPD) for the main analysis. Finally, while there is a requirement to treat all personal data securely, particular care is needed when processing, storing or transferring children's data, as this is a vulnerable group.

As personal data were collected and shared amongst different organisations, a detailed data-sharing plan was drawn up between the IES, the EEF and the delivery team, detailing which data were being shared with which organisations at which point, why it was necessary to

collect those data, how the data were being stored and how long for – all following GDPR guidelines. One important aspect of this was establishing the role of each of the organisations with regards to the data and their basis for processing the data. In the case of the evaluator, the basis was legitimate interests, as the evaluation team needed to process the data in order to evaluate the Flexible Phonics programme and conduct the RCT as they had been funded to do by the EEF. The evaluation team was deemed to be a joint controller of the data alongside the delivery team, but once the research was completed and the data was transferred to the EEF for archiving, the EEF then became the sole data owner. This arrangement was different to a lot of the projects that the evaluation team was used to working on; the evaluation team typically acts as a data processor, as it is conducting the research on behalf of another organisation, such as a government department or another client. As a joint data controller, the evaluation team was responsible for ensuring that the data was processed, stored, transferred and collected appropriately – that is, securely. The evaluation team would also be liable for any legal consequences if these requirements were not fulfilled.

Once the recruitment materials were finalised in January 2020, the IES applied for ethics approval through the internal IES Ethics Committee. This included a review by two senior staff within the IES and the chair of the ethics board, who is a member of the IES's management team. Firstly, headteachers, senior leaders or Early Years leaders received a flyer about the project. If they were interested, they received an information pack containing further details about the project and then signed a Memorandum of Understanding (MOU) to agree to take part in the project. The MOU detailed the responsibilities of all parties for the trial and data sharing. A project lead was designated in the MOU; this person then received further communications on the project as it progressed. Schools were also invited to webinars in July 2020 with the teams from UCL, the IES and Qa Research to hear more about the project and, in particular, changes to the protocol in light of Covid-19. The webinars also provided an opportunity for the teams to answer any questions raised by school staff, including those about how testing was to be undertaken.

Parents received an information sheet detailing the trial and information about data sharing. They had the chance to withdraw their child's data from being shared with the evaluation team at the start of the trial by telling the school. An accompanying letter contained a link to a privacy notice that explained how the data were being used, stored, shared and deleted. After randomisation, the parents could contact

the IES directly to have their child's data removed from the data stored by the evaluation team. There was also a privacy notice for school staff, explaining how the information collected was used and stored and communicating to participants their right to withdraw from data processing. This was available online, with the link provided in school and teacher letters and included in email briefings before staff took part in the surveys and interviews.

Teachers and teaching assistants had the chance to opt out of taking part in the implementation and process evaluation when they received the invitation to complete the survey or take part in an interview. The invitation contained information on what the research would contain and data sharing arrangements, as well as a link to the privacy notice. The delivery team applied for ethics approval through UCL's IOE (Faculty of Education and Society) for delivery of the trial, which was approved separately.

The statistical analysis plan

The next step in evaluation design is writing a detailed statistical analysis plan, which has many similarities to the protocol process. There is also an EEF template for the statistical analysis plan and the EEF require all evaluators to follow their analysis guidelines in drafting the document.[5] This is to aid comparability between EEF trials and ensure that there is consistency in the approach to the analysis.

The statistical analysis plan builds on the analysis section in the protocol and is a detailed document which ensures that decisions about how to deal with different scenarios are agreed before analysis commences. The main elements of the statistical analysis plan are the following:

- **Description of the intervention and the trial.** The statistical analysis plan provides a detailed description of the intervention to be tested in the trial and sets out the research questions. It identifies one or two primary research questions, which are the main focus of the study, and any secondary research questions which will also be considered. It also details key features of the trial design, such as the setting where the intervention will be tested, the approach to randomisation and the main methods of analysis.
- **Sample size calculations.** The statistical analysis plan includes the sample size calculations at the time the trial protocol was drafted, alongside those produced at the time of randomisation, based on the

units actually randomised to the intervention and control groups. It reports the MDES at each stage and the evidence underlying the assumptions made in the sample size calculations.

- **Primary and secondary analysis plans.** The statistical analysis plan describes the approach to the analysis of the primary and secondary outcome measures which will be used to answer the research questions. This includes describing the source of the outcome measures and explaining how they will be constructed. The plan includes details of the measure of prior attainment to be included in the analysis and sets out how the analysis will take into account the clustering of pupils within schools, for example through multi-level modelling. It also provides details of the software to be used to conduct the analysis and how the ICC will be calculated.
- **Subgroup analysis.** There may be an interest in knowing whether the effectiveness of an intervention varies between particular subgroups of pupils or types of school. If this is the case, the statistical analysis plan may set out plans to test whether the intervention has a stronger impact on the primary outcome for certain subgroups relative to others.
- **Longitudinal analysis.** As well as detailing plans for the initial analysis of impact, the statistical analysis plan may describe any future plans for analysis of secondary outcomes to see whether any impact from the intervention emerges sometime later or is sustained over time.
- **Imbalance at baseline.** If the composition of the intervention and control groups differs at baseline across key characteristics, outcomes for the control group may not be representative of the outcomes that would have been attained in the absence of the intervention. This could undermine the credibility of the impact estimate. The statistical analysis plan sets out the approach to assessing whether the characteristics of each trial arm are balanced at baseline and whether the impact estimates are likely to be biased.
- **Missing data.** The statistical analysis plan details how the completeness of the data used in the trial will be assessed. It may also describe the approach that will be used to impute missing data to explore whether the main findings of the analysis are likely to be affected by missing data.
- **Compliance.** Although the main analysis in trials is based on intention to treat (all enrolled schools are assumed to be running the intervention), the statistical analysis plan may also set out any plans to assess the impact of the intervention when it is implemented as

intended. This provides an indication of whether the intervention is effective in ideal circumstances. The statistical analysis plan explains what key features of the intervention must be implemented in order for delivery to be assessed as compliant. This could include distinguishing between elements of the intervention which are considered optional and those that are mandatory. The statistical analysis plan will also detail how compliance will be assessed and recorded and describe the methods which will be used to carry out the analysis, including the software to be used.

- **Effect size calculation.** To aid comparability with other trials, the statistical analysis plan sets out the approach to calculating the effect size given the nature of the primary outcome measure, such as whether it is continuous, ordinal or binary. It also explains how the statistical significance of the impact estimates will be assessed.
- **Sensitivity analysis.** There may be value in carrying out additional analyses to assess the robustness of the main findings. For example, if the analysis of imbalance at baseline suggests that there were substantial differences between the intervention and control groups prior to the intervention, the sensitivity analysis might explore whether including controls for these differences changed the main findings of the analysis. The statistical analysis plan provides details of any other analysis which may be used to understand the findings better, including using alternative measures of prior attainment, exploring different aspects of the approach to implementation or testing, and investigating the links between different facets of the intervention and particular outcomes.

The statistical analysis plan is reviewed by the EEF and the delivery team, so that all parties agree that the statistical analysis is appropriate to identify any impact from the intervention. This ensures that there is acceptance of the results if a trial finds no effect from the intervention. The completed statistical analysis plan is published on the EEF website and can be updated if there are any substantial changes to the plan, with the amendments highlighted so that it is clear what has changed.[6] Publishing the statistical analysis plan before the evaluator has access to outcomes data from the trial means that there is transparency over the planned analysis. This is important in order to demonstrate that the approach has not been changed in light of the findings in an attempt to demonstrate the effectiveness of the intervention.

Collecting outcome data

For this trial we worked alongside Qa Research (an external subcontractor), which collected data on the primary and secondary outcomes – measuring reading using the YARC Test full score[7] (Hulme et al., 2009) and Set-for-Variability using items selected from the Mispronunciation Communication Test developed by Tunmer and Chapman (2012). Data were collected at two time points: at the start of the trial, before randomisation in autumn 2020, and then at the end of the school year in June–July 2021. Qa Research collected the data using trained test administrators who were recruited from universities. The training of the administrators involved one of the reading test developers going through both of the tests in detail, explaining how to deliver it with children and how the responses should be scored. The data collection period at each time point was about six weeks, which was a challenge for the team, as there was a great deal of school disruption at this time and much rescheduling was thus required. Baseline testing was particularly challenging due to the pandemic restrictions.

Baseline testing took place in autumn 2020, when some schools were still very reluctant to allow external visitors, so 17 schools opted to utilise remote testing instead. The process for this involved someone at the school (often a teaching assistant) accompanying the child to a quiet area, where a laptop was set up with a virtual meeting with one of the test administrators. The assessment took place using video-conferencing software. Both the test administrator and the child had their camera and microphone on, and the test administrator could show materials on-screen when needed. A teaching assistant was always present with the children but they sat behind the child and were given instructions by the assessor not to interfere with the testing process or prompt the child. Assessors were briefed on strategies for responding appropriately if a staff member was interfering in the test, and the test administrators checked in with each assessor after testing to monitor for any interference from school staff. No staff disturbances were reported. This process was piloted by the test administrators with a few children of the same age at schools that were not taking part in the trial. The test administrators prepared instructions to send to schools in advance of the remote testing.

Report publication

Finally, at the end of the trial, the IES produced an evaluation report which was published on the EEF website.[8] This report follows an EEF

template and incorporates a detailed discussion of both the impact and the implementation and process evaluation method and results, alongside the cost evaluation. The reporting process takes approximately five months in most circumstances, from delivery of the first draft of the report to publication of the final version. The report goes through various stages of review, including review by EEF evaluation team members (to ensure that EEF guidance has been followed and that it relates to the protocol and statistical analysis plan), anonymous peer review by two external peer reviewers who work in relevant methodological or subject areas, and review by the delivery team (for any inconsistencies in understanding of the intervention and its delivery or for additional detail to be added where appropriate). The report must follow the protocol and statistical analysis plan that were outlined at the start of the project; in the event of any deviations, these must be clearly described, along with reasons for their occurrence.

This reporting process is led by the EEF which, at the end of reporting, gives the evaluation a rating indicating the robustness of the trial. This is known as a padlock rating and it is out of a possible five (as the maximum rating). The padlock rating looks at evaluation design (whether a project is an RCT or quasi-experimental design), the minimal detectable effect that the trial was powered to detect through the sample size calculations, and the attrition over the course of the trial on the primary outcome. Finally, it examines seven threats to validity, which are:

1. confounding effects – baseline differences
2. concurrent interventions
3. experimental effects, for example risk of contamination
4. implementation fidelity – whether the intervention was delivered as planned
5. missing data from the final analysis
6. measurement of outcomes, where these are appropriate for the sample and delivered robustly
7. selective reporting – whether there was evidence of findings being hidden.

This rating examines the primary outcome measure for all children who take part, including the subgroup of children who receive free school meals, as one of the EEF's main priorities is to establish if interventions are able to help 'close the gap' between disadvantaged children and their more advantaged peers. The EEF also prepare a press release for the project when the report is published on their website.

Partnership working

One of the unexpected outcomes of the pandemic was the positive impact on partnership working that was experienced. We worked closely with the delivery team (covered in more detail in Chapter 3), the EEF and the independent test administrators, Qa Research, to navigate the challenges of the pandemic together. We believe this forged more collaborative relationships, which enabled us to approach the challenges together whilst remaining independent and objective in our work on the evaluation. The relationship with schools was also critical to being able to carry on the research given the pandemic circumstances that they were facing.

Working with the delivery team and the EEF

These relationships were developed during the pandemic by regular in-depth meetings to discuss how the delivery and evaluation needed to adapt to changes and to actively problem-solve together to find a route forward that was suitable for schools and would reduce the burden on them as much as possible. Working on the Theory of Change model and TIDieR documents together helped build detailed understanding of the intervention among the evaluators, and also raised questions that the delivery team could consider when adapting to school and pandemic needs. The input of the delivery team and the EEF at each step of the evaluation process (including the IDEA workshops, protocol, statistical analysis plan and reporting) meant that they were given a voice and could reflect on their own experiences too as the project progressed.

Sometimes there were conflicting ideas amongst the three teams about the best ways to approach areas of research design, such as testing, or the types of analysis to be used. These were approached collaboratively and in open discussion. The EEF were very responsive to requests from the delivery and evaluation teams, but on occasion factors such as budgetary constraints meant that some ideas had to be scaled back. For example, it was decided to test only a subsample of the children on the secondary outcome (Set-for-Variability) rather than the whole sample of children, as the latter would have proved very time- and resource-intensive as well as placing extra burden on the schools.

Relationship with the test administrators

The evaluation team had previously worked with Qa Research on another trial for the EEF, called the Tips by Text trial (see Stokes et al., 2022). This was a parenting intervention for parents of children in Reception, focusing on giving tips to improve literacy, numeracy and socioemotional development. A strong relationship had thus been developed between the evaluation and test administration teams, as the Tips by Text trial had also run through the earlier stages of the pandemic (the trial ran from autumn 2019 to winter 2020). Close, regular communication was key to ensure that we could react quickly to adaptations needed for the testing process (such as remote testing, as described above). Very frequent contact was needed during the lead-up to testing and during the testing periods, so that all the teams were aware of the numbers of schools being tested and the delivery team could nudge less responsive schools to take part.

Relationships with schools

The pandemic also impacted the evaluation team's and delivery team's relationships with schools. Based on our experience in the Tips by Text trial, we expected to have larger than usual attrition at both testing points (baseline and endline) and for more schools to opt out of carrying on with the intervention. In the Tips by Text trial, approximately 40 per cent attrition was seen at endpoint testing, which took place in autumn 2020 (having been moved from summer 2020). However, the pupil-level attrition rate was only 20 per cent for the Flexible Phonics project.

We feel this reflects the resilience of schools and the dedication of teachers and teaching assistants to ensuring that pupils were offered the best chance to be supported with additional interventions following the lost learning due to the pandemic. The Tips by Text trial was more badly impacted by the initial response to the pandemic, having run mostly in 2020, whereas by 2021 teachers and teaching assistants were more able to be flexible in their response. Aside from adapting testing as described, we also ensured that we were as flexible as possible with schools when arranging interviews. This involved rearranging the majority of interviews, which increased the amount of time it took to arrange interviews.

Conclusions

The aim of this chapter was to describe the principled manner in which evaluation was undertaken in the Flexible Phonics trial. As the Flexible Phonics evaluation was funded by the EEF as an efficacy trial to explore whether the intervention improved children's reading in an RCT, three elements were carried out: an impact evaluation, an implementation and process evaluation and a cost evaluation.

Our approach to evaluation – while impacted by the Covid-19 pandemic in multiple ways, especially in terms of the delivery modes available – nevertheless reflected much of the broad methodological approach of more typical EEF-funded trials in other respects. It yielded interpretable data on how Flexible Phonics was implemented in practice and identified a number of key enablers and barriers to successful delivery. The likely cost was identified. The impact evaluation measured whether the intervention made a difference to children. For the Flexible Phonics trial, this was whether children's reading improved. A description of the results of the impact evaluation, and fuller descriptions of the implementation and process evaluation and cost evaluation, is the subject of the next chapter.

Notes

1 This also uses standard assumptions of 80 per cent power to detect effects and a 5 per cent significance level.
2 Available here: https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/protocol-study-plan-and-sap-templates
3 The evaluation protocol for the Flexible Phonics trial is available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/pages/projects/EEF_trial_protocol_flexible_phonics_updated_final.pdf?v=1701427207
4 This can be accessed at: http://www.isrctn.com/ISRCTN18428598
5 Available here: https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design
6 The statistical analysis plan for the Flexible Phonics trial is available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/pages/projects/20210609_EEF_SAP_Flexible__final.pdf?v=1701427207
7 Available from GL Assessment at: https://www.gl-assessment.co.uk/assessments/products/yarc/
8 https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/flexible-phonics

7
Evaluation findings

This chapter presents the main findings of the study. It sets out the main questions that the research sought to address and how the evidence produced in the evaluation answers them. It considers whether the intervention improved children's reading outcomes and whether children learnt to use the variability strategies taught in the Flexible Phonics approach. This involved drawing on the findings from the impact evaluation, which sought to provide answers to the research questions. The chapter also explores how the intervention was delivered and the extent to which it was delivered as intended, based on evidence from the implementation and process evaluation. These findings are linked to the themes set out previously and the wider literature.

The research questions

It is common for impact evaluations to seek to address a wide range of different questions and to explore the impact of the intervention on more than one outcome. However, as mentioned earlier, it is good practice to be clear about the primary objective of the intervention before starting data collection and analysis (Haynes et al., 2012; Torgerson & Torgerson, 2016).[1] This is to avoid any suggestion that the focus of the research has changed to suit what the data show.

In the current study, early workshops and discussions on the Theory of Change for Flexible Phonics, attended by both the delivery and evaluation teams, were used to agree on a set of research questions. This included identifying a single primary research question to be the main focus of the study. The agreed primary research question, RQ1, was as follows: 'Does the Flexible Phonics intervention improve Reception children's word reading ability as measured by the York Assessment of Reading for Comprehension (YARC) Early Word Recognition subscale?'

This was chosen as the focus of the study because the delivery and evaluation teams agreed that if the intervention did not have a discernible impact on word reading ability, it would be difficult to conclude that Flexible Phonics was effective.

After researching suitable ways of measuring word reading ability, performance on the YARC Early Word Recognition subscale was chosen as the preferred primary outcome measure. This decision considered existing evidence that the measure was appropriate to capture word reading ability among Reception-age children (age four to five) and the likelihood that it would be affected by an improvement in the skills that Flexible Phonics sought to impart.

The YARC assessment is suitable for use with children aged four to seven years (that is, starting with Reception year) and has been standardised with a UK cohort of 662 children (Hulme et al., 2009). As part of this process, each subscale was assessed for internal consistency using the Cronbach's Alpha measure.[2] For example, reliability for the Early Word Recognition subscale was 0.98. The YARC assessment is also widely used in research studies around Early Years reading (Colenbrander at al., 2017). As described in Chapter 2, theory and empirical research on Set-for-Variability allow us to predict that Set-for-Variability instruction should affect word reading acquisition for irregular words that contain graphemes that are inconsistent with standard phonic rules and may, according to some theories, also aid regular word blending. Direct Mapping should facilitate the acquisition of grapheme–phoneme correspondences (GPCs), which should then facilitate the reading of words.

