UNESCO MGIEP

unesco

Mahatma Gandhi Institute of
Education for Peace and
Sustainable Development

isee ASSESSMENT

WORKING GROUP 04

THE INTERNATIONAL SCIENCE AND EVIDENCE BASED EDUCATION ASSESSMENT

Education 2030

Published in 2022 by the United Nations Educational Scientific and Cultural Organization | Mahatma Gandhi Institute of Education for Peace and Sustainable Development, 35 Ferozshah Road, New Delhi 110001, India

ISBN 978-93-91756-04-8

The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The ideas and opinions expressed in this publication are those of the authors; they are not necessarily those of UNESCO and do not commit the Organization.

Cover illustration: Prasun Mazumdar Design | https://pmdindia.in/
Graphic design: Prasun Mazumdar Design | https://pmdindia.in/
Cover design: Prasun Mazumdar Design | https://pmdindia.in/
Illustrations: Prasun Mazumdar Design | https://pmdindia.in/
Typeset by: Prasun Mazumdar Design | https://pmdindia.in/
Printed by: Lustra Printing Press
Edited by: The Editorial Collective | http://www.editorialcollective.com.au/

Printed in New Delhi, India

The publication can be cited as: Duraiappah, A.K.,* van Atteveldt, N.M.,* Borst, G., Bugden, S., Ergas, O., Gilead, T., Gupta, L., Mercier, J., Pugh, K., Singh, N.C. and Vickers, E. A. (eds.) (2022) *Reimagining Education: The International Science and Evidence based Education Assessment.* New Delhi: UNESCO MGIEP.

* Indicates co-first editorship, as editors have contributed equally.

REIMAGINING EDUCATION: THE INTERNATIONAL SCIENCE AND EVIDENCE BASED EDUCATION ASSESSMENT

Citation: Duraiappah, A.K. and van Atteveldt, N.M.* (2022) Synopsis, Reimagining Education: The International Science and Evidence based Education Assessment. New Delhi: UNESCO MGIEP.*

**Indicates co-first authorship, as authors have contributed equally.*

THE INTERNATIONAL SCIENCE AND EVIDENCE BASED EDUCATION (ISEE) ASSESSMENT: WHY IS IT NECESSARY?

Education matters for people at all stages of life. But what is the purpose of education? This quintessential question must be asked before we can assess if our education systems are delivering on their promise. Should the goal of education be to develop human flourishing, or should it be to meet the demands of 'homo economicus'?

The way the future evolves very much depends on education. Today's mindsets on how we live, the economic and political systems we adopt, the formal and informal rules and regulations – the governance – that societies adopt, the way we perceive environmental and social problems are all very much influenced by the type (or lack) of education provided by past and present generations. The speed at which the world is changing, especially driven by technological progress and in transitioning from an industrial to a knowledge society, suggests that education can never be static and that the discourse on education, as **Dewey in 1923** asserted, 'should never come to an end'. It should be continuously evolving in response to the needs of society and the planet.

Therefore, now is the time to take stock and look ahead. A starting point is to ask two fundamental questions.

1. Are education systems serving the right purpose?

2. Are they equipped to address the pressing challenges we face today?

To answer these questions, a systematic assessment of the existing knowledge on education and learning is urgently needed. An assessment grounded in science[1] and evidence drawn from a multitude of disciplines, encompassing the entire complexity of learning and education, should consider the following:

- the goals of current education systems and their relevance to today's societal needs;

- the broad socio-political contexts in which education is embedded; and

- the state of the art for learning processes drawing from the sciences of learning.

While other reviews and reports have addressed pieces of this complex education ecosystem, a transdisciplinary approach drawing on science and evidence is urgently needed to understand the multifaceted complex education systems across the globe. The International Science and Evidence based Education (ISEE) Assessment is the first to use an integrated conceptual framework that requires the separate streams of knowledge to be integrated to answer the two overarching questions above.