In addition to the primary research question, a series of secondary research questions, largely driven by the theory and research described in chapters 2 and 3, were also agreed upon by the delivery and evaluation teams. These had several different objectives, including:

- whether Flexible Phonics had a discernible impact on measures of literacy, including the full YARC Early Word Reading scale (consisting of four subscales, including the Early Word Recognition subscale, which was used as the primary outcome measure; letter–sound knowledge; sound deletion; and sound isolation subscales). The YARC focuses primarily on letter knowledge and phonological awareness, so we would expect this to be most directly impacted by Direct Mapping and less directly by Set-for-Variability, which would be expected to impact on the irregular words included in the Early Word Recognition subscale. The mispronunciation correction test is

an oral language task designed to capture the impact of the Set-for-Variability strategy and was used to identify whether children were using this strategy (Tunmer & Chapman, 2012).
- whether there was any evidence that either the Direct Mapping or Set-for-Variability strategies had differential impacts
- whether the impact of Flexible Phonics differed between schools and pupils with different characteristics
- whether any impact that Flexible Phonics had on phonics skills was sustained over time.

The secondary research questions were thus formalised as follows:

- RQ2: Does the Flexible Phonics intervention improve Reception children's literacy outcomes, as measured by more general literacy tests?
- RQ3: What is the differential impact of Direct Mapping and Set-for-Variability skills on children's word reading ability?
- RQ4: Does the Flexible Phonics intervention provide value-added improvement to Reception children's word reading ability compared to good phonics teaching alone in schools identified with good phonics practice?
- RQ5: Does the Flexible Phonics intervention improve word reading ability differentially for children eligible for free school meals?
- RQ6: Does the Flexible Phonics intervention improve word reading ability differentially for children of low ability?
- RQ7: Does the Flexible Phonics intervention improve Reception children's phonics skills one year later at the end of Year 1?
- RQ8: Does the impact of the Flexible Phonics intervention differ depending on whether the school was participating in the Nuffield Early Language Intervention (NELI)?

In relation to the question of whether Flexible Phonics had a sustained impact on phonics skills (RQ7), the trial protocol noted that there was uncertainty over whether the results of the Year 1 phonics screening check (when all children will have turned 6), the measure used to assess whether impacts were sustained, would be affected by Flexible Phonics. This was because the phonics screening check measures the decoding of regular words, whereas the intervention focuses on words that do not adhere to phonics rules. However, as all state-funded schools are required to administer the test at the end of Year 1, no additional data collection was required and so it was thought to offer the most viable and

cost-effective way of exploring any potential longer-term impacts from the programme. In practice, Flexible Phonics did not have a clear impact on any of the primary or secondary outcome measures. It was therefore agreed between the evaluation team and the EEF that the additional work involved in gaining access to the phonics screening check data and conducting the analysis was not justified based on the findings at the end of the Reception year.

Rather than being based on existing theory, RQ4 was included because the EEF wished to explore whether Flexible Phonics had an impact on word reading ability in schools where phonics practice was already known to be good. In practice, it was not possible to answer RQ4 due to a lack of information on school performance in the Year 1 phonics screening check before the intervention. The EEF also require all evaluations to include subgroup analysis to explore the impact of the intervention on pupils eligible for free school meals. Again, RQ5 was included for this reason, rather than because theory suggested that the impact of Flexible Phonics would differ depending on whether pupils were eligible for free school meals.

RQ8 was added to the list of research questions when it became apparent that around two-fifths of schools in the trial were also participating in the NELI intervention as part of the government's Covid-19 support strategy. This made it relevant to consider whether Flexible Phonics was more or less effective in schools which received this additional support. The proportion of schools participating in NELI was broadly similar for both the intervention and control groups (26 out of 58 schools in the intervention group and 24 out of 60 schools in the control group), so any impact from NELI would be expected to affect both trial arms to a similar degree. However, the Flexible Phonics trial did not investigate how NELI was implemented in schools assigned to either trial arm. It is therefore possible that differences in the implementation of NELI between trial arms may have had an impact on the apparent effectiveness of Flexible Phonics.

In addition to the research questions to answer in the impact evaluation, the implementation and process evaluation considered 22 additional research questions covering a range of different aspects of delivery. These included:

- whether the Flexible Phonics strategies were being implemented as intended
- the likely feasibility of rolling out the intervention more widely
- engagement with the training

- the frequency and consistency with which the strategies were used
- ways in which the effectiveness of implementation and support for the strategies at all levels within schools might be improved
- the context in which the intervention was trialled, including the approach to phonics teaching in schools allocated to the control group
- the reasons behind schools adapting the intervention in any way.

These questions aimed to shed light on the findings of the impact evaluation, as well as identify ways in which the intervention might be improved.

The main findings from the impact evaluation and implementation and process evaluation

As described in the previous section, the main focus of the study was to investigate whether Flexible Phonics had an impact on children's reading outcomes. This was measured using the YARC Early Word Recognition subscale. To understand and inform findings from the impact evaluation, the implementation and process evaluation explored how schools were delivering Flexible Phonics, changes in practitioner behaviour and knowledge or understanding, and perceived outcomes. The following sections draw together evidence from these two strands of the evaluation to seek to answer the research questions and determine whether the findings support the assumptions underlying the Theory of Change model.

What impact did Flexible Phonics have on children?

The impact evaluation showed that there was no clear positive effect of the intervention on any of the primary and secondary outcomes. Box 7.1 provides a summary of the key findings.[3] This result held regardless of whether prior attainment was considered. The absence of any positive impact on these outcomes implies that the Set-for-Variability and Direct Mapping approaches are not effective in improving the primary and secondary outcome measures in intervention groups compared to control groups at post-intervention. It also means there was no evidence to support some of the key assumptions in the Theory of Change model. For example, pupils did not demonstrate greater word reading ability as a result of receiving the Flexible Phonics intervention and overall literacy and phonological awareness, as measured by the secondary outcome measures, were not affected.

Box 7.1: Summary of key findings from the impact analysis of the Flexible Phonics trial

The analysis used Bayesian multilevel modelling to take account of the clustering of pupils within schools.[4] Previous evidence from an earlier quasi-experimental study of Direct Mapping and Set-for-Variability in Canada was used to inform assumptions about the likely distribution of the primary outcome and improve the precision of the impact estimates (Savage et al., 2018). Bayesian 95 per cent credibility intervals were reported. There is a 95 per cent likelihood that the true impact of Flexible Phonics on the outcome measure is between the lower and upper ends of the reported credibility interval.

The effect of Flexible Phonics on the primary outcome measure of early word recognition was to reduce performance by 0.05 SDs. However, the lower and upper bounds of the 95 per cent credibility intervals ranged from –0.2 to 0.1, meaning it was highly uncertain whether the true impact of Flexible Phonics on the primary outcome was either positive or negative. This is illustrated in Figure 7.1, which shows the effect size (grey dot) and the lower and upper bounds of the 95 per cent credibility intervals (black squares) for the primary and secondary outcome measures. It is apparent from the chart that it was also uncertain whether Flexible Phonics had a positive or negative effect on the two secondary outcome measures.

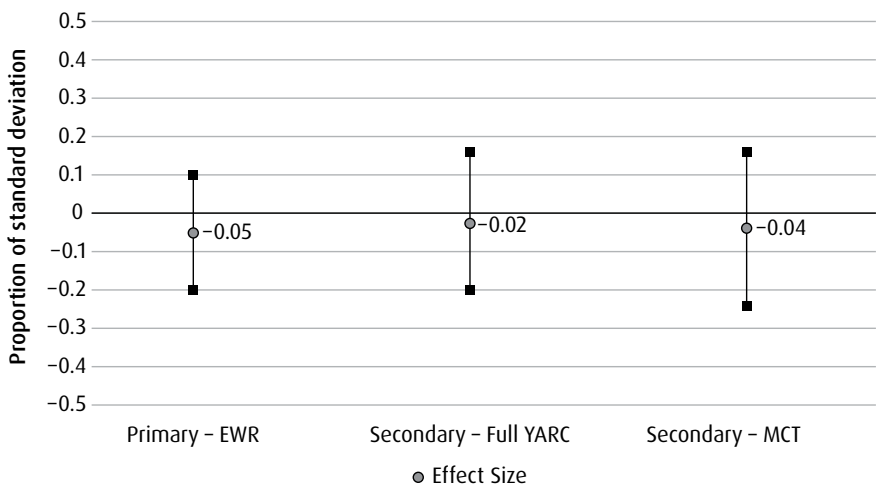


Figure 7.1: The effect sizes for Flexible Phonics on primary and secondary outcomes.

Note: EWR = Early Word Recognition; FSM = free school meals; MCT = mispronunciation correction test

In contrast to the impact evaluation findings, most Reception teachers and teaching assistants interviewed as part of the implementation and process evaluation case studies felt that there had been improvements in children's reading as well as increased confidence and engagement with phonics. This included anecdotal evidence of children sounding out new or exception words. However, when teachers and teaching assistants were asked to report whether they felt that children's engagement with phonics had increased over the year in the final survey, there was no difference between control and intervention groups. As the aim of the Flexible Phonics programme is to build on existing phonics teaching and further extend phonics skills, it is possible that the programme enriches children's phonics skills but that the benefits are not substantial enough to create a significant difference when compared with general phonics teaching. In England, teaching phonics in the Reception year and using a validated systematic synthetic phonics programme has been recommended by the government as good practice for several years, and phonics teaching is well established in English schools. Previous studies that have found significant effects of the Direct Mapping and Set-for-Variability approaches took place in other English-speaking countries, where phonics teaching may not be as firmly established in the curriculum.

One of the key elements of Flexible Phonics, Set-for-Variability, introduces advanced strategies for reading words with an irregular phoneme-to-grapheme mapping, which requires a good understanding of general phonics as a basis. This requirement was set out in the intervention training and resources. If there were children in this sample who were still working on these foundations at post-test, then we would not expect them to have acquired the Set-for-Variability strategies. This is likely to have affected the overall impact of Flexible Phonics on reading outcomes. The results of the mispronunciation correction test, which was designed to capture the development of Set-for-Variability, found no difference in performance between children in the intervention and control groups by the end of the delivery period. However, it is possible that children in the intervention group may continue to develop this skill over time, which could lead to improved reading performance in the future by facilitating the reading of new irregular words through strategies and increased confidence to attempt new words. We will discuss the implications of this non-effect further in Chapter 8.

It is also possible that the context of the Covid-19 pandemic may have reduced the impact that the Flexible Phonics programme was able to have. In the final survey, teachers and teaching assistants reported mixed

impacts of the pandemic on children's phonics skills: just over half thought that some children were delayed while others were at the usual level or above; nearly one-quarter felt that the children were delayed by two to three months. Also, some teachers and teaching assistants interviewed as part of the case studies reported that they were delayed in their phonics teaching due to disruptions to teaching. Research by Bowyer-Crane et al. (2021) with a sample of schools in England also found that children starting school in autumn 2020 were reported – by approximately 90 per cent of those schools – to have delays in communication and language, literacy and personal and social development. Similarly, Rose et al. (2021) found that children in Year 1 were three months behind in reading at summer 2021. Set-for-Variability requires children to have mastered basic phonics skills first and some teachers and teaching assistants were still catching up with general phonics at the start of the Flexible Phonics delivery period. It is difficult to say whether the impact of Flexible Phonics on children's outcomes might have been different if the trial had taken place in a year when children's language and communication and other key aspects of development were not affected by the pandemic. Discussion of how adaptations made to the intervention because of the pandemic may have influenced the possible impact of Flexible Phonics is included in the section 'Limitations of the research findings', later on in this chapter.

Did Flexible Phonics have an impact on particular groups of children?

Flexible Phonics was intended to be a universal intervention appropriate for children of all phonics abilities. Nevertheless, the EEF require that all evaluations explore whether the impact of the intervention on pupil outcomes vary between those who are eligible for free school meals and those who are not. They also require the analysis to consider whether the impact of the intervention varies depending on pupil attainment before receiving the intervention. The impact evaluation found that Flexible Phonics did not have an impact on the primary outcome for children who were eligible for free school meals or who had a lower score on pre-intervention tests of reading ability. In addition, the impact of Flexible Phonics on the primary outcome did not vary between pupils based on their eligibility for free school meals or prior reading ability. More details of the analysis are shown in Box 7.2.

Box 7.2: Effect sizes for Flexible Phonics in subgroups of children

Flexible Phonics raised performance on the primary outcome measure of early word recognition for pupils eligible for free school meals by 0.02 standard deviations. However, as Figure 7.2 shows, there was a 95 per cent chance that the true impact of Flexible Phonics was between −0.21 and 0.25 standard deviations, so it was unclear whether the intervention had a positive or negative impact for this group. Further analysis also found that the impact of Flexible Phonics did not vary between pupils depending on whether they were eligible for free school meals.

Pupils who had lower levels of performance on the combined early word recognition and letter–sound knowledge subscales (below the median) prior to the intervention experienced a reduction of 0.08 standard deviations in their performance on the primary outcome after taking part in the Flexible Phonics programme. Again, however, the 95 per cent credibility intervals spanned zero, meaning that the true impact of Flexible Phonics for pupils with below-average pre-test performance was uncertain. There was also no evidence that the impact of the Flexible Phonics intervention varied depending on performance in the pre-test.

In schools where no pupils were participating in NELI, Flexible Phonics was associated with a reduction in the primary outcome measure of 0.18 standard deviations compared to schools running both NELI and Flexible Phonics simultaneously. There was a 95 per cent chance that the true impact of Flexible Phonics lay between −0.39 and 0.03 for this subgroup. Further analysis found marginal evidence that Flexible Phonics was more effective in schools where some pupils received support from NELI than schools which did not participate in NELI. The 95 per cent credibility intervals reported also spanned zero. The analysis provided marginal evidence that Flexible Phonics was more effective in schools which participated in NELI, as the lower bound was very close to zero (the lower and upper limits were −0.048 and 4.21 respectively). This provides some evidence that the strongest effect on reading outcomes is evident where NELI and Flexible Phonics interventions were both delivered.

Findings from the implementation and process evaluation research suggest that while some staff had concerns that the programme may not work as well with certain groups, there were examples of schools using Flexible Phonics successfully with all the groups mentioned. In the final survey, most staff felt that Flexible Phonics worked well for advanced or average readers and about half felt that it worked well for struggling

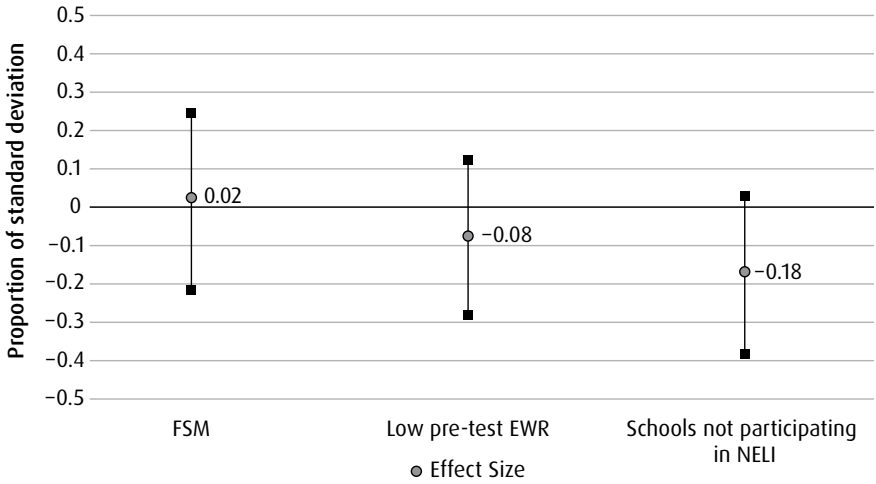


Figure 7.2: Effect sizes for Flexible Phonics in subgroups of children.
Note: FSM = free school meals; EWR = Early Word Recognition

readers. However, fewer staff felt that it worked well for pupils where English was an additional language or those who had special educational needs or a disability.

Consistent with the concerns above that much distinct Set-for-Variability content was not covered before post-testing, staff noted that where children were behind in phonics, they were not able to engage with the Set-for-Variability element. However, others described using the strategies for struggling readers and finding these useful. Although Flexible Phonics was intended as a universal intervention, some staff described using it exclusively to target specific groups of learners, including both higher-performing readers and children who were struggling with phonics. In particular, activities around teaching continuous phonation aimed to help children who were struggling with blending, whereas the Direct Mapping element would be expected to benefit children of all levels.

Did Flexible Phonics have an impact on particular types of schools?

As the NELI intervention was rolled out on a national level across England during the delivery period of the Flexible Phonics trial, the evaluation needed to explore the possible impacts of this and incorporate this factor into the Theory of Change model. In the 'Enabling factors/conditions for success' section of the model, it was suggested that schools taking part

in NELI might have fewer resources to take part in Flexible Phonics, such as less staff time for planning and/or delivering Flexible Phonics. It was therefore thought that Flexible Phonics may be less likely to have a positive impact on pupil outcomes in schools participating in NELI. However, the impact evaluation found some evidence that Flexible Phonics was more effective in schools that participated in NELI (see Box 7.2 for details).

As the NELI intervention targets children struggling with spoken language in Reception, through teaching vocabulary and phonological awareness, it is possible that the direct vocabulary instruction may facilitate Set-for-Variability use. It is also possible that children at schools that were participating in NELI and implementing the Flexible Phonics intervention may have caught up with general phonics faster and benefited more from the Flexible Phonics strategies. For example, evidence presented in Chapter 2 suggests that the Direct Mapping strategy for learning GPCs is closely tied to greater phonological ability. Similarly, the earlier children master general phonics, the earlier they can be taught the Set-for-Variability and mispronunciation correction strategies and the earlier this can impact their word reading. If this is the case, it suggests that Flexible Phonics might have had a greater impact in a year when children were less delayed in their general phonics learning.