Science and evidence are now widely accepted as a necessary condition for most policy-making. The success of the Intergovernmental Panel on Climate Change (IPCC) in influencing policy by bringing the best science and evidence to the table has been instrumental in shaping climate change policy. However, the road has not been smooth, with many

[1] We define science as the pursuit and application of knowledge and understanding of the natural and social world following a systematic methodology based on evidence (The Science Council, https://sciencecouncil.org/about-science/our-definition-of-science/).

critics questioning the validity of the science and the evidence provided. The same can be said of the Millennium Ecosystem Assessment, which brought to the fore the power of multidisciplinary science and evidence in informing policy-making for the sustainable use of biodiversity and ecosystem services for the well-being of humanity.

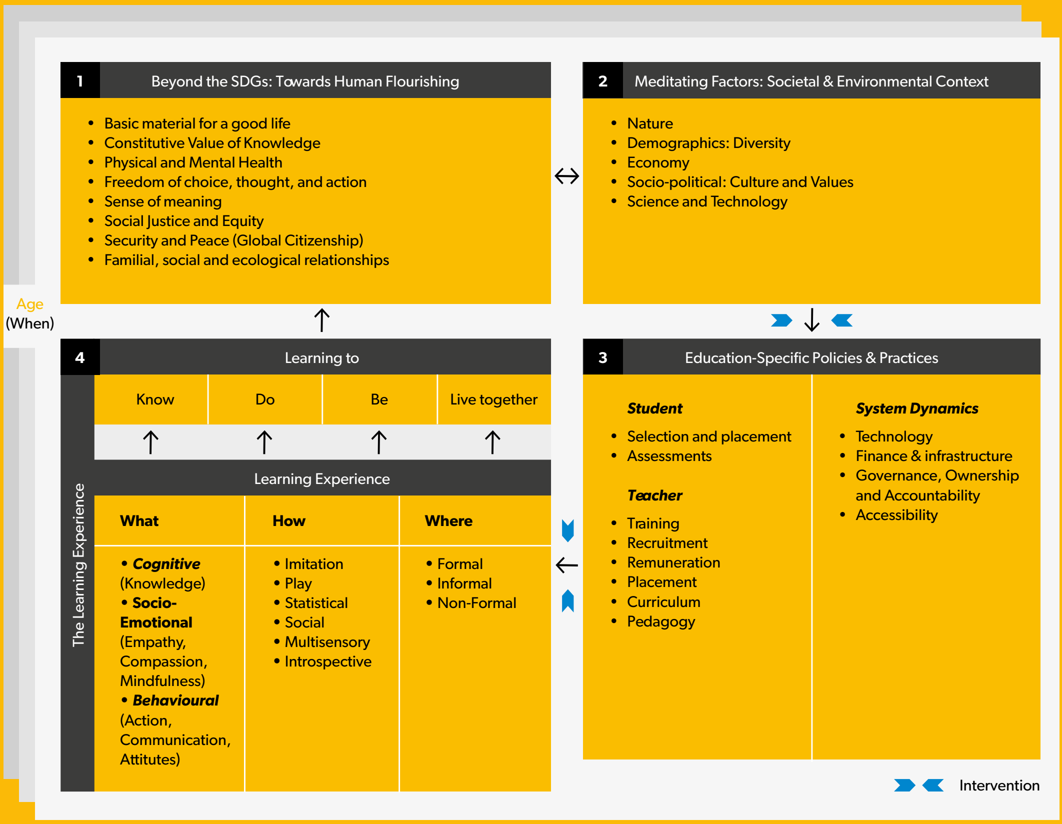
The field of education is no different. However, unlike in the environmental field, no previous attempts have been made to undertake an integrated transdisciplinary international assessment of science and evidence in the field of education. Education policy has been widely influenced by anecdotal information and is seldom backed up by transdisciplinary consensus science and evidence. However, our knowledge of learning processes and their bidirectional relationship with their contexts is rapidly increasing due to advancements in all disciplines addressing educational issues, and particularly over the past two decades by research from the field of mind, brain and education. But the exchange of knowledge and information across the various disciplines working on education is challenging, as is the translation of new findings from this transdisciplinary research into educational policy.