Alternatively, it is possible that schools that were participating in both the Flexible Phonics and NELI trials had greater motivation or resources to support language development than schools that were not participating in both trials. It may also be the case that some activities or aspects of the two programmes were similar and so the strategies were doubly reinforced, or that strategies from both interventions were used alongside each other with the whole class or with groups of struggling readers. However, in case study interviews with schools implementing NELI, staff did not mention using NELI strategies with the whole class. For example, in one school, the teacher taught Flexible Phonics and the teaching assistant taught NELI, so neither was familiar with the other programme. The evaluation and delivery teams were unable to directly observe delivery in schools due to pandemic restrictions on visits.

Did Flexible Phonics have an impact when it was implemented as intended?

It is possible that Flexible Phonics had no effect on the primary or secondary outcome measures because of some schools not implementing the strategies as intended. If this were the case, it could reduce the average impact of Flexible Phonics, making it appear that the techniques

were ineffective, when in practice they worked in the subset of schools that were using them. A compliance analysis was used to explore whether Flexible Phonics had an impact on the primary outcome in schools where the delivery team judged that the strategies were being used. However, as the delivery team was unable to visit schools to support and observe delivery, the assessment of compliance made by the delivery team was based on teachers' self-reports. On this measure, two-thirds (67 per cent) of schools were thought to be delivering the Flexible Phonics programme to the required standard. By contrast, a later survey showed that half of all teachers and teaching assistants reported that they were using flashcards to teach exception words by sight memorisation, in direct contradiction to the Flexible Phonics approach. Based on the compliance measure agreed between the delivery and evaluation teams, it appeared that even when the impact evaluation focused on this subset of schools that were implementing Flexible Phonics as intended, it did not have a discernible effect on the primary outcome. However, the survey results suggest that even in schools that were judged to be compliant, some aspects of teaching practice may have deviated from the intended approach.

How easy was it for schools to implement Flexible Phonics?

The majority of schools reported that it was straightforward to incorporate Flexible Phonics activities and strategies into their usual phonics approach. The programme was designed to be flexible, so schools were able to make small adaptations, such as changing the suggested order in which sounds were taught to fit with their existing phonics teaching. The implementation and process evaluation did highlight the value of clearly stating that Flexible Phonics is consistent with government guidance for schools in England around using a validated systematic synthetic phonics programme, as well as allowing the flexibility to suit existing practice when initiating discussions with schools about implementation. The Department for Education (DfE) updated its essential core criteria for systematic synthetic phonics programmes in April 2021, while delivery was still underway (DfE, 2023a), as well as its guidance around choosing a phonics teaching programme (DfE, 2023b). This most likely contributed to uncertainty or confusion among schools around the suitability of Flexible Phonics.

Intervention schools used a range of programmes for their general phonics teaching and there was no systematic evidence that Flexible Phonics was incompatible with any of these approaches. Indeed, some

school staff and delivery team support partners noted that the Sounds-Write phonics programme had similarities to Flexible Phonics in that it also did not teach exception words through sight learning.

What did it cost schools to implement Flexible Phonics?

The main challenge faced by schools in implementing the Flexible Phonics programme was the time required to work out how best to integrate Flexible Phonics into existing teaching practice, or to add in a daily Direct Mapping reading session if needed. However, the amount of time varied by school and no schools participating in the trial indicated that this was impossible or too onerous. A further challenge identified during interviews was that there were few instances of the target phoneme in some of the Direct Mapping books. However, overall, school staff were positive about the books provided and the way children in their class engaged with them.

The monetary costs of participating in the Flexible Phonics programme were relatively low, with schools generally using or adapting existing phonics teaching resources to teach the Flexible Phonics strategies. In addition to this, schools were given free access to books used for Direct Mapping, a training manual and an online portal where they could share resources with other schools. In the trial, the cost of these resources was covered by the EEF and the Institute of Education of UCL, but if the Flexible Phonics programme was rolled out more widely it would be necessary to charge a programme fee to cover these costs. However, calculations of costs produced as part of the evaluation suggested that even if schools paid for these resources, Flexible Phonics would remain a low-cost intervention.

Did the Flexible Phonics programme affect teacher and teaching assistant practice or confidence?

Practice

The final survey of teachers and teaching assistants explored any changes to practice related to phonics, reading comprehension, spelling and handwriting. The areas explored were drawn from the Ofsted (2017) report titled 'Bold Beginnings: The Reception Curriculum in a Sample of Good and Outstanding Primary Schools'. Statistical tests found no difference between practice in the intervention and control groups on these activities. The case studies, however, did find that most schools were incorporating aspects of Flexible Phonics in their daily phonics teaching.

Most staff reported that they were implementing the Direct Mapping element, and some schools were using texts other than the books provided as part of their delivery of this element. The majority were also using mispronunciation correction activities to introduce children to the idea of trying to correct a deliberately mispronounced word. Less than half of case study schools had introduced Set-for-Variability when reading exception words. Staff at some case study schools also described teaching other strategies such as 'flipping sounds', where a grapheme is mapped to more than one sound. Several case study schools reported that they now embedded phonics activities throughout the day, such as playing phonics-based games when lining up and when writing on the whiteboard or reading with the class.

Most teachers and teaching assistants interviewed intended to continue with the Flexible Phonics activities they were using, as they felt that they worked well and children engaged with them well. The practices and behaviours explored in the endline survey were quite general phonics activities, such as how often they undertook activities involving reading exception words. So while the Flexible Phonics programme did not result in high-level, broad changes to phonics teaching, qualitative evidence from the case study schools suggested that teachers and teaching assistants did incorporate activities and strategies from the programme into their phonics teaching.

Confidence

The Flexible Phonics intervention aimed to enrich the understanding of staff and their confidence in phonics by introducing two novel strategies – Direct Mapping and Set-for-Variability – as well as other approaches and advice for teaching general phonics, which incorporated the latest findings from the research literature. These strategies were intended to enhance teachers' and teaching assistants' ability to be flexible and adaptive when teaching phonics. In addition, the Flexible Phonics programme included advice around the order in which to teach GPCs and why, as well as strategies to support children who were struggling with blending. While not being a core focus of the Flexible Phonics programme, the evaluation also explored whether learning about this range of strategies and the underlying language processing during reading affected staff's overall confidence in phonics.

In the baseline survey at the start of the year, the vast majority of teachers and teaching assistants participating in the trial agreed that they felt confident in their ability to teach phonics or support phonics teaching with Reception children. This suggests that teachers and teaching

assistants in Reception were already quite confident at teaching phonics. The final survey showed that staff in both intervention and control group schools did not differ in their level of confidence in teaching phonics compared with the start of the year, suggesting that the Flexible Phonics programme did not affect confidence. The government encourages schools in England to teach systematic synthetic phonics programmes, so phonics teaching is already well established. As a result, it is possible that the Flexible Phonics training did positively affect confidence, but that changes in confidence could not be measured, as scores were already at a ceiling.

Limitations of the research findings

As the trial ran during the 2020–1 academic year, the Covid-19 pandemic had a substantial impact on the day-to-day running of primary schools; the everyday lives of families; the training, support and delivery of the Flexible Phonics programme; and the delivery of the evaluation. The delivery period was reduced from five months (20 weeks) to approximately three-and-a-half months (14 weeks) and training in Flexible Phonics techniques was changed from a face-to-face format to online. Pandemic-related restrictions that limited access to schools also meant that the delivery support team and the evaluation team were unable to enter schools to observe the delivery of Flexible Phonics in practice. This meant that both delivery and evaluation teams were reliant on educators' self-reported descriptions of how they delivered the intervention and how they perceived children's engagement with the approach. This may have limited the delivery team's ability to fully support the needs of the teachers and schools delivering Flexible Phonics and to accurately assess the degree of compliance in delivering the programme. The delivery team reported that some schools did not start delivering Flexible Phonics until after the Easter holidays, so the total intervention time for those schools was closer to 10 weeks. It also meant that the evaluation team had a more limited picture of how schools were delivering Flexible Phonics and were unable to access a broad range of contextual information about the school and local environment.

The Covid-19 pandemic impacted the collection of child assessment data to a limited extent. It was mostly possible to conduct the YARC and mispronunciation correction test assessments in person both before and after the intervention, but some schools requested that assessments take place online. The need for classes (teachers and pupils) to isolate at home

when someone in the class tested positive for Covid-19 meant that some pupil data could not be collected, as it was not possible to return to test the class within the assessment period. To mitigate this, the assessment team allowed extra time towards the end of the assessment period to return to schools or conduct additional visits if many children were absent during previous visits. This minimised the potential impact of the pandemic on the level of attrition in the study, which, at 20 per cent, was still slightly larger than anticipated. The rate of attrition was nevertheless lower than that seen in other recent studies and there were no differences between the intervention and control groups. It is assumed that this missingness is random, but analysis of the missing data suggested that the post-intervention primary outcome measure was more likely to be missing for those eligible for free school meals. However, the likelihood that the post-test score was not recorded did not vary between trial arms; in addition, a sensitivity analysis, in which the missing scores were imputed, produced findings consistent with the main analysis.

It is likely that the impact of the pandemic would be similar across intervention and control groups, given that schools were allocated to either trial arm at random. However, the disruption to children's education and development, as well as to the delivery of the Flexible Phonics programme, may have affected the level of impact achieved and findings may have been different if this disruption had not occurred.

Analysis of the pre-intervention test data for the Early Word Recognition subscale suggested that they may not be suited to measuring pre-test reading ability for Reception-aged children. Pre-test scores in particular were heavily concentrated around zero, with a mean average score of 3.75 for the intervention group and 4.08 for the control group from a total possible score of 30. There was also limited variability between pupils, indicating that it did little to distinguish between pupils' reading ability at baseline. This is likely to explain why the findings were similar when the main analysis was repeated without controlling for the pre-test score. It is possible that the low scores were partly due to the delays in language development observed during the pandemic. However, the extent to which the test failed to capture variation between Reception-aged pupils was unforeseen.

Without further research it is not possible to say whether the increased effect of Flexible Phonics for schools also participating in the NELI programme was because of the focus on struggling learners, the strategies and activities taught in the NELI programme, or the possilibity that these schools had greater motivation or resources for language support. Assuming that at least some of the effect was not driven by school

motivation or resources, it would have been beneficial to explore which aspects of the NELI programme worked to complement and support the delivery of Flexible Phonics.

Lessons learnt for future delivery

While the Flexible Phonics programme was delivered in exceptional circumstances which involved significant changes to how the programme was delivered as well as children's learning and home environments, the evaluation did identify some general points that might increase the likelihood of the intervention being effective in any future delivery. These potential improvements are set out below.

The evaluation highlighted the need for the delivery team to continue to emphasise the importance of teaching assistants receiving training to allow them to ensure effective delivery of Flexible Phonics. In some cases, teaching assistants were unable to attend all training and support sessions, limiting their proficiency in using the strategies. While delivery team attendance data indicates that 90 per cent of teaching assistants received training at the sessions, the delivery team were aware of teaching assistants not attending because they were covering classes. Where staff were unable to attend a session, they were sent a recording afterwards. Also, teachers sometimes dominated support sessions, with lower levels of engagement by teaching assistants. This suggests that offering separate support sessions for teaching assistants might increase the use of Flexible Phonics by teaching assistants and improve the consistency of delivery.

While most teachers and teaching assistants reported that they understood how the Flexible Phonics intervention worked and the underlying language processes, some felt the training could have used less technical language and fewer references to theory. Some participants also felt that the training was quite long and may have benefited from including practical demonstrations with children or videos of elements being delivered with children. This was not possible during the trial, as restrictions introduced in response to the pandemic made it difficult to try out strategies with children as part of the training or to film a demonstration. The initial training and resources included video demonstrations where members of the delivery team role-played as a child and a teacher; later, during the delivery period, one of the support partners was able to film demonstrations with her child. However, practical demonstrations would be a valuable addition to future training

if this were delivered in person. It is possible that reassessing the balance of theory and practical elements, and the length of the training, might improve the effectiveness of Flexible Phonics. The mode of training delivery – that is, in person or online – could also be considered.

As noted earlier, during the trial one or two schools needed reassurance on how Flexible Phonics would fit alongside existing phonics programmes or the requirement to use a systematic synthetic phonics programme in England after the Department for Education updated its essential core criteria for systematic synthetic phonics programmes in April 2021 (DfE, 2023a). To increase the likelihood of schools taking up the intervention in any future roll-out, it would be helpful to continue to emphasise that Flexible Phonics can be delivered alongside a school's usual phonics programme and can be adapted to teach sounds in the same order as their existing phonics practice.

A third of schools were identified by the delivery team as non-compliant and continued to teach some aspects of phonics in a way that conflicted with the Flexible Phonics approach. Additionally, half of staff surveyed at endline reported teaching some pupils exception words by sight. Also, some were uncertain about which elements of the programme were compulsory. These deviations from the intended approach to using Flexible Phonics may have reduced the effectiveness of the programme. Future training and follow-up support could be improved by giving greater clarity on which aspects of the programme are mandatory. More tailored guidance could be provided according to schools' existing approaches or preferred programmes for teaching phonics and the observed needs of children as better support for differentiation. Consistent with this, some teachers and teaching assistants would have liked a member of the delivery team to observe their delivery of Flexible Phonics in person to give feedback on whether they were implementing the programme correctly. This was not possible during the pandemic but might be feasible for future delivery. The delivery team invited staff to send a recording of themselves delivering a Flexible Phonics session for observational feedback, but no schools took up this offer.

Finally, reviewing the effectiveness of the Flexible Phonics resources might improve this aspect, given the feedback that some of the books supplied for Direct Mapping included few examples of the target sounds, especially compared with the reading scheme books that some schools may have been using. It may be possible to identify additional books with a higher frequency of some sounds. Also, engagement with the online portal with resources to support teachers and teaching assistants in implementing Flexible Phonics was low. As there was a need to prevent

schools allocated to the control group from accessing the intervention materials or viewing discussion boards on Flexible Phonics, it was necessary to restrict access to the portal by requiring users to log on. This barrier may have reduced the use of the portal. In a wider roll-out of Flexible Phonics in future, materials could be hosted on an open website, increasing the likelihood that teachers and teaching assistants will make use of the available resources. As engagement with the portal was low, the delivery team shared information and resources via monthly email newsletters and through support partners working directly with schools. It may be helpful to consider incorporating these and other alternative ways of sharing peer learning and raising awareness of resources in future delivery.

Conclusions

This study found that Flexible Phonics was no more effective in improving reading skills than best-practice phonics teaching when taught to children aged four to five in Reception classes in England. It therefore did not replicate the findings of Savage et al. (2018), who found that this approach taught to struggling readers aged five to seven in Canada had resulted in improved reading skills, with additional positive outcomes for spelling and reading. As phonics teaching in Canada starts later, the expectation was that children in England aged four to five would be at a similar developmental phase of reading as those who participated in the Canadian study. However, it is possible that the age difference may have affected the level of impact.

Previous studies have shown that explicitly linking phonics learning with a relevant reading task (Hatcher et al., 1994, 2004, 2006; Shapiro & Solity, 2008) or with Direct Mapping (Chen & Savage, 2014) was more effective than regular phonics teaching or a vocabulary learning task. While there was no clear impact from the intervention, as noted above, some staff reported that the linked sound did not occur very frequently in certain texts and a few teachers and teaching assistants noted that they already read books with children to reinforce learning as part of their existing phonics programme. Either of these factors could have reduced the potential effect of this aspect of Flexible Phonics.

Wider evidence from the literature suggests that teaching Set-for-Variability is more effective than standard phonics teaching for reading irregular words (Zipke, 2016; Dyson et al., 2017) or than reading all words, including those with regular pronunciation (Elbro et al., 2012;

Kearns at al., 2016; Elbro & de Jong, 2017). Although the intervention had no impact, some teachers and teaching assistants at case study schools described children as being more willing to attempt reading new words and to try different approaches, and less concerned about getting the word 'wrong'.

Research on the impact of the pandemic on primary school children has consistently found pupil development delayed across the board by several months (Bowyer-Crane et al. 2021; Rose et al., 2021), which aligns with the delays in language and communication development for the pupils in this evaluation reported by teachers and teaching assistants. Most recently, the 2021–2022 Year 1 phonics screening check confirmed an impact on language skills for this cohort at the end of Year 1 as well: the percentage of children meeting the expected standard fell from 82 per cent in 2019 to 75 per cent in 2022 (DfE, 2022).

Set-for-Variability is a relatively advanced strategy that is taught after children have learnt synthetic phonic blending of GPCs to derive simple (consonant-vowel-consonant) syllable pronunciations. As the pandemic meant that children participating in the trial were delayed in their language and communication development, they completed their phonics teaching later than usual. This is likely to have resulted in a large number of children learning Set-for-Variability strategies at a later point in the academic year than would be typical, which will have limited the potential impact of a core element of the programme. Such effects may be evident in delayed post-test assessments of reading.

Finally, the wider context of the pandemic may have impacted the personal, social and emotional development of pupils (Bowyer-Crane et al., 2021), affecting concentration and school attendance. It was also necessary to reduce the length of the Flexible Phonics programme due to the disruption to teaching caused by Covid-19. These deviations from the intended design of the programme and the context in which it was implemented are likely to have had an impact on the effectiveness of Flexible Phonics.

Notes

1 See also EEF (2024) for guidance on specific aspects of evaluation design.

2 Letter Sound Knowledge (Core) 0.95, Letter Sound Knowledge (Extended), Early Word Recognition 0.98, Sound Isolation 0.88, Sound Deletion 0.93, Sound Deletion and Sound Isolation Combined 0.95.

3 Further details of the analysis can be found in the evaluation report (Dawson et al., 2024). The Flexible Phonics efficacy trial was funded by the EEF. The Flexible Phonics evaluation report is available under the Open Government Licence © Crown copyright 2023. You may re-use this information (excluding logos) free of charge in any format or medium, under the terms

of the Open Government Licence. To view this licence, visit: https://nationalarchives.gov.uk/doc/open-government-licence or email: psi@nationalarchives.gsi.gov.uk. Where we have identified any third-party copyright information, you will need to obtain permission from the copyright holders concerned.

4 The analysis was based on a Markov chain Monte Carlo sample size of 100,000 with a burn-in period of 2,500, repeated for ten chains.