Recognizing the need for, but absence of, a transdisciplinary approach to education and the limited use of science and evidence in education policy-making further strengthens the need for the ISEE Assessment. The term 'assessment' here refers to a critical evaluation of the state of existing knowledge on education and learning by a team of independent experts drawn from a broad range of relevant disciplines and from across the world. The knowledge base is peer-reviewed scientific literature, but also includes credible grey literature. The Assessment report consists of 25 chapters, which have undergone a blind peer-review process. It assesses findings from across disciplines through deliberative discussions amongst the team of diverse experts throughout the project. The accompanying Summary for Decision-Makers (SDM) addresses overarching key questions and translates the answers into policy-relevant recommendations. In addition, the Assessment highlights gaps in knowledge and suggests potential future research agendas. To be clear, the ISEE Assessment is of a very different nature from international large-scale student assessments, such as the Programme for International Student Assessment (PISA). Assessments like the one we present here have proved extremely fruitful in other domains (e.g. IPCC) to synthesize information available from a wide range of disciplines. This has never before been performed for education.

THE ISEE ASSESSMENT CONCEPTUAL FRAMEWORK AND STRUCTURE

The ISEE Assessment launched in September 2019 with an expert meeting hosted by the Chief Scientist's Office, Quebec, Montreal and included approximately 20 scientists from around the world. Expertise was drawn from a range of education-related disciplines, such as international comparative education, human developmental and education psychology, neuroscience, cognitive science, economy and philosophy. This group gathered over three days to deliberate if an assessment of education would be beneficial, what it could contribute to education and what should be the conceptual framework. Although there were many disagreements among the experts, two common findings emerged: the need for an assessment of this nature; and the need for a transdisciplinary, multicultural and multiperspective lens to rethink the education agenda for the twenty-first century.

Developing a conceptual framework is an essential first step when undertaking an assessment of this nature. The



1 Beyond the SDGs: Towards Human Flourishing

- Basic material for a good life
- Constitutive Value of Knowledge
- Physical and Mental Health
- Freedom of choice, thought, and action
- Sense of meaning
- Social Justice and Equity
- Security and Peace (Global Citizenship)
- Familial, social and ecological relationships

2 Meditating Factors: Societal & Environmental Context

- Nature
- Demographics: Diversity
- Economy
- Socio-political: Culture and Values
- Science and Technology

Age (When)

4 Learning to

Know | Do | Be | Live together

The Learning Experience

Learning Experience

What	How	Where
• ***Cognitive*** (Knowledge) • **Socio-Emotional** (Empathy, Compassion, Mindfulness) • ***Behavioural*** (Action, Communication, Attitutes)	• Imitation • Play • Statistical • Social • Multisensory • Introspective	• Formal • Informal • Non-Formal

3 Education-Specific Policies & Practices

Student

- Selection and placement
- Assessments

Teacher

- Training
- Recruitment
- Remuneration
- Placement
- Curriculum
- Pedagogy

System Dynamics

- Technology
- Finance & infrastructure
- Governance, Ownership and Accountability
- Accessibility

Intervention

Figure 1. The ISEE Assessment Conceptual Framework of lifelong learning

ISEE Assessment Conceptual Framework (CF) aims to capture the key interlinkages between critical components of the education and learning system as understood by the education community represented by the group of experts convened at the first expert workshop. The CF presented above provides the basis for understanding and unpacking the complexity of the knowledge on education and learning across the world.