8

Lessons to be learnt for greater impact

This final chapter naturally seeks to pull together all the work in the book to reflect on the most important points and assess what has been learnt. Our focus is on three broad questions: firstly, what have we learnt about improving reading attainment from this specific trial? Secondly, sitting alongside this: what are the implications for theories of typical and atypical reading development? Finally, we explore a broader issue: what can be learnt about the complex but important task of delivering educational randomised controlled trials (RCTs) from the detailed overview of this trial? Our stated perspective in Chapter 1 was one of exploring RCTs as a particular form of complex cultural human activity shaped by a design dichotomy (treatment activity versus control activity). What has been learnt from this RCT activity?

This book has been built on a series of components. Chapter 1 presented the case that RCTs – and 'realist' trials more specifically (Bonell et al., 2012) – with modifications are needed to explore how real-life variance in response to interventions is associated with strongly theorised predictors of outcomes. We presented the view that RCTs have unique methodological strengths for hypothesis testing and programme evaluation. Their strength is in supporting causal inference through randomisation. We have noted that this strength is also, however, a potential weakness if used uncritically: the control that RCTs offer for extraneous variables (EVs) may lead to the simplification of the complex, multiple and distal (indirect) causes, including structural influences, that may exist in the real world outside RCTs. RCTs are at their strongest, however, under certain conditions:

- when they form part of a mosaic of methods for scientific understanding

- when they are closely linked to strongly theorised scholarship and research that is often based on non-RCT methods
- when they are augmented with statistical path analyses, including mediation, and qualitative pilots and process analyses
- when there is clear pre-specification of key theorised processes and functions within an intervention.

As such, strong RCT designs – including this one, Flexible Phonics – have much in common with realist RCTs (Bonell et al., 2012).

We also argue in Chapter 1 that educational RCTs may need to be somewhat reconstrued beyond considering only indices of average treatment effects. This is because there are plenty of reasons for thinking that the most effective delivery of educational change is *inherently dimensional* – varying across contexts and participants – as opposed to being uniform or akin to a pill in pharmaceutical research. Effective educational interventions must accommodate a range of reading levels within a given class. Teachers will likely vary in their uptake of (and methods for delivery of) any intervention, given that children's needs vary and teachers have autonomy and expertise. Indeed, without this variation, the intervention will likely not succeed (as well) at scale. Homogeneity is not, even in principle, desirable. Dimensional – not just factorial – RCT intervention designs are needed to achieve this and inform questions about for whom and when interventions in a given class may add value.

Thus, any teacher intervention education should, we argue, start with these assumptions. We favour interventions based on a clear and potentially falsifiable Theory of Change and which use change models that can accommodate treatment effect heterogeneity (varying impacts) across multiple settings, along with strong mediation-based models of RCT analysis (Bonell et al., 2012). Related to this, when constructing professional education development, we have argued that it is a reasonable assumption that teachers have greater knowledge and expertise in their classrooms (Boldrini et al., 2023), which has led us to explore the theory and practice of collaborative working with partner schools in our RCT.

We also noted the potentially confounding expectation effects in non-blinded RCTs exploring the average effects of an intervention. Expectation-based effects, including generic Hawthorne effects, do not predict a theorised mediation effect (another reason to use such models). Expectation effects also do not predict the longer-term effects of an intervention in delayed post-intervention tests, because Hawthorne is construed as a short-term motivational advantage only. We had hoped to

use such mediation models here, but this was precluded by null patterns of overall main effect results. Such models could, however, potentially be explored for reported *interaction effects* between Flexible Phonics and the NELI outcomes, described in Chapter 7. Theory predicts that baseline Set-for-Variability measures should be a predictor of responsiveness to the intervention in this specific case. What also remains to be explored here is an exploration of key *moderating effects* (for example of initial attainment levels in reading, phonological awareness and Set-for-Variability). We believe this strongly theorised, more dimensional conception for the assessment of an effective intervention potentially contributes to a still fledgling implementation science of reading.

Chapter 2 provided a thorough and timely narrative review of all evidence on Set-for-Variability and Direct Mapping. This included evidence of the importance of these from wider psychological and other research published since the 1960s in the fields of cognitive and experimental psychology, linguistics, computational cognitive science, pedagogical task analysis, word corpus analyses and, still in its infancy, educational neuroscience. At the centre of this is the notion that generative self-teaching tools beyond phonic blending alone are required to master the opaque 'outlier' spelling system of English. We concluded that Direct Mapping and, most especially, Set-for-Variability are promising theory-based tools that can help with this computational challenge. Preliminary evidence reviewed here suggests that this strategic flexibility is, indeed, teachable, but that much more work is needed to explore impacts on outcomes.

Chapter 3 explored the complexities of developing the Flexible Phonics trial. We noted that partnership working, most particularly with the Department for Education, was often the art of realpolitik as much as the pursuit of a science of reading. Consequently, we suggest, as scientists, the importance of expert independent peer review of proposals and government stakeholder education on the role of novel RCTs in the improvement of public services. Here the science needs to be allowed to speak if it is to serve the public good as fully as possible.

The partnership working with the EEF and independent evaluation teams involved many funding and timing constraints, in part reflecting the EEF's role and evaluator remit, and at other times reflecting the context of the Covid-19 pandemic, though sensible compromise was largely possible through a process of collective problem-solving. Some primary analyses that were mooted early on – many in the spirit of realist RCTs – but were not possible became some of the candidates for secondary analyses, and are discussed further below. Mediation analysis, at least from post-test

oral Set-for-Variability tests, was planned for. One important feature of this EEF trial was the pilot study approach, which was used to inform the subsequent RCT. This pre-RCT situated approach aided the RCT by identifying needed modifications to the intervention content and the trial delivery. The results suggest the importance of social validity (teacher acceptance of the value of our intervention content) and of our co-expert intervention approach in intervention uptake (Savage & Fox, n.d.). These are all potential implications for subsequent EEF-funded RCTs.

Chapter 4 focused on the content of the Flexible Phonics training and the post-training practitioner support model and how these elements were managed in the context of the pandemic. Next we turned to the post-training intervention support model, where a four-stage flexible framework based on the solution-focused coaching model was utilised by the support partners during structured online support calls with schools. The allocation of a dedicated support partner to each school during the intervention implementation phase appeared to be mutually beneficial for building trusting working relationships. It also enabled the support partners to obtain a more detailed understanding of the school context, so that they could provide tailored help and support and facilitate schools' implementation and delivery of the intervention and track and monitor progress between meetings.

Due to partial school closures during the Covid-19 pandemic, planned observations of teachers delivering Flexible Phonics in classrooms were abandoned. Instead, treatment fidelity was rated using the global treatment fidelity rubric (GTFR) tool, an adapted version of the ABRA implementation fidelity measure rubric developed by Savage et al. (2013). Where intervention delivery was more extensive based on the ratings from the GTFR tool, the whole school team were fully invested in delivering the intervention, open to new approaches to teaching exception words and well supported by school senior management. In contrast, intervention delivery was less extensive in many one-form entry schools. These schools mentioned they found the pressures of planning the entire curriculum, together with additional workload-related challenges caused by the Covid-19 pandemic, heavily impacted their ability to implement the intervention.

Chapter 5 explained the importance of active collaboration between the delivery teams, the evaluation teams and schools. Working in partnership with practitioners, and utilising their knowledge and expertise about what works in practice in the classroom, helped to strengthen the design and implementation of all aspects of the trial, from intervention content to the design of the intervention training and

the follow-up support provided post-training. The complexities of the pandemic brought about the need for agile adaptions to be made to all aspects of the trial. Strong partnership working with schools – keeping their needs and concerns at the forefront of any trial adaptations, together with clear and transparent communication – was key to successfully running the trial with minimal attrition.

Based on the arguments and delivery contexts outlined above, Chapter 6 discussed the formal evaluation process, including working with the funder to set up a trial, writing a protocol and statistical analysis plan, gaining ethics approval, handling data management issues, planning and carrying out the work, and reporting. Chapter 6 discussed in detail the design of the impact evaluation and what features teams can think about when preparing sample size calculations for their trials. It also outlined our implementation and process evaluation and how this had to be updated due to the constraints of Covid-19 (mainly a move to virtual interviews and observations, which both worked well). The chapter also explored changes over the course of the trial to the EEF's guidance and approach to implementation and process evaluation research; a wider issue in RCTs is that requirements around implementation and process evaluation research and reporting are less rigorous than those for the impact evaluation. Finally, the chapter covered partnership working, with both the delivery team and funder, from the evaluation team's perspective and how this worked successfully despite the challenges that the pandemic had created. In addition, our way of working with the test administrators and schools was also detailed. This allowed us to maintain good relationships throughout the trial, reduce the burden on the schools and achieve a reduced attrition from data collection.

Chapter 7 reported the quantitative analyses and qualitative analysis as run by the evaluation team. The planned quantitative analyses showed that there was no statistically significant impact of the intervention on the Early Word Reading raw score – the primary outcome of interest. Further analyses showed the result was robust to several sensitivity tests. Additionally, the intervention also had no significant impact on either of the secondary outcomes of reading comprehension and the candidate moderator variable – the mispronunciation correction test. There was, however, some evidence that the intervention may have been more effective in schools participating in the Nuffield Early Language Intervention (NELI) in combination with the Flexible Phonics intervention.

In this context, additional and unplanned preliminary exploration of raw score growth for irregular words (words which do not follow primary phonic rules, such as *island* and *muscle*) which might particularly benefit

from an Set-for-Variability intervention was subsequently undertaken and is reported here for the first time. This showed larger effect sizes for irregular words when both NELI and Flexible Phonics interventions were delivered (d = .95), as compared with NELI alone (d = .65). The equivalent contrast for regular words that follow phonic rules was d = 1.42 versus d = 1.48, respectively. This analysis shows that the effect is specific to irregular words, a pattern expected based on many theoretical views of where Set-for-Variability might operate and that we considered in Chapter 2 (Venezky, 1999; Dyson et al., 2017). There was no evidence to suggest that Flexible Phonics had a differential impact based on whether children's reading ability was above or below average at pre-test. Equally, eligibility for free school meals was not a significant predictor of intervention outcomes.

Turning to implementation and process analysis data, teachers and teaching assistants reported that it was relatively straightforward to deliver Flexible Phonics and integrate it into their existing phonics teaching, across a range of phonics programmes. Costs and additional time needed for planning and delivery were minimal. School staff felt that children engaged well with activities; some observed that children were more willing to attempt reading new or difficult words and felt less worried about 'getting it wrong'. Teachers and teaching assistants surveyed showed no change in their confidence or practice regarding teaching phonics, though it was noted that levels of confidence in teaching phonics were already high at the start of delivery.

A minority of schools did not teach Flexible Phonics to all pupils in Reception. Data from virtual support meetings with schools suggested this was often because the schools were short-staffed due to Covid-19 rather than being due to staff uncertainty about Flexible Phonics content. Some teachers and teaching assistants were confused as to which were the key elements of the programme, so future delivery could seek to clarify these aspects of the work. It was noted that some teaching assistants were unable to attend training and/or support outside of teaching hours. Future intervention delivery could thus usefully explore how best to facilitate teaching assistant engagement.

Pulling all results together, the formal report of this work (Dawson et al., 2024) concludes by noting that while the predicted main effects of the intervention were not found, the context of the Covid-19 pandemic caused significant disruption to the delivery of Flexible Phonics and to children's education more generally. This disruption may have impacted the effectiveness of the intervention.

Take-home messages from this book

Interpreting findings in the context of this book

It might seem perverse to devote an entire book to considering a trial that has in some important sense 'failed' in its primary conceptual objective of evidencing a theorised link between Flexible Phonics pedagogies, such as Set-for-Variability, and improved word reading outcomes in Reception classrooms. A hypothesised significant impact of teaching Flexible Phonics strategies on a a pre-specified, standardised primary outcome that was both valid and reliable – that is, word reading – was *not* achieved here. Such findings are humbling. They suggest that Flexible Phonics content does not easily add value to reading ability in Reception classrooms when delivered largely as a whole-class intervention.

However, here, as in many cases, the failure to find an effect is perhaps the point of greatest learning, and in these other senses we argue the trial is not a failure at all. How, then, might this trial be understood? One way to approach this is to go back to the process outlined in Chapter 3, where we considered the process evaluation, to revisit whether we can distinguish between:

1. theory failure (the intervention does not work as predicted)
2. implementation failure (it was not implemented as intended)
3. methodology failure (inadequate evaluation methods were selected, or suitable evaluation methods were used inadequately).

Taking these hypotheses in reverse order, in the broadest sense we have no reason to doubt the specific methodologies used. The approaches used here to assess reading professional development and the measures used to assess outcomes are appropriate. We also know that other approaches from our stable of interventions, such as ABRACADABRA (Savage et al., 2013; Johnson et al., 2019; Vousden et al., 2022; Bell et al., 2022), when scaled across multiple sites in randomised evaluation trials following standard EEF trial methodology protocols (such as those used here), can produce robust, statistically significant impacts on standardised reading outcomes in early primary school settings in England when compared with control schools.

A greater concern here, perhaps, is how methodology interfaces with implementation. We do know the intervention was reduced from 20 to 14 weeks due to the Covid-19 pandemic. Beyond this, we can unfortunately only surmise as to the 'failure of implementation

hypothesis', as not one member of the UCL delivery team (principal investigator, project manager, support partners) ever set foot in a single school or met even a single teacher or teaching assistant face to face in the entire trial – because of the pandemic. Implementation in classrooms was never observed directly by the delivery team. It is thus not known how well the intervention was delivered. It was also impossible to calibrate support to teachers based on observed practice, which may have been key to producing a greater impact.

What evidence is there that the study was well implemented? On the positive side, as described in chapters 4 and 5, we developed a strong and clear rubric that was used to assess implementation fidelity. This measure had high internal reliability. Rich evidence from the teacher self-reports in Chapter 7 suggests Flexible Phonics *was* implemented. Certainly, the support provided to teachers by support partners was well executed, as we can directly testify. Other evidence from the pilot study reported by Savage and Fox (n.d.), and summarised in Chapter 3, suggested that the content and approach to shared co-expert delivery of the intervention were well received by teachers and teaching assistants (albeit in its traditional face-to-face delivery format). Attendance rates, engagement and responsiveness to online training in the Reception trial – and the fact that Year 1 teachers from 29 of 59 intervention condition schools attended follow-up training to take Flexible Phonics ideas into Year 1 teaching – speak to enthusiasm among school staff for the Flexible Phonics RCT.

The very fact that many schools added NELI alongside Flexible Phonics to their regular workload arguably speaks to the sense of urgency and purpose in finding all ways to support children during the delivery phases of the study. All of these actions suggest that all educators were really trying their best for young children and may even have redoubled their efforts during the pandemic to support pupils and prevent learning loss from the partial school closures. It is also possible that this unique sense of urgency impacted the activity of control schools, thus representing a form of John Henry effect wherein control schools redouble efforts because they know they are a control. There is of course also a possible selection effect, wherein schools motivated to improve phonics teaching are more likely to sign up for the Flexible Phonics trial in the first place than the average school.

On the other hand, we may need to consider the enthusiasm and commitment to the study from the perspective of *quality* of delivery. Unfortunately, this was impossible to assess directly due to the context of the pandemic. It is quite likely that the professional capacity of school staff

to change their practice was (quite understandably) much reduced in the pandemic context, with its multiple impacts on teachers' workloads and professional lives. Indeed, it was commendable that they even considered such professional learning at this time.

Direct observations of teachers of older struggling readers that we undertook after the pandemic have occasionally suggested that there may be misunderstandings about the role of Set-for-Variability. It is a strategy for learning (acquisition of) new, unfamiliar irregular words, but some professionals have construed it as a strategy for *overlearning* of irregular words already known. A further concern that is noted in the formal report (Dawson et al., 2024) is the reported finding that 50 per cent or more of the intervention group educators who responded to the endline survey reported that they continued to teach exception words as sight words, as they had always done. The qualitative comments also contained further clues suggesting that there were misunderstandings. For example, one educator noted they had all the information they needed after only the first online training session, whereas much of the key content essential for understanding Flexible Phonics (especially Set-for-Variability) was delivered in the second and third online training sessions. It is also possible that effects were diluted in schools that were delivering multiple interventions (Flexible Phonics and NELI). However, the larger effect sizes for combined Flexible Phonics and NELI interventions, as noted earlier, potentially argue against this view.

Looking more broadly, we know that the 2020–1 Covid-19 pandemic produced overall learning loss: an estimated learning loss of two to three months for literacy (Thorn & Vincent-Lancrin, 2021; Betthäuser et al., 2023) and widening gaps between children based on features of family background (Achtaridou, et al., 2022; Betthäuser et al., 2023). There is also now wider evidence internationally that the partial school closures meant that home pedagogical influences, rather than school pedagogical influences, impacted outcomes (Dunn et al., 2023). This all speaks to the need to explore the impacts of intervention on the weakest readers.

Turning to the 'failure of theory' hypothesis, if we assume for the sake of discussion that the results as reported do represent an accurate assessment of underlying intervention effects, and not the effects of the pandemic context on delivery efficacy, then several interesting observations can be made. Theoretically, the results may show that not all children seem to need the additional direct instruction in managing grapheme-to-phoneme inconsistency offered by Flexible Phonics. Revisiting the evidence and data presented in Chapter 2, it might be that many children work out the inconsistencies of English

using Set-for-Variability for themselves, while others may not. The body of evidence showing concurrent and longitudinal predictive validity for individual differences in Set-for-Variability measures (Tunmer & Chapman, 2012; Steacy, Compton et al., 2019; Steacy, Wade-Woolley et al., 2019; Edwards et al., 2021) and Direct Mapping measures (Savage et al., 2020) on reading might be interpreted thus. Further (envisaged) analysis in this intervention sample, showing that baseline Set-for-Variability measures are a strong unique predictor of reading outcomes, would support this view. If so, the results here might indicate that Set-for-Variability is important in development, but that it is not necessary to teach it to all children.