WORKING GROUP 1: EDUCATION AND HUMAN FLOURISHING

Working Group 1 on human flourishing unpacks **Box 1** and explores the interdependency between **Boxes 1 and 4** in the CF. **Chapter 1** provides an overview of the working group and the rationale for the chapters presented in the volume. **Chapter 1** also evaluates the concept of human flourishing and explores whether a definition can be used in education systems that allows context-sensitivity but still offers a common set of parameters. A main finding is that any education system for the future must acknowledge that volatility, uncertainty, complexity and ambiguity are central characteristics of our world, and education systems must rise to meet these challenges. **Chapter 2** reports that since the Second World War, educational policy and, in particular, education's role in human development has advanced along two parallel tracks with the dominant pathway focusing on the economy, while the other track, which takes a broader humanistic view emphasizing non-economic and non-instrumental objectives for human flourishing, is relegated. **Chapter 3** presents recent advances in cognitive and affective science that demonstrate the skills associated with flourishing can be cultivated through education, in the same way as literacy and numeracy. The chapter also outlines that about 82 per cent of teachers in teacher surveys consider there is a disproportionate focus on exams in education in contrast to the well-being of students. A similar observation emerged with 73 per cent of parents preferring to send their children to a school where they would be happy even if their exam results were not as good as those achieved in high-stress exam oriented schools. Most students (81 per cent) indicated they wanted to learn more about how to look after their mental well-being.

Chapter 4 presents some perspectives and suggestions on curriculum, assessment and teaching reforms towards an education for flourishing following six curricular domains and six learning trajectories: learning to know and think, learning to do and evaluate, learning to learn, learning to live together, learning to live with nature and learning to be and become. This chapter recommends a slight adaptation of UNESCO's four pillars of education by introducing two additional pillars to equip education systems to better address today's societal and environmental challenges. Chapter 5 completes the work of this working group by providing recommendations for strengthening schools towards an education for flourishing based on an assessment of existing school practices and environments.

WORKING GROUP 2: EDUCATION AND CONTEXT

Working Group 2 on contexts aims to understand how our social, economic and political systems influence, and are influenced by, our education systems **(the interdependent link between Box 2 and Box 3 in the CF)**. Furthermore, it examines how these contextual factors relate to diverse conceptions of the purpose of education **(the interdependent link between Box 1 and Box 2)**. The first four chapters look at the macro level: the social, political, economic and environmental contextual factors the group considers as having a critical influence in the design of education systems across the globe. The group looked at the political economy of education, as well as how global social phenomena such as colonialism and more recently climate change and sustainability issues have influenced education systems. These chapters look at how equitable education systems have been over the past fifty years and develop interesting insights into how meritocracy – frequently touted today as the great equalizer – actually threatens the equity and sustainability of education systems, fuelling acute competitive intensity and narrowing the experience of learning for millions. The concept of 'hereditary meritocracy' is shown to be a rising trend among Ivy League educational institutions in the United States, where the majority of the students are from the top 1 per cent of the income distribution while a minority come from households in the bottom 60 per cent. In addition, the chapter informs how socio-economic disparities affect the learning of the over 1 billion children who are impacted by poverty.

Chapter 2 on environmental contexts highlights the limitations of approaches to 'education for sustainable development', given that education remains wedded to a fundamentally human capital oriented vision looking at nature purely from an instrumentalist view rather than as an existential and intrinsic element of human flourishing. An important dimension in today's education systems is the notion of conflict and its implications for education. **Chapter 5** reports that the psychological impact of conflict (and related, trauma and poverty) on learning is huge and that, as far as possible, education systems must recognize and accommodate these impacts when designing curriculum, assessments and teacher training. Approximately 37 per cent of primary school aged refugee children are out of school, while only 24 per cent have access to secondary education and a dismal 3 per cent to higher education. Both **Chapters 5 and 8** (on curriculum) stress the role that education can and often does play in causing conflict, through fostering intolerance, xenophobia and societal division.

Chapters 6 and 7 of Working Group 2 then address the nature and extent of recent advances in neuroscience and technology as these relate to education, assessing how developments in these fields have both influenced, and have been influenced by, contextual factors (political, commercial, cultural, etc.). The final set of three chapters assesses how contexts have shaped, and are shaped by, key institutional features of our education systems that include curriculum and pedagogy **(Chapter 8)**, assessment **(Chapter 9)** and the teaching profession **(Chapter 10)**. These chapters elaborate how curriculum, assessment and teacher training are influenced by the political, social and economic climate in which education systems are embedded. Taken as a whole, the analysis presented in **Working Group 2**, while underlining the crucial importance of education in today's world, also reminds us of education's darker aspects (e.g. its potential to fuel conflict, as well as ameliorate it) and of its limitations as a resource for solving the world's problems if the contextual factors are not aligned towards peace and sustainability. A key conclusion is the need to balance hope in education's transformative potential with awareness that fully realizing its capacity to promote human flourishing requires far-reaching changes in our political and socio-economic order.

WORKING GROUP 3: EDUCATION AND THE LEARNING EXPERIENCE

Working Group 3 on the learning experience assesses the relationship between the 'what', 'how', 'where' and 'when' of learning, and how they relate to UNESCO's pillars of education, in light of state-of-the-art evidence from the science of learning, and studies of the socio-economic, environmental and other challenges we face today **(the interdependent links between Box 4 with Boxes 3 and 1 in the CF)**. Building on the definition of education and learning as a 'relational' process **(Working Group 1)** and insights from brain imaging studies, the role of social and emotional learning (SEL) is incorporated into all four aspects of learning. **Chapter 4** on social and emotional foundations of learning highlights that the learning experience at the individual level is intrinsically cognitive, emotional and social, as there is no clear dissociation between cognitive and emotional functions of the brain; rather learning occurs from the interconnectedness of neural networks across many functions. The chapter reports that although SEL improves learning outcomes by 7 to 11 per cent, it only constitutes about 7 and 4 per cent of learning in primary and secondary education respectively.

Chapter 2 on brain development and maturation highlights the non-linear nature of brain development and learning as a result of a lifelong dynamic and mutually interacting interplay between nature and nurture, contrary to the long-held belief in the competing forces between biology and culture. Although the themes of individual differences and learning differences overlap to some extent, experts from **Working Group 3** strongly felt that separate chapters on individual differences and learning differences and disabilities were needed. Therefore, **Chapter 3** provides new evidence demonstrating that individual differences in human development and learning arise from reciprocal interactions between biological, psychological and sociological factors. It calls for an integrated multidisciplinary approach to the study of human development, and its conceptualization in education. **Chapter 4** provides details of SEL, what it entails and offers to the learning experience. The chapter underscores the high returns to investment in SEL and its contribution to not only academic achievement but also to social issues such as bullying, substance abuse, aggression, and depression, among others. **Chapter 5** emphasizes the importance of building a strong foundation of academic skills, such as literacy and numeracy, to scaffold other skills and develop flourishing. This underscores the importance of the integration of SEL with the more traditional competencies of literacy and numeracy within education systems to reach for human flourishing, which we call the 'whole-brain approach'. The chapter also emphasizes the importance of mother tongue instruction in the first formative years before second languages are introduced to achieve the best possible learning outcomes while highlighting the findings of the **2016 UNESCO Global Monitoring Report** that about 40 per cent of the global population does not have access to instruction in the language they understand.

Chapter 6 raises important questions relating to inclusive education versus special needs education and presents findings suggesting that care should be taken when designing inclusive education policies. Emphasizing that one in every five to ten children expresses some form of learning difference such as dyslexia or dyscalculia, it highlights that particular attention should be given to disabilities that are invisible but significantly affect learning. About 40 per cent of countries do not collect data on prevalence, school attendance and school completion for students with disabilities/differences, limiting informed and effective policy-making to close gaps in access and learning under the inclusive education umbrella. The call for universal, preventive screening emerges as a clear policy recommendation, while also recognizing that careful implementation is essential. **Chapter 7** addresses 'where we learn' and explores how built spaces, natural spaces and digital spaces affect learning. It looks at the roles of these different kinds of spaces for learning, attainment, interpersonal relationships, skills development, well-being and behaviours across UNESCO's four pillars of education. The chapter also explores how learning spaces can be actively shaped, felt and understood through practices and policies that occur within and around them.