This view regarding the need for Set-for-Variability teaching can only be stated very cautiously, because we know the pandemic affected teaching and learning significantly. We also know from the implementation and process evaluation that at least one setting reported that they were still catching their class up on basic phonics in the summer term and so could not even start on the Set-for-Variability-related activities. The project evaluation report (Dawson et al., 2024) noted in conclusion that the majority of teachers rated at least some of the children in their class as behind in phonics compared to a typical year. Wider data on developmental delays in reading in England due to Covid-19 (for example Bowyer-Crane et al., 2021) is consistent with this view. Without foundational phonics, Flexible Phonics will not add value. Arguably, then, what is really needed to establish the patterns of impacts of Flexible Phonics firmly is a study where Set-for-Variability is introduced at the right time developmentally (when children have the prerequisite synthetic phonics and vocabulary in place) and perhaps also for the right students – those who show, through their mispronunciations of irregular words, that they will benefit from the Set-for-Variability approach.

A modification of the theory regarding the need for direct instruction is thus that the Flexible Phonics intervention is best targeted at certain poor readers rather than whole samples of representative Reception children. Chapter 7 showed that sub-analyses on children in receipt of free school meals and analyses of the lower 50 per cent on pre-test early word recognition did not alter headline null effects of intervention. However, a further nuanced sub-analysis of the existing data could still be useful. As noted above, we also know that the pandemic has had uneven generalised effects on learner outcomes, where school closures foregrounded the effects of home education (Dunn et al., 2023). Children from low socioeconomic contexts and those with fewer family learning opportunities were the most badly affected in terms of learning

outcomes (Betthäuser et al., 2023). Such findings suggest a nuanced investigation of the very poorest readers from low socioeconomic contexts (across treatment and control groups). Here we might emphasise that it is important to investigate standard score results to consider *growth* versus *catch-up* learning, given this complex pandemic delivery context. It might usefully be noted that this theorising is not so much a 'failure' of the original theory as a quite standard scientific process of principled theory revision in light of empirical data emerging from energetic efforts at theory falsification – that is, thoroughly testing the theory and its assumptions. This revised theory makes a potentially interesting prediction that phonic inflexibility rather than just phonics deficit may be evident in some poor readers.

Finally, it might be that the Reception year is too early for Flexible Phonics work. It is possible that phonic flexibility comes into play the more children meet novel words in connected texts unsupported by teachers supplying whole-word pronunciations, and with growing knowledge of variant GPCs, experience in reading and wider language development. Further research might usefully explore all these issues.

Interpreting the absence of effects

One null finding that deserves some reflection was the absence of any reliable impact of the intervention on the oral Set-for-Variability task. Oral Set-for-Variability operated as a theorised potential mediator of the effects of Set-for-Variability instruction on word reading outcomes among a subset of children tested at post-test. Other results, such as improvements in oral Set-for-Variability in the absence of improvement in word reading or improvement in word reading with no improvement in oral Set-for-Variability, would provide direct falsification of a strong view of Set-for-Variability instruction as a necessary and sufficient cause of improvement in word reading. The absence of post-test differences in oral Set-for-Variability is hard to interpret, but leaves open the possibility that there was insufficient pedagogical activity to improve oral Set-for-Variability and thereby to cause improvements in word reading.

More broadly, in this book we have emphasised the importance of realist RCT trials that include such direct falsification (Bonell et al., 2012). While not possible here in the absence of significant effects, other future EEF trials might usefully embrace fully implemented mediation as the strongest test of a formal Theory of Change. A good example of this was the study by Vousden et al. (2022), who extended an EEF trial of ABRACADABRA to show that mooted causal forces

behind ABRACADABRA (word decoding and listening comprehension) mediated treatment effects on English national literacy assessments evident at a delayed post-test. Several benefits might accrue to the use of such mediation more broadly by trial funders such as the EEF. Selecting only strongly theorised and falsifiable interventions for funding may aid in both finding more positive effects in EEF trials than currently reported (for example Lortie-Forgues & Inglis, 2019) and the replicability of those successful trials as they more likely reflect causal mechanisms.

Another line of interpretation speaks to what can be made of the null effects of the intervention on the primary outcome. The Flexible Phonics group showed one month less progress than control schools. While this data summary of months of progress is eye-catching, it is situated in the context of wide variation either side of zero and a very small negative effect size, showing that post-test treatment group versus control group differences are unlikely to be of practical importance. In short, there was broad *equivalence across the intervention and control conditions*. From this perspective, while none of the outcome reading measures suggested positive impacts, there were also no statistically reliably negative effects either.

The reader will recall the evidence provided in Chapter 3 of the highly cautious stance of some in government to particulars of current practice in early reading instruction. Broad equivalence here tentatively shows that while Flexible Phonics was not more effective in formal terms, it was in fact equally effective. This may show there is more than one way to support early literacy, with assumptions and practices quite different to much current systematic synthetic phonics. Given there were no significant negative effects of the use of Flexible Phonics approaches, including using real books as opposed to books from reading schemes – which policy in England at the time of the trial firmly states must be avoided – an implication is that approaches linking phonics and real books in the Reception year are not detrimental to attainment. This non-detrimental effect of the intervention on phonics attainment has been reported in the literature before, outside the UK, including in well-executed systematic reviews (Jenkins et al., 2004). Its confirmation here potentially paves the way for better and more varied practices in this domain. It should of course be noted that there is always a financial cost to any professional development involving new techniques, and this is not itself evidenced by null effects in formal contrasts of treatment and control groups as shown here.

This non-effect of real book use is arguably one of several 'dogs that didn't bark in the night' (that is, effects some might expect but which did not materialise) in reported findings. For example, there is also no reason not to adopt other orders of GPC teaching (we note it has been a part of English policy for schools not to adopt alternatives to approved systematic synthetic phonics schemes unless schools can prove equivalence). Based on these findings about Flexible Phonics, schools could reasonably adopt many aspects of Flexible Phonics even under current DfE recommendations, given that Flexible Phonics does not seek to replace foundational synthetic phonics. Furthermore, most teachers in the trial found it possible to connect Flexible Phonics to their existing schemes (and to NELI) and nearly all thought it had added value for children, which speaks to the viability of these alternative pathways. Flexible Phonics is also cheap to deliver in absolute terms, as the cost analysis in Chapter 7 confirms. All this speaks to its potential usefulness in English schools, though we also note that long-term effects on attainment are yet to be established.

We should note this equivalence was achieved by Flexible Phonics in its very first year of implementation 'at scale'. Many approaches, including the current systematic synthetic phonics programmes, have needed loops of training over years, with direct hands-on support in some cases, and a wide array of resources and whole-school implementation, to achieve current levels of delivery and outcomes (see, for example, Johnson et al., 2019). This broad equivalence of reading outcomes across schools was achieved despite multiple challenges with delivery, such as the intervention being introduced mid-year during a pandemic and through remote delivery by video-conferencing. Finally, the effect was achieved in the absence of a whole-school approach to practise changes for Flexible Phonics, and with no directly observed support for – or observation of – instruction. In our previous successful scaled trials (Savage et al., 2013; Johnson et al., 2019; Bell et al., 2022), we were able to do direct coaching work with educators after having directly observed their practice. Both our wider work and the qualitative analyses here suggest these features of support may be needed to produce changed practice. Additionally, small-group or individual tutorial-style delivery of the Flexible Phonics intervention to children may also be important, as it seemed to be for NELI. There are thus many reasons for thinking that this equivalence of outcomes for Flexible Phonics and business-as-usual provision, despite all challenges, is quite an impressive achievement.

Future studies

What are the next steps in terms of basic research here? The first step concerns the analysis of the existing data from the trial. Some of us will seek to link the trial data to national Key Stage 1 teacher-assessed outcomes to explore the longer-term impacts of the intervention. This follow-up may be important, as there is now wider data suggesting that Set-for-Variability effects are strongest at delayed post-tests after the taught strategy has been available to children so that they can benefit from print experiences over time (Dunn et al., 2024; Savage et al., 2018, 2024). It is also important to explore follow-up post-tests, given (as noted in Chapter 4) that 29 intervention condition schools also received Year 1 training in Flexible Phonics and indicated that they wish to continue Flexible Phonics into Year 1 teaching. This is a modest move towards more school-wide implementation and should be evaluated. Finally, delayed-post-test analyses can be key to establishing intervention effects independent of Hawthorne and John Henry confounds, as noted in Chapter 1.

Once access to the secondary database has been secured by members of the current team, several further analyses can be undertaken. Our first suggestion utilises the word recognition subtest from the York Assessment of Reading for Comprehension (YARC). The primary outcome word recognition measure was designed to allow sub-contrasts of matched regular and irregular word reading. Given the theoretical importance of this contrast, where Flexible Phonics should uniquely enable reading of irregular words, the evaluation of this contrast is paramount to testing theory. The preliminary data showing larger effect sizes for just these words in the context of NELI, described earlier, can then be formally investigated.

The existing analyses reflect the EEF's focus on average effects for omnibus outcome measures only. There are, however, also several longitudinal analyses that can be undertaken to look at the predictive validity of the Set-for-Variability measure on later reading outcomes, including Key Stage 1 teacher assessment data made at the end of Key Stage 1 and considering possible interaction effects between Flexible Phonics and NELI. Another domain to explore is the lowest-performing 20 per cent or so of poor readers. The existing sub-analyses on children in receipt of free school meals and analyses of the lower 50 per cent on pre-test early word recognition may not index the particular needs of children in the lowest band of attainment, who are often considered the most at risk and who have also responded to interventions involving

Set-for-Variability and Direct Mapping. As we saw in Chapter 2, existing work quite consistently shows the effects of Set-for-Variability intervention among at-risk and struggling readers (Savage et al., 2018; Côté et al., 2021; Dunn et al., 2023; Basma et al., 2024; Savage et al., 2024), with effects evident on standardised reading and spelling outcome measures.

Current theory also predicts individual differences in the responsiveness of children to GPC-based interventions, moderated by baseline phonological abilities, and this pattern is supported empirically by the results of Savage et al. (2020) and Yeung and Savage (2020). For the same reasons, analogous analyses should be undertaken using baseline phonic abilities as a continuous predictor (moderator) of varied reading outcomes across treatment and control conditions; following Savage et al. (2024), post-test Set-for-Variability may predict variance in outcomes at a delayed follow-up post-test. Such analyses allow for a more nuanced exploration of when and for whom the effects of intervention are evident. This is consistent with our underlying thesis regarding the importance of acknowledging dimensionality and theorised mechanisms in realist educational RCTs (Bonell et al., 2012).

Turning to other prospective work, considering the lingering uncertainty around the interpretation of the null results given the pandemic context of this present trial, we likely need further investigations of Wave 1 (Tier 1) interventions that use Flexible Phonics content in regular classrooms. These interventions should be carried out under more typical (or optimal) conditions of delivery and implementation should be supported through face-to-face meetings in schools informed by direct observation of classroom teaching using Set-for-Variability. Under such circumstances, both implementation fidelity and implementation quality can be directly and reliably assessed. A more modestly scaled, but more detailed, realist classroom-level RCT trial (Bonell et al., 2012) is thus suggested in the first instance. A nuanced assessment of expected effects for matched irregular versus regular words, ideally through multiple measures allowing for more robust latent variable analyses, may be important for testing theory most effectively.

A very key steer from the present results given the apparent impact of NELI is to evaluate the inclusion of a more robustly delivered core vocabulary intervention as part of Flexible Phonics. There was reluctance among the delivery team – in the context of remote delivery during a pandemic – to insist on the vocabulary instruction element of the Flexible Phonics intervention, for fear it might overwhelm teachers. Suggested best practices in vocabulary instruction were, however, discussed with

teachers during the intervention training. The status of this vocabulary element may need to be more detailed and somewhat more 'mandatory' in nature (emphasised as crucial to teachers during the intervention training) given the preliminary analyses suggesting possible patterns of interaction between Flexible Phonics and the vocabulary aspect of NELI. Analyses of combined fidelities of separate phonics, vocabulary and Set-for-Variability strategy intervention sub-components may also be useful. More generally, we already have data from other studies exploring Set-for-Variability in contrast to interventions targeting morphology among poor readers (Dunn et al., 2024; Savage et al., 2024). In these studies, morphology instruction seeks to alert children to distributed word meanings of base morphemes – such as *sign* shared across a family of words such as *design*, *signature*, *resignation*, *designate* –where the base morpheme pronunciation is not always shared.

It might be argued that the updated essential core criteria for systematic synthetic phonics programmes that were published by the government in April 2021 and subsequently updated (DfE, 2023a) may pose a challenge to any future evaluation of Flexible Phonics in England. The updated criteria include elements which are at first glance similar to key features of the Flexible Phonics programme. Previously, the Set-for-Variability strategies in Flexible Phonics stood in contrast to some approaches, such as learning exception words by sight; now, systematic synthetic phonics programmes must ensure that children are taught to 'decode and spell common exception words' (DfE, 2023a, supporting documentation). Further to this, systematic synthetic phonics programmes are expected to use decodable reading material closely matched to the level of children's phonics attainment. While the Direct Mapping approach specifies that when children learn a new GPC they should read a text containing that GPC on the same day, it is possible (indeed it is to be hoped) that validated systematic synthetic phonics programmes will lead to similar practice.

While it is likely an improvement in practice to move from sight word instruction to decoding of irregular words, under the new guidance children are still not taught how to manage mismatches between pronunciation strings derived from decoding and conventional irregular word pronunciations. It is simply assumed in the DfE guidance that with the accumulation of more diverse GPC rules, irregular words will become decodable. However, as we noted in chapters 2 and 3, hundreds of candidate GPC rules are needed to map English fully, more than ever can be reasonably taught (see Gontijo et al., 2003, who suggest there are 461 GPCs in written English). What is a child to do with all the variant

GPCs? Consider the letter *a*. While the grapheme *a* is consistent in *had* (and even assuming the highly unlikely context that a child has been taught all the phoneme variants of *a*), how does a young child, when confronted with the unknown irregular word *wasp*, choose between the */a /* phoneme from the eight phoneme variants that exist in written English? (For example, consider: *about*, *after*, *taken*, *was*, *all*, *many*, *manager*, *various*). Such problems for *a* are far from being an outlier case – most GPCs have variants. In short, there clearly remains room for teaching an implementable Set-for-Variability strategy to young children to manage this spelling complexity, a strategy that also necessarily embraces word meanings.

Implementation science

Stepping beyond specifics of theory and cognitive processes in reading for this Flexible Phonics trial, we have also explored an important and more general question on implementation: working with current notions of compliance to an RCT protocol versus something more organic – 'permitted' co-construction within the intervention arm of the RCT. There were clear advantages in adopting this approach such as access, goodwill created among educators, and evidence of social validity of content. However, there were potential disadvantages – it is possible that aspects of NELI made up for gaps arising from the non-mandatory vocabulary aspects of the intervention, for example. Further work is needed under more normal delivery conditions to systematically contrast the effects of traditional notions of intervention compliance with those of flexibility-based intervention models built around differentiation (the avoidance of one-size teaching in favour of carefully tailoring teaching to meet local or individual needs). Here, we again note that differentiation within a classroom is a core evidenced feature of effective reading education (for example Dunn et al., 2024). In the Flexible Phonics trial, we sought to unite teachers' expertise in creative, effective content delivery and their knowledge of their classrooms and pupils with the expertise of university-based researchers in methodology, theory and empirical evidence regarding literacy (Boldrini et al., 2023). Greater flexibility in intervention delivery, aligned to variation in children's needs – as assessed by teachers, who are experts in local within-class delivery – does not, for us, sit comfortably with the classic medical model of fixed 'dosage' that features in many universal interventions.

Consider some of the delivery complexities in the present trial. Flexible Phonics benefits from a strong underlying theoretical conceptual and empirical model, drawn from basic research and outlined in Chapter 2. We thus know from this work that effective Set-for-Variability intervention means, at a minimum, ensuring that children:

1. have foundational *item-level* knowledge – that is, the ability to phonically blend a specific given word (itself based on foundational phonemic awareness and GPC knowledge)
2. have knowledge of item-level word meaning(s)
3. can use item Set-for-Variability
4. can bring all three above to bear on any given irregular word (for example Steacy et al., 2016; Steacy, Compton et al., 2019; Edwards et al., 2021).

Sentence-level comprehension and comprehension monitoring are also potentially involved in successful mispronunciation correction. Consider further a given irregular word – such as *pint*, *muscle*, etc – which is considered a to-be-learnt 'item'. We know from replicated basic research that both item-level and child-level particularities matter in Set-for-Variability use (Steacy, Compton et al., 2019; Steacy, Wade-Woolley et al., 2019; Edwards et al., 2021). This means that what is required from a teacher to allow a given child to achieve item-level success will always vary. For some words, it will be the GPCs, for others the vocabulary, for others mispronunciation correction as children try to bridge the varying distance between a spelling pronunciation and a conventional pronunciation. Furthermore, for some children, a word will be part of their oral (listening and/or speaking) vocabulary, whereas for others it will not. Phonics ability or another ability potentially used in Set-for-Variability (for example comprehension monitoring in sentences) varies across children and may necessitate support.

Furthermore, while some of these component abilities benefit from direct guidance from established wider science of reading research literature on what is maximally effective (for example phonic blending), some components do not similarly benefit from clear research-based item-level prescriptions about pedagogical specifics (for example item vocabulary, syntax, comprehension monitoring). Teachers are thus required to draw on their expertise and experience to decide these elements of delivery. Indeed, teachers, with their varied skill sets, will likely always play a role in devising the most engaging and effective delivery of pedagogical ideas and broad methods generated by scientific

research on reading. They are best placed to know the responsiveness of their pupils to different variants of any given approach, which points to the benefits of co-construction. It is also possible that the teaching assistant-delivered small-group and tutorial formats used in NELI will be important parts of future Flexible Phonics interventions, especially given the access issues for teaching assistants to training as noted in the implementation analyses above. It has long been known that one-to-one tutoring adds clear value in evidenced reading interventions (for example Stuebing et al., 2008).

More broadly, we note that the field has yet to fully realise the benefits of uniting what we know of the theory with what we know of the practice of teacher professional development (Collin & Smith, 2021; Basma & Savage, 2023). Theory of professional development suggests that school-wide support, use of technology, promotion of reflective practice, reporting measures of teacher change, and delivery in the context of sustained school–university partnerships (as opposed to one-time involvements) are all important to raising all aspects of pupil attainment. Models of teacher uptake of interventions suggest that evidence of early, visible value added can improve teachers' engagement with the intervention, their motivation to engage with the intervention and their sustained intervention use. Objections to interventions can sometimes be conceptual or philosophical rather than relating to the intervention specifics (for example Savage et al., 2013). We have also argued elsewhere that understanding teachers' decisions to implement interventions, together with a thick description (that is, rich detailed analysis) of teaching activity before and after an intervention, is crucial to understanding change (Savage, 2012).