WORKING GROUP 4: EDUCATION - DATA AND EVIDENCE

The ISEE Assessment was initiated with the idea of using science and evidence as its founding pillars. However, we soon noticed that the terms evidence and data prompted a slew of questions and clarifications that we did not anticipate. Recognizing the diversity of views and perspectives of what a science and evidence based assessment means, a small group of experts was commissioned to provide more clarity and guidance on what evidence means and how data can and should be used in education practice and policy-making. This working group's focus is on seeking the best way to provide answers to the questions: what worked?; what is working best generally?; and will a given intervention work here and now? A new taxonomy of eight tiers or levels of evidence guides matching available evidence to these questions and assess the strength of this evidence. The experts in this group provide a deeper understanding of how effect size and consistency of effect sizes influence learning outcomes, and how they can − and cannot − be used in practice and policy guidance. They also illustrate the potential of this modern approach to evidence based education by discussing the EEF (Education

Endowment Fund) Evidence Database, effectively providing a proof of concept regarding some of the key ideas put forward as the new norm.

Working Group 4, in particular **Chapter 3**, highlights the importance of understanding and interpreting uncertainty. The concepts of p-values and statistical significance, together with confidence intervals, are explained and recommended as the new standard practice to be used when presenting empirical evidence in support of practice and policy-making. The core finding from **Working Group 4** is that science and evidence based education practice and decision-making are evolving into a more complex set of questions, but are potentially very fruitful undertakings, for which it is key to understand the limitations of extant data and evidence in striving to create, obtain and use recent evidence. A clear and transparent discourse surrounding the assumptions and caveats in the analysis should always be provided so that practitioners and decision-makers are aware of limitations and uncertainties.

GOVERNANCE AND SOCIAL PROCESS OF THE ISEE ASSESSMENT

The ISEE Assessment is a first of its kind for the field of education. Most studies reviewing education and learning primarily take a single disciplinary lens with very little collaboration, especially across traditional educational study disciplines and the newer science of learning disciplines. A key component for a successful endeavour of this nature is mutual respect and acceptance of multiple perspectives and a culture of ‘agree to disagree’. In addition, an open culture is needed in which experts keep an open mind, truly listen to others and are fearless in asking questions to ensure transparency in assumptions and terminology. Finally, there must be a process in place to facilitate consensus building across all experts in order to create a synthesis of findings to be used by policy-makers. Achieving the above will strengthen education systems and facilitate learning for the benefit of the individual and society.

An Advisory Board guided by two co-chairs was formed, comprising eminent persons from academia, business and policy, to provide support and guidance to the Assessment. The primary function was to ensure the relevance and credibility of the Assessment exercise. The overall scientific work of the Assessment was guided by the two Assessment co-chairs, one from the social sciences and the other from the natural sciences. The primary responsibility of the Assessment co-chairs was to ensure smooth collaboration across the various disciplines within and across working groups and to ensure the strictest scientific rigour was applied to the Assessment exercise. The co-chairs also were responsible for synthesizing the Assessment findings in the SDM document and a shorter headliners document that conveys the key messages and policy recommendations from the ISEE Assessment.

Each working group had two senior co-chairs supported by a junior co-chair, always combining experts from traditional educational studies and the sciences of learning community. Recruitment for these positions was a non-trivial process. Many early invitations were politely rejected because the work was outside those individuals' comfort zones, as well as requiring them to find common ground and come to shared consensual conclusions with experts and scientists outside their own communities and bubbles. This in itself was an important finding as a new social contract for education is designed and implemented by member countries in response to UNESCO's Futures of Education report released in November 2021.

Once the group leaders were identified, the arduous process of identifying the authors and structure of the chapters for the various working groups took place. The tendency to identify familiar faces and colleagues was only natural and therefore stringent requirements for each chapter to ideally have at the minimum two disciplines represented were established, alongside the strong recommendation to reach a representative author team in terms of geographic location and gender. However, the process was not always perfect and sometimes a chapter has leaned further towards a particular discipline or perspective than we ideally would have liked.

In order to minimize disciplinary bias but also to ensure scientific credibility, a blind peer-review process was put in place. Review editors, again from different disciplines, were identified to oversee the review process to ensure legitimacy, credibility and the optimal selection of the most appropriate reviewers for each of the chapters across all four working groups. The secretariat overseeing the logistics of the assessment was responsible for compiling the review comments and supporting the review editors to ensure all comments were adequately addressed by the respective chapter authors before they were approved for publication.

THE OUTPUTS

The results of the ISEE Assessment are presented in four volumes, each presenting the findings from each of the four working groups. As mentioned earlier, three working groups present state-of-the-art knowledge on education and learning based on the CF developed for the ISEE Assessment, and one on the meaning and use of data and evidence. Needless to say, there are many interlinkages across these working groups and attempts have been made to insert cross-references where necessary.

The SDM is an essential output from the ISEE Assessment. The SDM is presented not as a summary of each working group, but as a synthesis across all the working groups. The SDM is structured along five key questions of relevance for policy-makers. This involved ‘harvesting’ the answers to each question from all four volumes and presenting them in an integrated fashion that reflects the complexity and interconnectedness among the various components within the education sector. The SDM presents the overarching key messages, findings and recommendations that emerge from the full ISEE Assessment report.

A headliners document forms part of the overall package, providing a brief overview and reflecting the key take-home messages and policy recommendations. It is meant to offer a snapshot of the ISEE Assessment and is a quick reference primarily for decision-makers and policy-makers.

CONCLUDING REMARKS

The ISEE Assessment is a first for the education sector. It brings together a critical mass of experts and scientists working in the field of education. The process of bringing together over 300 experts and scientists from a range of disciplines has been a challenging task but offers an exciting learning experience of transdisciplinary collaboration within education. The two-and-a-half year journey produced new insights but, more importantly, provides the basis for future such assessments. The assessment process and the findings suggest that transdisciplinary research and collaboration is a necessary condition for any education policy-making, especially at the global level. The insights emerging when a range of disciplines combine their relevant research and perspectives are invaluable, offering understandings that sometimes contradict conventional intuitions. It is also important to emphasize the process of consensus building among experts coming from multiple disciplines on findings which might be controversial or uncertain.

This first assessment highlights the richness of evidence and data on learning and education systems, but it also demonstrates how fragmented and compartmentalized these are across the world. Another key observation from the Assessment is that many of the experts and scientists were uncomfortable assigning confidence levels to the findings and the subsequent recommendations. This will need attention if we are to ground the science of learning into education policy-making. An international science organization representing multiple disciplines with a mandate on education should ideally carry out an assessment like the ISEE Assessment periodically in the future.

In 2021 UNESCO called for a new social contract in ‘Reimagining Our Futures together: A New Social Contract for Education’. We are optimistic that the take-home messages, key findings and policy recommendations put forward by the ISEE Assessment will guide countries across the globe when designing the blueprint for this new social contract. An education for human flourishing using a whole-brain, learner-centric approach acknowledges the interconnectedness between cognitive, social and emotional dimensions, and how these are influenced heavily by societal and contextual factors. Furthermore, recognizing and understanding the vast individual differences in development and learning is key when designing any social contract on education in any part of the world.