More broadly speaking still, most high-quality and impactful research agendas are programmatic in nature. That is, any one individual study leads to certain insights but also further unresolved questions, typically requiring further study. The research agenda around intervention teaching of Set-for-Variability is of such a nature, where the questions described above now require empirical evidence. It is likely that policy-relevant scale-up of modified and more specifically targeted versions of any intervention will be needed as well in due course. A related question highlights the need for replication and the evidence from meta-analyses (or even mega-analyses), which can reflect both the cumulative evidence available and, via moderation analyses, some of the specifics of where, when and for whom interventions are most effective. Meta-analytic data might be more fully used to drive scaled-up EEF RCT studies. To the same end, the use of some of the multi-phase optimisation

and SMART RCT trial methodologies described in Chapter 1 – rather than (or perhaps better, as well as) the simple two-arm treatment-control RCTs most often used in EEF trials – may be more informative both of learning gains made and to policy.

In summing up the work described in this book, we suspect that there is value in viewing strongly theorised educational interventions such as this one as a particular kind of complex and inherently varied human cultural activity shaped by variation in children, words, classrooms and schools. Such variation in the maximally effective activity in the treatment arm of a trial sits firmly within an underlying randomised treatment–control condition dichotomy that must be rigorously maintained through a trial in order to illuminate possible causal paths.

The work here clearly reflects just the start of such an analysis. Much important implementation science work awaits – and, more broadly, for reading interventionists seeking to impact at scale. Impactful evidenced interventions must unite schools, universities, funders, policymakers and methodologists in a common purpose. In closing, however, we hope this book serves multiple purposes. Firstly, we have sought to illuminate for readers how EEF evaluation trials are run in a general sense and the many influences that shape a trial such as this. To this end, we have considered the details of the Flexible Phonics trial as an example case. This trial had notable atypicalities, not least in its delivery during the Covid-19 pandemic, the use of pilot studies and other features. It was also atypical in that it tested a well-developed theory and sought to advance a nuanced notion of scaled implementation, as described in this book. Our hope is that this book succeeds, firstly, in explaining how RCTs are run in EEF trials. Further, we hope it will stimulate further activity around the nuancing of strongly theorised RCT methodologies in the field of early reading, but also in other domains well beyond this, to best evidence improved outcomes for young people.

References

Achtaridou, E., Mason, E., Behailu, A., Stiell, B., Willis, B., & Coldwell, M. (2022). School recovery strategies. Department for Education: https://www.gov.uk/government/publications

Altman, D. G., Moher, D., & Schulz, K. F. (2012). Improving the reporting of randomised trials: the CONSORT Statement and beyond. *Statistics in Medicine*, *31*(25), 2985–97. https://doi.org/10.1002/sim.5402

Athey, S., & Imbens, G. (2017). The econometrics of randomized experiments. In A. J. Banerjee & E. Duflo (Eds), *Handbook of Economic Field Experiments*, pp. 73–140. Amsterdam: Elsevier. https://doi.org/10.1016/bs.hefe.2016.10.003

Basma, B., & Savage, R. S. (2023). Teacher professional development and student literacy in middle and high school: A theoretical review and meta-analysis. *Journal of Teacher Education*, *74*(3). https://doi.org/10.1177/00224871231153084

Basma, B., Savage, R., Bertone, A., & Luk, G. (2024). Reading disability in children: Exploring the N400 and its associations with Set-for-Variability. *Developmental Neuropsychology*, 1–19. https://doi.org/10.1080/87565641.2024.2418063

Bell, K., Ainsworth, H., Elliott, L., Fairhurst, C., Gascoine, L., Hanley, P. et al. (2022). ABRACADABRA (ABRA) and reading and understanding in Key Stage 1 (RUKS). Education Endowment Foundation. Accessed 5 September 2024. https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/ABRA-Report-Final.pdf?v=1725662787

Betthäuser, B. A., Bach-Mortensen, A. M., & Engzell, P. (2023). A systematic review and meta-analysis of the evidence on learning during the COVID-19 pandemic. *Nature Human Behaviour*, *7*, 375–85. https://doi.org/10.1038/s41562-022-01506-4

Biemiller, A., & Boote, C. (2006). An effective method for building meaning vocabulary in primary grades. *Journal of Educational Psychology*, *98*(1), 44–62. https://doi.org/10.1037/0022-0663.98.1.44

Bishop, D. V. M. (1997). *Uncommon Understanding: Development and disorders of language comprehension in children.* London: Psychology Press.

Bodmer, W., Bailey, R. A., Charlesworth, B., Eyre-Walker, A., Farewell, V., Mead, A., & Senn, S. (2021). The outstanding scientist, R.A. Fisher: His views on eugenics and race. *Heredity*, *126*, 565–76. https://doi.org/10.1038/s41437-020-00394-6

Boldrini, G., Fox, A., & Savage, R. S. (2023). Flexible Phonics: A complementary 'next generation' approach for teaching early reading. *Literacy*, *57*(1), 72–86. https://doi.org/10.1111/lit.12308

Bonell, C., Fletcher, A., Morton, M., Lorenc, T., & Moore, L. (2012). Realist randomised controlled trials: A new approach to evaluating complex public health interventions. *Social Science & Medicine*, *75*(12), 2299–306. https://doi.org/10.1016/j.socscimed.2012.08.032

Bowyer-Crane, C., Bonetti, S., Compton, S., Nielsen, D., D'Apice, K., & Tracey, L. (2021). The impact of Covid-19 on school starters: Interim Briefing 1 – Parent and school concerns about children starting school. Education Endowment Foundation. Accessed 5 September 2024. https://www.niesr.ac.uk/publications/impact-covid19-school-starters-interim-briefing-1-parent-and-school-concerns-about-children-starting-school?type=report

Bumbarger, B., & Perkins, D. (2008). After randomised trials: Issues related to dissemination of evidence-based interventions. *Journal of Children's Services*, *3*(2), 55–64. https://doi.org/10.1108/17466660200800012

Bryant, P., & Bradley, L. (1985). *Children's Reading Problems*. Oxford: Blackwell.

Brysbaert, M., & Biemiller, A. (2017). Test-based age-of-acquisition norms for 44 thousand English word meanings. *Behavior Research Methods*, *49*(4), 1520–23. https://doi.org/10.3758/s13428-016-0811-4

Byrne, B. (2000). *The Foundation of Literacy: The child's acquisition of the alphabetic principle*. London: Psychology Press.

Cain, T. (2015). Teachers' engagement with published research: Addressing the knowledge problem. *The Curriculum Journal*, *26*(3), 488–509. https://doi.org/10.1080/09585176.2015.1020820

Callréus, T. (2022). The randomised controlled trial at the intersection of research ethics and innovation. *Pharmaceutical Medicine*, *36*, 287–93. https://doi.org/10.1007/s40290-022-00438-8

Campbell, R. (1982). *Dear Zoo*. London: Pan Macmillan.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research.* Boston, MA: Houghton & Mifflin.

Carlgren, I. (2012). The learning study as an approach for 'clinical' subject matter didactic research. *International Journal for Lesson and Learning Studies*, *1*(2), 126–39. https://doi.org/10.1108/20468251211224172

Chandler, K. (2016). Using breakout rooms in synchronous online tutorials. *Journal of Perspectives in Applied Academic Practice*, *4*(3), 16–23. https://doi.org/10.14297/jpaap.v4i3.216

Chawla, B. (2018). Bench to bedside: Translational research demystified. *Official Scientific Journal of the Delhi Ophthalmological Society*, *29*(1), 4–5.

Cheatham, J. P., & Allor, J. H. (2012). The influence of decodability in early reading text on reading achievement: A review of the evidence. *Reading and Writing: An Interdisciplinary Journal*, *25*, 2223–46. https://doi.org/10.1007/S11145-011-9355-2

Chen, V., & Savage, R. S. (2014). Evidence for a simplicity principle: Teaching common complex grapheme-to-phonemes improves reading and motivation in at-risk readers. *Journal of Research in Reading, 37*, 196–214. https://doi.org/10.1111/1467-9817.12022

Chi, M. T. (2021). Translating a theory of active learning: An attempt to close the research–practice gap in education. *Topics in Cognitive Science*, *13*(3), 441–63. https://doi.org/10.1111/tops.12539

Cochran-Smith, M., & Lytle, S. L. (1999). Chapter 8: Relationships of knowledge and practice: Teacher learning in communities. *Review of Research in Education*, *24*, 249–305. https://doi.org/10.3102/0091732X024001249

Colenbrander, D., Kohnen, S., Beyersmann, E., Robidoux, S., Wegener, S., Arrow, T., Nation, K., & Castles, A. (2022). Teaching children to read irregular words: A comparison of three instructional methods. *Scientific Studies of Reading*, *26*(6), 545–64. https://doi.org/10.1080/10888438.2022.2077653

Colenbrander, D., Nickels, L., & Kohnen, S. (2017). Similar but different: Differences in comprehension diagnosis on the Neale Analysis of Reading Ability and the York Assessment of Reading for Comprehension. *Journal of Research in Reading*, *40*(4), 403–19. https://doi.org/10.1111/1467-9817.12075

Collin, J., & Smith, E. (2021). Effective professional development. Education Endowment Foundation. Accessed 5 September 2024. https://educationendowmentfoundation.org.uk/education-evidence/guidance-reports/effective-professional-development

Collins, L. M., Murphy, S. A., & Strecher, V. (2007). The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): New methods for more potent ehealth interventions. *American Journal of Preventative Medicine*, *32*(5), S112–S118. https://doi.org/10.1016/j.amepre.2007.01.022

Compton, D. L., Miller, A. C., Elleman, A. M., & Steacy, L. M. (2014). Have we forsaken reading theory in the name of 'quick fix' interventions for children with reading disability? *Scientific Studies of Reading*, *18*(1), 55–73. https://doi.org/10.1080/10888438.2013.836200

Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., Underwood, P., & Morrison, F. J. (2009). Individualizing student instruction precisely: Effects of child × instruction interactions on first graders' literacy development. *Child Development*, *80*(1), 77–100. https://doi.org/10.1111/j.1467-8624.2008.01247.x

Côté, M-F., Savage, R. S., & Petscher, Y. (2021). Cross-linguistic transfer of reading between English and French among Grade 1 students attending French immersion programs. *Scientific Studies of Reading*, *25*(5), 383–96. https://doi.org/10.1080/10888438.2020.1817027

Coyne, M. D., McCoach, D. B., & Kapp, S. (2007). Vocabulary intervention for kindergarten students: Comparing extended instruction to embedded instruction and incidental exposure. *Learning Disability Quarterly*, *30*(2), 74–88. https://doi.org/10.2307/30035543

Daniels, P. T., & Share, D. L. (2018). Writing system variation and its consequences for reading and dyslexia. *Scientific Studies of Reading*, *22*(1), 101–16. https://doi.org/10.1080/10888438.2017.1379082

Darnell, C., Solity, J., & Wall, H. (2017). Decoding the phonics check. *British Educational Research Journal*, *43*(3), 505–27. https://doi.org/10.1002/berj.3269

Dawson, A., Huxley, C., Gray, H., Alexander, K., Edney, C., & Bamford, S. (2024). Flexible phonics, a two-armed cluster randomised trial: Evaluation report. Education Endowment Foundation. Accessed 5 September 2024. https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Flexible_Phonics_Report_Final_2024-03-05-142402_acpo.pdf?v=1717676270

Dawson, A., Yeomans, E., & Brown, E. R. (2018). Methodological challenges in education RCTs: reflections from England's Education Endowment Foundation. *Educational Research*, *60*(3), 292–310. https://doi.org/10.1080/00131881.2018.1500079

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

DfE. (Department for Education). (2014). Phonics: Choosing a programme. London: Department for Education.

DfE (2018). Schools, pupils and their characteristics: January 2018. Statistical publication, 28 June 2018. Accessed 9 September 2024. https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2018

DfE. (2022). Academic Year 2021/22: Key Stage 1 and phonics screening check attainment. Accessed 10 September 2024. https://explore-education-statistics.service.gov.uk/find-statistics/key-stage-1-and-phonics-screening-check-attainment/2021-22

DfE. (2023a). Guidance: Systematic synthetic phonics (SSP) teaching programme validation. Accessed 5 September 2024. https://www.gov.uk/government/publications/phonics-teaching-materials-core-criteria-and-self-assessment

DfE. (2023b). Guidance: Choosing a phonics teaching programme. Accessed 5 September 2024. https://www.gov.uk/government/publications/choosing-a-phonics-teaching-programme/list-of-phonics-teaching-programmes

Diener, E., Northcott, R., Zyphur, M. J., & West, S. G. (2022). Beyond experiments. *Perspectives on Psychological Science*, *17*(4), 1101–19. https://doi.org/10.1177/17456916211037670

Donaldson, J. (1999). *The Gruffalo*. London: Macmillan.

Drake, D. A., & Ehri, L. C. (1984). Spelling acquisition: Effects of pronouncing words on memory for their spellings. *Cognition and Instruction*, *1*, 297–320. https://doi.org/10.1207/s1532690xci0103_2

Duff, D., Tomblin, J. B., & Catts, H. (2015). The influence of reading on vocabulary growth: A case for a Matthew effect. *Journal of Speech, Language, and Hearing Research*, *58*(3), 853–64. https://doi.org/10.1044/2015_JSLHR-L-13-0310

Dunn, K., Georgiou, G., Inoue, T., Savage, R., & Parrila, R. (2023). Home and school interventions aided at-risk students' literacy during Covid-19: A longitudinal analysis. *Reading and Writing*, *36*(1), 449–66. https://doi.org/10.1007/s11145-022-10354-7

Dunn, K., Georgiou, G., Savage, R., & Parrila, R. (2024). Efficacy of small group reading intervention for grade 2 and 3 struggling readers: A randomised controlled trial. *Journal of Learning Disabilities, 0*(0). https://doi.org/10.1177/00222194241297058

Dyson, H., Best, W., Solity, J., & Hulme, C. (2017). Training mispronunciation correction and word meanings improves children's ability to learn to read words. *Scientific Studies of Reading*, *21*(5), 392–407. https://doi.org/10.1080/10888438.2017.1315424

Edwards, A. A., Steacy, L. M., Seigelman, N., Rigobon, V. M., Kearns, D. M., Rueckl, J. G., & Compton, D. L. (2021). Unpacking the unique relationship between set for variability and word reading development: Examining word- and child-level predictors of performance. *Journal of Educational Psychology*, *114*(6), 1242–56. https://doi.org/10.1037/edu0000696

EEF. (Education Endowment Foundation). (2017). Improving literacy in Key Stage 1. Accessed 5 September 2024. https://educationendowmentfoundation.org.uk/education-evidence/guidance-reports/literacy-ks-1

EEF. (2019). Implementation and process evaluation guidance for EEF evaluations. Accessed 9 September 2024. https://d2tic4wvo1iusb.cloudfront.net/documents/evaluation/evaluation-design/IPE_guidance.pdf

EEF. (2022). Implementation and process evaluation guidance for EEF evaluations. Accessed 5 September 2024. https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-IPE-Guidance-August-2022.pdf?v=1692218994

EEF. (2023). Cost evaluation guidance for EEF evaluations. Accessed 5 September 2024. https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/Cost-Evaluation-Guidance-Feb_2023.pdf?v=1692218994

EEF. (2024). Evaluation design. Accessed 5 September 2024. https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

Elbro, C., & de Jong, P. F. (2017). Orthographic learning is verbal learning: The role of spelling pronunciations. In K. Cain, D. Compton, & R. Parrila (Eds), *Theories of Reading Development*, pp. 169–90. Amsterdam: John Benjamins. https://doi.org/10.1075/swll.15.10elb

Elbro, C., de Jong, P., Houter, D., & Nielsen, A.-M. (2012). From spelling pronunciation to lexical access: A second step in word decoding. *Scientific Studies of Reading*, *16*(4), 341–59. https://doi.org/10.1080/10888438.2011.568556

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh, Scotland: Oliver & Boyd.

Fletcher, J. M., Savage, R. S., & Vaughn, S. (2021). A commentary on Bowers (2020) and the role of phonics instruction in reading. *Educational Psychology Review*, *33*, 1249–74. https://doi.org/10.1007/s10648-020-09580-8

Foorman, B. R., Francis, D. J., Davidson, K. C., Harm, M. W., & Griffin, J. (2004). Variability in text features in six Grade 1 basal reading programs. *Scientific Studies of Reading*, *8*(2), 167–97. https://doi.org/10.1207/s1532799xssr0802_4

Forman, S. G., Olin, S. S., Hoagwood, K. E., Crowe, M., & Saka, N. (2009). Evidence-based interventions in schools: Developers' views of implementation barriers and facilitators. *School Mental Health*, *1*, 26–36. https://doi.org/10.1007/s12310-008-9002-5

Furlong, L., Serry, T., Bridgman, K., & Erickson, S. (2021). An evidence-based synthesis of instructional reading and spelling procedures using telepractice: A rapid review in the context of COVID-19. *International Journal of Language & Communication Disorders*, *56*(3), 456–72. https://doi.org/10.1111/1460-6984.12619

Gibson, E. J. (1965). Learning to read: Experimental psychologists examine the process by which a fundamental intellectual skill is acquired. *Science*, *148*(3673), 1066–72. https://doi.org/10.1126/science.148.3673.1066

Gibson, E. ., & Levin, H. (1975). *The Psychology of Reading*. Boston, MA: MIT Press.

Gontijo, P. F. D., Gontijo, I., & Shillcock, R. (2003). Grapheme–phoneme probabilities in English. *Behaviour Research Methods, Instruments, & Computers*, *35*(1), 136–57. https://doi.org/10.3758/BF03195506

Gonzalez-Frey, S., & Ehri, L. (2021). Connected phonation is more effective than segmented phonation for teaching beginning readers to decode unfamiliar words. *Scientific Studies of Reading*, *25*(3), 272–85. https://doi.org/10.1080/10888438.2020.1776290

Goodman, K. S. (1967). Reading: A psycholinguistic guessing game. *Literacy Research and Instruction*, *6*(4), 126–35. https://doi.org/10.1080/19388076709556976

Gorard, S., See, B. H., & Siddiqui, N. (2020). What is the evidence on the best way to get evidence into use in education? *Review of Education*, *8*(2), 570–610. https://doi.org/10.1002/rev3.3200

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*(1), 6–10. https://doi.org/10.1177/074193258600700104

Grant, A. M. (2012). Making positive change: A randomized study comparing solution-focused vs. problem-focused coaching questions. *Journal of Systemic Therapies, 31*(2), 21–35. https://doi.org/10.1521/jsyt.2012.31.2.21

Green, S. C., Bavelier, D., Kramer, A. F., Vinogradov, S., Ansorge, U., Ball, K. K., et al. (2019). Improving methodological standards in behavioral interventions for cognitive enhancement. *Journal of Cognitive Enhancement*, *3*(1), 2–29. https://doi.org/10.1007/s41465-018-0115-y

Gu, Q., & Ahn, M.Y. 2020. The research schools network: Supporting schools to develop evidence-informed practice, Evaluation report.

Gutiérrez, K. D., & Penuel, W. R. (2014). Relevance to practice as a criterion for rigor. *Educational Researcher*, *43*(1), 19–23. https://doi.org/10.3102/0013189X13520289

Hadley, E. B., Dickinson, D. K., Hirsh-Pasek, K., & Golinkoff, R. M. (2019). Building semantic networks: The impact of a vocabulary intervention on pre-schooler's depth of word knowledge. *Reading Research Quarterly*, *54*(1), 41–61. https://doi.org/10.1002/rrq.225

Hanley, R., Masterson, J., Spencer, L., & Evans, D. (2004). How long do the advantages of learning to read a transparent orthography last? An investigation of the reading skills and reading impairment of Welsh children at 10 years of age. *The Quarterly Journal of Experimental Psychology*, *57*(8), 1393–1410. https://doi.org/10.1080/02724980343000819

Hargreaves, A. (1996). Transforming knowledge: Blurring the boundaries between research, policy, and practice. *Educational Evaluation and Policy Analysis, 18*(2), 105–22. https://doi.org/10.3102/01623737018002105

Harm, M. W., McCandliss, B. D., & Seidenberg, M. S. (2003). Modeling the successes and failures of interventions for disabled readers. *Scientific Studies of Reading*, *7*(2), 155–82. https://doi.org/10.1207/S1532799XSSR0702_3

Hatcher, P. J., Hulme, C., & Ellis, A. W. (1994). Ameliorating early reading failure by integrating the teaching of reading and phonological skills: The phonological linkage hypothesis. *Child Development*, *65*(1), 41–57. https://doi.org/10.1111/j.1467-8624.1994.tb00733.x

Hatcher, P. J., Hulme, C., Miles, J. N. V., Carroll, J. M., Hatcher, J., Gibbs, S., et al. (2006). Efficacy of small-group reading intervention for beginning readers with reading delay: A randomised trial. *Journal of Child Psychology and Psychiatry*, *47*(8), 820–27. https://doi.org/10.1111/j.1469-7610.2005.01559.x

Hatcher, P. J., Hulme, C., & Snowling, M. J. (2004). Explicit phoneme training with phonic reading instruction helps young children at risk of reading failure. *Journal of Child Psychology and Psychiatry*, *45*(2), 338–58. https://doi.org/10.1111/j.1469-7610.2004.00225.x

Haynes, L., Service, O., Goldacre, B., & Torgerson, D. (2012). Test, learn, adapt: Developing public policy with randomised controlled trials. Cabinet Office, Behavioural Insights Team. https://doi.org/10.2139/ssrn.2131581

Hilte, M., & Reitsma, P. (2007). Spelling pronunciation and visual preview both facilitate learning to spell irregular words. *Annals of Dyslexia*, *56*, 301–18. https://doi.org/10.1007/s11881-006-0013-3

Hirschkorn, M., & Geelan, D. (2008). Bridging the research–practice gap: Research translation and/or research transformation. *Alberta Journal of Educational Research*, *54*(1), 1–13. https://doi.org/10.55016/ojs/ajer.v54i1.55207

Hoffmann, T., Glasziou, P., Boutron, I., Milne, R., Perera, R., Moher, D., et al. (2014). Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. *British Medical Journal*, *348*, g1687. https://doi.org/10.1136/bmj.g1687

Hulme, C., Bowyer-Crane, C., Carroll, J. M., Duff, F. J., & Snowling, M. (2012). The causal role of phoneme awareness and letter–sound knowledge in learning to read: Combining intervention studies with mediation analyses. *Psychological Science*, *23*(6), 572–7. http://doi.org/10.1177/0956797611435921

Hulme, C., Snowling, M. J., Stothard, S. E., Clarke, P., Bowyer-Crane, C., Harrington, A., Truelove, E., & Nation, K. (2009). *The York Assessment of Reading for Comprehension (YARC)*. London: GL Assessment.

Hume, D. (1748/2007). *An Enquiry Concerning Human Understanding*. Edited by P. Millican. Oxford: Oxford University Press. https://doi.org/10.1093/owc/9780199549900.001.0001

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016). Implementation and process evaluation (IPE) for interventions in education settings: An introductory handbook'. Education Endowment Foundation. Accessed 10 September 2024. https://www.researchgate.net/publication/328761292_Humphrey_N_Lendrum_A_Ashworth_E_Frearson_K_Buck_R_Kerr_K_2016_Implementation_and_process_evaluation_IPE_for_interventions_in_educational_settings_An_introductory_handbook

Husain, F., Wishart, R., Marshall, L., Frankenberg, S., Bussard, L., Chidley, S., Hudson, R., Votjkova, M., & Morris, S. (2018). Family skills: Evaluation report and executive summary. Education Endowment Foundation. Accessed 5 September 2024. https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Family_Skills.pdf?v=1692219563

Jenkins, J. R., Peyton, J. A., Sanders, E. A., & Vadasy, P. F. (2004). Effects of reading decodable texts in supplemental first-grade tutoring. *Scientific Studies of Reading*, *8*(1), 53–85. https://doi.org/10.1207/s1532799xssr0801_4

Johnson, J. L., Adkins, D., & Chauvin, S. (2020). Review of the quality indicators of rigor in qualitative research. *American Journal of Pharmaceutical Education*, *84*(1), 7120, 138–46. https://doi.org/10.5688/ajpe7120

Johnson, H., McNally, S., Rolfe, H., Ruiz-Valenzuela, J., Savage, R. S., Vousden, J., & Wood, C. (2019). Teaching assistants, computers and classroom management. *Labour Economics*, *58*, 21–36. https://doi.org/10.1016/j.labeco.2019.02.006

Joyce, K. E., & Cartwright, N. (2020). Bridging the gap between research and practice: Predicting what will work locally. *American Educational Research Journal*, *57*(3), 1045–82. https://doi.org/10.3102/0002831219866687

Kearns, D., Rogers, H. J., Koriakin, T., & Al Ghanem, R. (2016). Semantic and phonological ability to adjust decoding: A unique correlate of word reading skill? *Scientific Studies of Reading*, *20*(6), 455–70. https://doi.org/10.1080/10888438.2016.1217865

Kieran, C., Krainer, K., & Shaughnessy, J. M. (2013). Linking research to practice: Teachers as key stakeholders in mathematics education research. In M. A. Clements, A. Bishop, C. Keitel-Kreidt, J. Kilpatrick, & F. Koon-Shing Leung (Eds), *Third International Handbook of Mathematics Education*, pp. 361–92. Dordrecht: Springer International. https://doi.org/10.1007/978-1-4614-4684-2_12

Kirby, J., & Savage, R. S. (2008). Can the simple view deal with the complexities of reading? *Literacy*, *42*(2), 75–82. https://doi.org/10.1111/j.1741-4369.2008.00487.x

Kline, R. B. (2015). The mediation myth. *Basic and Applied Social Psychology*, *37*(4), 202–13. https://doi.org/10.1080/01973533.2015.1049349

Klingner, J. K., Boardman, A. G., & McMaster, K. L. (2013). What does it take to scale up and sustain evidence-based practices? *Exceptional Children*, *79*(3), 195–211. https://doi.org/10.1177/001440291307900205

Kohnen, S., Banales, E., & McArthur, G. (2020). Videoconferencing interventions for children with reading and spelling difficulties: A pilot study. *Telemedicine and e-Health*, *27*(5). https://doi.org/10.1089/tmj.2020.0061

Levesque, K. C., Breadmore, H. L., & Deacon, H. S. (2021). How morphology impacts reading and spelling: Advancing the role of morphology in models of literacy development. *Journal of Research in Reading*, *44*(1), 10–26. https://doi.org/10.1111/1467-9817.12313

Levin, H., Baum, E., & Bostwick, S. (1963). The learning of variable grapheme-to-phoneme correspondences: Comparison of English and Spanish speakers. In H. Levin, E. Gibson, A. L. Baldwin, E. Gibson, C. F. Hockett, H. Riccutti, & G. J. Suci (Eds), *A Basic Research Program on Reading: Final report on Cooperative research project no 639*, Chapter 14. US Office of Education; Department of Health, Education, and Welfare.

Levin, H., & Watson, J. (1963). The learning of variable grapheme-to-phoneme correspondences. In H. Levin, E. Gibson, A. L. Baldwin, E. Gibson, C. F. Hockett, H. Riccutti, & G. J. Suci (Eds), *A Basic research program on reading. Final report on Cooperative research project no 639*, Chapter 12. US Office of Education; Department of Health, Education, and Welfare.

Liberman, I. Y., & Shankweiler, D. (1991). Phonology and beginning reading: A tutorial. In L. Rieben & C. A. Perfetti (Eds), *Learning to Read: Basic research and its implications*, pp. 3–17. New York: Lawrence Erlbaum Associates.

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, *48*(3), 158–66. https://doi.org/10.3102/0013189X19832850

Lovett, M. W., Lacerenza, L., & Borden, S. L. (2000). Putting struggling readers on the PHAST track: A program to integrate phonological and strategy-based remedial reading instruction and maximise outcomes. *Journal of Learning Disabilities*, *33*(5), 458–67. https://doi.org/10.1177/002221940003300507

Malin, J. R., & Brown, C. (2019). Joining worlds: Knowledge mobilization and evidence-informed practice. In J. R. Malin & C. Brown (Eds), *The Role of Knowledge Brokers in Education: Connecting the dots between research and practice*, pp. 1–12. London: Routledge. https://doi.org/10.4324/9780429462436-1

Manyak, P. C., & Latka, M. (2020). Engaging vocabulary units: A flexible instructional model. *The Reading Teacher*, *73*(4), 501–12. https://doi.org/10.1002/trtr.1852

Marett, K., & Joshi, K. D. (2009). The decision to share information and rumors: Examining the role of motivation in an online discussion forum. *Communications of the Association for Information Systems*, *24*(1), #4. https://doi.org/10.17705/1CAIS.02404

Masterson, J., Stuart, M., Dixon, M., & Lovejoy, S. (2010). Children's printed word database: Continuities and changes over time in children's early reading vocabulary. *British Journal of Psychology*, *101*(2), 221–42. https://doi.org/10.1348/000712608X371744

Maxwell, B., Stevens, A., Demack, S., Coldwell, M., Wolstenholme, C., Reaney-Wood, S., Stiell, B., & Lortie-Forges, H. (2021). Review: EEF implementation and process evaluation (IPE) quality pilot. Accessed 9 September 2024. https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-evaluation-reports-and-research-papers/syntheses-of-eef-evaluations/eef-implementation-and-process-evaluation-ipe-quality-pilot

McGeown, S., Oxley, E., Ricketts, J., & Shapiro, L. (2023). Working at the intersection of research and practice: The love to read project. *International Journal of Educational Research*, *117*, #102134. https://doi.org/10.1016/j.ijer.2022.102134

McGuinness, C., McGuinness, D. & McGuinness, G. (1996). Phono-GraphixTM: A new method for remediating reading difficulties. *Annals of Dyslexia*, *46*, 73–96. https://doi.org/10.1007/BF02648172

McIntyre, D. (2005). Bridging the gap between research and practice. *Cambridge Journal of Education*, *35*(3), 357–82. https://doi.org/10.1080/03057640500319065

McNally, S., Ruiz-Valenzuela, J., & Rolfe, H. (2016). ABRA: Online reading support. Evaluation report and executive summary. Education Endowment Foundation. Accessed 5 September 2024. https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/ABRA_with_addendum.pdf?v=1692219276

Milkman, K., L., Gromet, D., Ho, H., Kay, J. S., Lee, T. W., Pandiloski, P., et al. (2021). Megastudies improve the impact of applied behavioural science. *Nature*, *600*, 478–83. https://doi.org/10.1038/s41586-021-04128-4

Moher, D., Schulz, K. F., & Altman, D. G. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel group randomized trials. *Annals of Internal Medicine*, *134*(8), 657–94. https://doi.org/10.7326/0003-4819-134-8-200104170-00011

Moir, T. (2018). Why is implementation science important for intervention design and evaluation within educational settings? *Frontiers in Education*, *3*, #61. https://doi.org/10.3389/feduc.2018.00061

Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin*, *137*(2), 267–96. https://doi.org/10.1037/a0021890

Møller, H. L., Mortensen, J. O., & Elbro, C. (2022). Effects of integrated spelling in phonics instruction for at-risk children in kindergarten. *Reading & Writing Quarterly, 38*(1), 67–82. https://doi.org/10.1080/10573569.2021.1907638

Murray, L., Wegener, S., Wang, H.-C., Parrila, R., & Castles, A. (2022). Children processing novel regular words during reading: An eye-tracking study. *Scientific Studies of Reading*, *26*(5), 417–31. https://doi.org/10.1080/10888438.2022.2030744

National Reading Panel. (2000). Teaching children to read: Reports of the subgroups. Accessed 5 September 2024. https://www.nichd.nih.gov/publications/product/247

Niemi, H. (2008). Advancing research into and during teacher education. In B. Hudson & P. Zgaga (Eds), *Teacher Education Policy in Europe: A voice of higher education institutions*, pp. 183–208. Umeå: University of Umeå.

Oakley, B., Felder, R., Brent, R., & Elhajj, I. (2004). Turning student groups into effective teams. *Journal of Student-Centered Learning*, *2*(1), 9–34.

Ofsted. (2017). Bold beginnings: The Reception curriculum in a sample of good and outstanding primary schools. Report no. 170045, November 2017. Accessed 5 September 2024. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/663560/28933_Ofsted_-_Early_Years_Curriculum_Report_-_Accessible.pdf

Olin, A., Almqvist, J., & Hamza, K. (2023). To recognize oneself and others in teacher–researcher collaboration. *Educational Action Research*, *31*(2), 248–64. https://doi.org/10.1080/09650792.2021.1897949

Parsons, S. (2021). The importance of collaboration for knowledge co-construction in 'close-to-practice' research. *British Educational Research Journal*, *47*(6), 1490–9. https://doi.org/10.1002/berj.3714

Pearl, J., & Mackenzie, D. (2018). *The Book of Why*. New York: Basic Books.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*(1), 22–37. https://doi.org/10.1080/10888438.2013.827687

Petscher, Y., Cabell, S. Q., Catts, H. W., Compton, D. L., Foorman, B. R., Hart, S. A., et al. (2020). How the science of reading informs 21st-century education. *Reading Research Quarterly*, *55*(S1), S267– S282. https://doi.org/10.1002/rrq.352

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56–115. doi.org/10.1037/0033-295X.103.1.56

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., et al. (2020). Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *New England Journal of Medicine*, *383*(27), 2603–15. https://doi.org/10.1056/NEJMoa2034577

Pritchard, S. C., Coltheart, M., Marinus, E., & Castles, A. (2018). A computational model of the self-teaching hypothesis based on the dual-route cascaded model of reading. *Cognitive Science*, *42*(3), 722–70. https://doi.org/10.1111/cogs.12571

Puzio, K., Colby, G. T., & Algeo-Nichols, D. (2020). Differentiated literacy instruction: Boondoggle or best practice? *Review of Educational Research*, *90*(4), 459–98. https://doi.org/10.3102/0034654320933536

Race, A. (2010). Science for non-scientists: Communicating the importance of child development to business and policy leaders. [PowerPoint presentation]. Cambridge, MA: Center on the Developing Child, Harvard University.

Ricketts, J., Dawson, N., & Davies, R. (2021). The hidden depths of new word knowledge: Using graded measures of orthographic and semantic learning to measure vocabulary acquisition. *Learning and Instruction*, 74, #101468. https://doi.org/10.1016/j.learninstruc.2021.101468

Robinson-Smith, L., Menzies, V., Cramman, H., Wang, Y., Fairhurst, C., Hallett, S., et al. (2019). Easypeasy: Learning through play, evaluation report. Education Endowment Foundation. Accessed 9 September 2024. https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/EasyPeasy.pdf?v=1725612781

Rose, S., Twist, L., Lord, P., Rutt, S., Badr, K., Hope, C., & Styles, B. (2021). Impact of school closures and subsequent support strategies on attainment and socio-emotional wellbeing in Key Stage 1: Interim Paper 2. Education Endowment Foundation. Accessed 5 September 2024. https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/nfer-impact-of-school-closures-and-subsequent-support-strategies-on-attainment-and-socioemotional-wellbeing-in-key-stage-1

Rosell-Aguilar, F. (2018). Twitter: A professional development and community of practice tool for teachers. *Journal of Interactive Media in Education*, 1, #6. https://doi.org/10.5334/jime.452

Royer, H. (2021). Benefits of megastudies for testing behavioural interventions. *Nature Behavioral Science, News and Views, 600*, 392–93. https://doi.org/10.1038/d41586-021-03400-x

Rueckl, J. G., Zevin, J. D., Wolf, V., & II, H. (2019). Using computational techniques to model and better understand developmental word reading disorders (i.e., dyslexia). In J. Washington & D. Compton (Eds), *Dyslexia: Revisiting etiology, diagnosis, treatment, and policy*, pp. 57–70. Towson, MD: Brookes Publishing.

Rycroft-Smith, L. (2022). Knowledge brokering to bridge the research–practice gap in education: Where are we now? *Review of Education*, *10*(1), #e3341. https://doi.org/10.1002/rev3.3341

Samarasekera, U. (2022). Martin Landray: Harnessing the power of good clinical trials. Profile. *The Lancet*, *399*(10331), P1217. https://doi.org/10.1016/S0140-6736(22)00527-X

Savage, R. S. (2012). Evidence-based reading interventions: Implementation issues for the 21st century. In B. Kelly & D. F. Perkins (Eds), *Handbook of Implementation Science for Psychology in Education*, pp. 277–97. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139013949.021

Savage, R. S. (2019). Literacy, basic processes, and interventions for struggling readers. *In Oxford Research Encyclopedia of Education*. Oxford: Oxford University Press. https://doi.org/10.1093/acrefore/9780190264093.013.925

Savage, R. S. (2022). Teaching children to read. In M. J. Snowling, C. Hulme, & K. Nation (Eds), *The Science of Reading: A Handbook*, pp. 209–34. Second edition. London: Wiley. https://doi.org/10.1002/9781119705116.ch10

Savage, R. S., Abrami, P., Piquette-Tomei, N., Wood, E., Sanghera-Sidhu, S., Burgos, G., & Deleveaux, G. (2013). A pan-Canadian cluster randomized control effectiveness trial of the ABRACADABRA web-based literacy program. *Journal of Educational Psychology*, *105*(2), 310–28. https://doi.org/10.1037/a0031025

Savage, R. S., & Fox, A. (n.d.). Evidence of the social validity of the 'BLINDED INTERVENTION' intervention. Unpublished paper.

Savage, R., Georgiou, G., Inoue, T., Dunn, K., & Parrila, R. (2024). Set-for-Variability predicts responsiveness to tier 2 reading interventions. *Scientific Studies of Reading*, 1–23. http://dx.doi.org/10.1080/10888438.2024.2418940

Savage, R.S., Georgiou, G., A. Parrila, R., & Maiorino, K. (2018). Preventative reading interventions teaching direct mapping of graphemes in texts and set-for-variability aid at-risk learners. *Scientific Studies of Reading*, *22*(3), 225–47. https://doi.org/10.1080/10888438.2018.1427753

Savage, R. S., Georgiou, G., Parrila, R., Maiorino, K., Dunn, K., & Burgos, G. (2020). The effects of teaching complex grapheme–phoneme correspondences: Evidence from a dual site cluster trial with at-risk Grade 2 students. *Scientific Studies of Reading*, *24*(4), 321–37. https://doi.org/10.1080/10888438.2019.1669607

Scanlon, D. M., & Anderson, K. L. (2020). Using context as an assist in word solving: The contributions of 25 years of research on the interactive strategies approach. *Reading Research Quarterly*, *55*(S1), 19–34. https://doi.org/10.1002/rrq.335

Scanlon, D., Schumaker, J. B., & Deshler, D. D. (1994). Collaborative dialogues between teachers and researchers to create educational interventions: A case study. *Journal of Educational and Psychological Consultation*, *5*(1), 69–76. https://doi.org/10.1207/s1532768xjepc0501_5

Schenke, W., Van Driel, J. H., Geijsel, F. P., & Volman, M. L. (2017). Closing the feedback loop: A productive interplay between practice-based research and school development through cross-professional collaboration in secondary education. *Professional Development in Education*, *43*(5), 860–80. https://doi.org/10.1080/19415257.2016.1258654

Schulz, K. F., Altman, D. G., Moher, D., & CONSORT Group. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *BMC Medicine*, *8*, #18. https://doi.org/10.1186/1741-7015-8-18

See, B. H., Gorard, S., & Siddiqui, N. (2016). Teachers' use of research evidence in practice: A pilot study of feedback to enhance learning. *Educational Research*, *58*(1), 56–72. https://doi.org/10.1080/00131881.2015.1117798

Seidenberg, M. S., Cooper Borkenhagen, M., & Kearns, D. M. (2020). Lost in translation? Challenges in connecting reading science and educational practice. *Reading Research Quarterly*, 55(S1), S119– S130. https://doi.org/10.1002/rrq.341

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*(4), 523–68. https://doi.org/10.1037/0033-295X.96.4.523

Seymour, P. H., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, *94*(2), 143–74. https://doi.org/10.1348/000712603321661859

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston, MA: Houghton-Mifflin.

Shapiro, L. R., & Solity, J. (2008). Delivering phonological and phonics training within whole-class teaching. *British Journal of Educational Psychology*, *78*(4), 597–620. https://doi.org/10.1348/000709908X293850

Share, D. L. (1995). Phonological recoding and self-teaching: *Sine qua non* of reading acquisition. *Cognition*, *55*(2), 151–218. https://doi.org/10.1016/0010-0277(94)00645-2.

Sibieta, L., Kotecha, M., & Skipp, A. (2016). Nuffield Early Language Intervention: Evaluation report and executive summary. Education Endowment Foundation. Accessed 5 September 2024. https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/EEF_Project_Report_Nuffield_Early_Language_Intervention.pdf?v=1725844485

Siddiqui, N., Gorard, S., & See, B. H. (2018). The importance of process evaluation for randomised control trials in education. *Educational Research*, *60*(3), 357–70. https://doi.org/10.1080/00131881.2018.1493349

Silver, E. A., & C. Lunsford. (2017). Linking research and practice in mathematics education: Perspectives and pathways. In J. Cai & National Council of Teachers of Mathematics (Eds), *Compendium for Research in Mathematics Education*, pp. 28–47. Reston, VA: National Council of Teachers of Mathematics.

Sims, S., Anders, J., Inglis, M., & Lortie-Forgues, H. (2020). Quantifying 'promising trials bias' in randomised controlled trials in education. Centre for Education Policy and Equalising Opportunities (CEPEO) Working Paper no. 20-16. UCL. Accessed 5 September 2024. https://EconPapers.repec.org/RePEc:ucl:cepeow:20-16

Skivington, K., Matthews, L., Simpson, S. A., Craig, P., Baird, J., Blazeby, J. M., et al. (2021). Framework for the development and evaluation of complex interventions: Gap analysis, workshop and consultation-informed update. *Health Technology Assessment*, *25*(57). https://doi.org/10.3310/hta25570

Snow, C. E. (2015). 2014 Wallace Foundation distinguished lecture: Rigor and realism: Doing educational science in the real world. *Educational Researcher*, *44*(9), 460–6. https://doi.org/10.3102/0013189X15619166

Solari, E. J., Terry, N. P., Gaab, N., Hogan, T. P., Nelson, N. J., Pentimonti, J. M., et al. (2020). Translational science: A road map for the science of reading. *Reading Research Quarterly*, *55*(Suppl1), S347–S360. https://doi.org/10.1002/rrq.357

Steacy, L. M., Compton, D. L., Petscher, Y., Elliott, J. D., Smith, K., Rueckl, J. G., Sawi, O. Frost, S. J., & Pugh, K. (2019). Development and prediction of context-dependent vowel pronunciation in elementary readers. *Scientific Studies of Reading*, *23*(1), 49–63. https://doi.org/10.1080/10888438.2018.1466303

Steacy, L. M., Elleman, A. M., Lovett, M. W., & Compton, D. L. (2016). Exploring differential effects across two decoding treatments on item-level transfer in children with significant word reading difficulties: A new approach for testing intervention elements. *Scientific Studies of Reading*, *20*(4), 283–95. https://doi.org/10.1080/10888438.2016.1178267

Steacy, L. M., Wade-Woolley, L., Rueckl, J. G., Pugh, K. R., Elliott, J. D., & Compton, D. L. (2019). The role of set for variability in irregular word reading: Word and child predictors in typically developing readers and students at-risk for reading disabilities. *Scientific Studies of Reading*, *23*(6), 523–32. https://doi.org/10.1080/10888438.2019.1620749

Stokes, L., Dawson, A., Williams, C., Alexander, K., Akehurst, G., Manzoni, C., Runge, J., & Xu, L. (2022). Tips by Text: Evaluation report. Education Endowment Foundation. Accessed 10 September 2024. https://d2tic4wvo1iusb.cloudfront.net/documents/projects/Tips-by-Text-Final.pdf?v=1651212334

Stuebing, K. K., Barth, A., Cirino, P., Francis, D., & Fletcher, J. M. (2008). A response to recent reanalyses of the National Reading Panel report: Effects of systematic phonics instruction are practically significant. *Journal of Educational Psychology*, *100*(1), 123–34. https://doi.org/10.1037/0022-0663.100.1.123

Sullivan, G., Blevins, D., & Kauth, M. R. (2008). Translating clinical training into practice in complex mental health systems: Toward opening the 'Black Box' of implementation. *Implementation Science*, *3*(1), 1–7. https://doi.org/10.1186/1748-5908-3-33

Tashakkori, A., & Teddlie, C. (2003). The past and future of mixed methods design: From data triangulation to mixed methods designs. In A. Tashakkori & C. Teddlie (Eds), *Sage Handbook of Mixed Methods in Social & Behavioral Research*, pp. 671–702. Thousand Oaks, CA: Sage.

Terry, N. P., Petscher, Y., Gaab, N., & Hart, S. (2021). Researchers translating the science of reading: Widening the lens of translational science through team science. *The Reading League Journal*, *2*(1), 46–53.

Thaler, V., Reitsma, P., & Landerl, K. (2008). Spelling pronunciations: Transforming irregularity into regularity. *Learning and Instruction*, *18*(3), 295–308. https://doi.org/10.1016/j.learninstruc.2007.10.001

Thorn, W., & Vincent-Lancrin, S. (2021). Schooling during a pandemic: The experience and outcomes of schoolchildren during the first round of COVID-19 lockdowns. Paris: OECD Publishing. https://doi.org/10.1787/1c78681e-en

Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*, 19(6), 344–57. https://doi.org/10.1093/intqhc/mzm042

Torgerson, C., Brooks, G., Gascoine, L., & Higgins, S. (2018). Phonics: Reading policy and the evidence of effectiveness from a systematic 'tertiary' review. *Research Papers in Education, 34*(2), 208–38. https://doi.org/10.1080/02671522.2017.1420816

Torgerson, C. J., & Torgerson, D. J. (2016). Randomised trials in education: An introductory handbook. Education Endowment Foundation. Accessed 5 September 2024. https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/Randomised_trials_in_education-revised031213.pdf?v=1702555641

Tse, L., & Nicholson, T. (2014). The effect of phonics-enhanced big book reading on the language and literacy skills of 6-year-old pupils of different reading ability attending lower SES schools. *Frontiers in Psychology*, *5*, #1222. https://doi.org/10.3389/fpsyg.2014.01222

Tunmer, W. E., & Chapman, J. W. (2012). Does set for variability mediate the influence of vocabulary knowledge on the development of word recognition skills? *Scientific Studies of Reading*, *16*(2), 122–40. https://doi.org/10.1080/10888438.2010.542527

UCL Division of Biosciences. (2024). Ronald Aylmer Fisher (1890–1962). Accessed 9 September 2024. https://www.ucl.ac.uk/biosciences/gee/ucl-centre-computational-biology/ronald-aylmer-fisher-1890-1962

Vadasy, P. F., & Sanders, E. A. (2021). Introducing grapheme-phoneme correspondences (GPCs): Exploring rate and complexity in phonics instruction for kindergarteners with limited literacy skills. *Reading and Writing*, *34*, 109–38. https://doi.org/10.1007/s11145-020-10064-y

Varghese, C., Bratsch-Hines, M., Aiken, H., & Vernon-Feagans, L. (2021). Elementary teachers' intervention fidelity in relation to reading and vocabulary outcomes for students at risk for reading-related disabilities. *Journal of Learning Disabilities*, *54*(6), 484–96. https://doi.org/10.1177/0022219421999844

Venezky, R. L. (1967). English orthography: Its graphical structure and its relation to sound. *Reading Research Quarterly*, *2*(3), 75–105. https://doi.org/10.2307/747031

Venezky, R. L. (1999). *The American Way of Spelling: The structure and origins of American English orthography*. New York, NY: Guilford Press.

Verhoeven, L., & Van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology*, *22*(3), 407–23. https://doi.org/10.1002/acp.1414

Vousden, J., Cunningham, A., Johnson, H., McNally S., Waldron, S., Ammi, S., Pillinger, C., Savage, R. S., & Wood, C. (2022). Decoding and comprehension skills mediate the link between a small-group reading programme and English national literacy assessments. *British Journal of Educational Psychology*, *92* (1), 105–03. https://doi.org/10.1111/bjep.12441

Vousden, J. I., Ellefson, M. R., Solity, J., & Chater, N. (2011). Simplifying reading: Applying the simplicity principle to reading. *Cognitive Science*, *35*(1), 34–78. https://doi.org/10.1111/j.1551-6709.2010.01134.x

Walker, M., Nelson, J., Bradshaw, S., & Brown, C. (2018). Teachers' engagement with research: What do we know? A research briefing. Education Endowment Foundation. Accessed 5 September 2024. https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-evaluation-reports-and-research-papers/methodological-research-and-innovations/teachers-engagement-with-research

Walker, B., Redmond, I., & Lengyel, A. (2013). Are they all the same? Lurkers and posters on the net. *eCULTURE*, *3*(1), 155–65.

Wasserstein, R., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond '$p < 0.05$'. *The American Statistician*, *73*(1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Watson, J. B. (1919). Psychology from the standpoint of a behaviorist. Philadelphia, PA: J B Lippincott. https://doi.org/10.1037/10016-000

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, *92*(4), 548–73. https://doi.org/10.1037/0033-295X.92.4.548

Wenger, E. (1998). *Communities of Practice: Learning, meaning, and identity.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511803932

Wenger, E., McDermott, R., & Snyder, W. M. (2002). Seven principles for cultivating communities of practice. In *Cultivating Communities of Practice: a guide to managing knowledge*, pp. 1–19. Boston, MA: Harvard Business School Publishing.

What Works Clearinghouse. (2016). Foundational skills to support reading for understanding in Kindergarten through 3rd Grade. Educators' Practice Guide. NCEE 2016-4008. US Department of Education, Institute of Education Sciences.

Wheatley, C., Beale, N., Wassenaar, T., Graham, M., Eldridge, E., Dawes, H., & Johansen-Berg, H. (2020). Fit to study: Reflections on designing and implementing a large-scale randomized controlled trial in secondary schools. *Trends in Neuroscience and Education*, *20*, #100134. https://doi.org/10.1016/j.tine.2020.100134

Williams, C. W., Lees-Haley, P. R., & Brown, R. S. (1993). Human response to traumatic events: An integration of counterfactual thinking, hindsight bias, and attribution theory. *Psychological Reports*, *72*(2), 483–94. https://doi.org/10.2466/pr0.1993.72.2.483

Woolf, S. H. (2008). The meaning of translational research and why it matters. *Jama*, *299*(2), 211–3. https://doi.org/10.1001/jama.2007.26

Wyse, D., & Bradbury, A. (2022). Reading wars or reading reconciliation? A critical examination of robust research evidence, curriculum policy and teachers' practices for teaching phonics and reading. *BERA Review of Education*, *10*(1), 1–53. https://doi.org/10.1002/rev3.3314

Yadav, D. (2022). Criteria for good qualitative research: A comprehensive review. *Asia-Pacific Education Research*, *31*, 679–89. https://doi.org/10.1007/s40299-021-00619-0

Yeung, S., & Savage, R. S. (2020). Teaching grapheme–phoneme correspondences using a Direct Mapping approach for at-risk second language learners: A randomized control trial. *Journal of Learning Disabilities*, *53*(2), 131–44. https://doi.org/10.1177/0022219419894563

Ziegler, J. C., Perry, C., & Zorzi, M. (2020). Learning to read and dyslexia: From theory to intervention through personalized computational models. *Current Directions in Psychological Science*, *29*(3), 293–300. https://doi.org/10.1177/0963721420915873

Zimbardo, P. G., Haney, C., Banks, W. C., & Jaffe, D. (1974). The psychology of imprisonment: Privation, power and pathology. In Z. Rubin (Ed.), *Doing unto Others: Explorations in social behavior*, pp. 61–73. Englewood Cliffs, NJ: Prentice-Hall.

Zipke, M. (2016). The importance of flexibility of pronunciation in learning to decode: A training study in set for variability. *First Language*, *36*(1), 71–86. https://doi.org/10.1177/0142723716639495

Index

In 2021-22, the Flexible Phonics reading intervention, a large-scale randomised controlled trial, took place in 118 schools in England and involved nearly 3,000 children. This study aimed to provide valuable insights into the success of largely school-based education trials. The Flexible Phonics trial proved to be not only an important experiment in improving children's literacy, but a case study in which the methodology of single randomised controlled trials in education can be considered.

Reading Randomised Controlled Trials investigates the complexities of conducting randomised controlled trials in the field of education and how they can be seen as a cultural activity. The book emphasises the human operations, decision-making, and actions that drive such trials, which are often overlooked in published reports. Through the case study of Flexible Phonics, the book highlights some of the key differences between scaled educational trials and other types of trials, such as pharmaceutical trials. The additional focus on early childhood literacy is of significant educational importance, particularly in the context of UK school partial closures in 2020 and 2021 during the Covid-19 pandemic. *Reading Randomised Controlled Trials* is ultimately a unique resource on the implementation of randomised controlled trials in authentic school contexts and for the sustained improvement of practice in education.

Robert Savage is Dean of the Faculty of Education, York University, Canada.

Amy Fox is Senior Research Fellow at the Institute for Employment Studies, Brighton.

Anneka Dawson is Head of IES Trials Unit and Pre-16 Education at the Institute for Employment Studies, Brighton.

Helen Gray is Chief Economist at Learning and Work Institute, Leicester.

Clare Huxley is Senior Research Fellow at the Institute for Employment Studies, Brighton.

Free open access version available from www.uclpress.co.uk

Image credit: Book composition with open book © Freepik.

Cover design: www.hayesdesign.co.uk

UCLPRESS

ISBN: 978-1-80008-279-3 90000

9 781800 082793