AUTHORS

Abigail Hackett,
Manchester Metropolitan University
a.hackett@mmu.ac.uk

Adam Wood,
University College London
a.wood@ucl.ac.uk

Adriano Linzarini,
UNESCO MGIEP
adriano.linzarini@protonmail.com

Akiyoshi Yonezawa,
Tohoku University
akiyoshi.yonezawa.a4@tohoku.ac.jp

Alaidde Berenice Villanueva Aguilera, Durham University
alaidde.b.villanueva@durham.ac.uk

Alejandra Cristina Mizala,
Universidad de Chile
amizala@uchile.cl

Alejandrina Cristia,
Ecole normale supérieure (ENS)
alecristia@gmail.com

Alex Wilson,
University of Saskatchewan
alex.wilson@usask.ca

Alida Hudson,
The Ohio State University,
hudson.634@osu.edu

Amber Gove,
RTI International
agove@rti.org

Amy Ellis-Thompson,
Education Endowment Foundation
amy.ellis-thompson@eefoundation.org.uk

Anantha K. Duraiappah,
UNESCO MGIEP
ak.duraiappah@unesco.org

Andrew Fuligni,
University of California, Los Angeles
afuligni@ucla.edu

Angela H. Gutchess,
Brandeis University,
gutchess@brandeis.edu

Angela Page,
University of Newcastle
apage1@newcastle.edu.au

Angus Hikairo Macfarlane,
University of Canterbury
angus.macfarlane@canterbury.ac.nz

Ann Dowker,
University of Oxford ann.dowker@psy.ox.ac.uk

Anna Bull,
University of York
anna.bull@york.ac.uk

Anna H. Miller,
Vanderbilt University
anna.h.miller@vanderbilt.edu

Anna Hickey-Moody,
RMIT University
anna.hickey-moody@rmit.edu.au

Anna Lucia Campos,
Educational Association for Human Development and IMCE - Instituto Mente, Cerebro & Educación
acampos@imce.la

Anne Castles,
Macquarie University
anne.castles@mq.edu.au

Annelinde Vandenbroucke,
Leiden University
a.r.e.vandenbroucke@fsw.leidenuniv.nl

Annouchka Bayley,
University of Cambridge
acb218@cam.ac.uk

AUTHORS

Antoni Verger Planellsdni,
Universitat Autònoma de
Barcelona antoni.verger@uab.cat

Anya Chakraborty,
UNESCO MGIEP
anya.c.calcutta@gmail.com

Arjen Wals,
Wageningen University,
arjen.wals@wur.nl

Arnaud Cachia,
LaPsyDÉ, Université de Paris
arnaud.cachia@parisdescartes.fr

Arthur Graesser,
University of Memphis
graesser@memphis.edu

Baiba Martinsone,
University of Latvia
baiba.martinsone@lu.lv

Bassel Akar,
Notre Dame University
bassel.akar@gmail.com

Beatrice Ávalos-Bevan,
Universidad de Chile
bavalos254@gmail.com

Ben Williamson,
University of Edinburgh
ben.williamson@ed.ac.uk

Berna Güroğlu,
Leiden University
bguroglu@@fsw.leidenuniv.nl

Betony Clasby,
University of Sheffield
beclasby1@sheffield.ac.uk

Bilen Mekonnen Araya,
Queen's University
18bma7@queensu.ca

Bowen Paulle,
University of Amsterdam
b.paulle@uva.nl

Brianna Doherty,
University of California,
San Francisco
brianna.doherty@ucsf.edu

Carlo Perrotta,
Monash University
carlo.perrotta@monash.edu

Carmel Cefai,
University of Malta
carmel.cefai@um.edu.mt

Carmen Strigel,
RTI International
cstrigel@rti.org

Carolien Rieffe,
Leiden University;
University of Twente, Enschede;
and University College London
crieffe@fsw.leidenuniv.nl

Catherine Burke,
University of Cambridge
cb552@cam.ac.uk

Catherine Elizabeth Draper,
University of the Witwatersrand
catherine.draper@wits.ac.za

Catherine McBride,
Chinese University of Hong Kong
cammiemcbridechang@gmail.com

Catriona O'Toole,
Maynooth University
Catriona.a.otoole@mu.ie

Caylee Cook,
University of the Witwatersrand
caylee.cook@wits.ac.za

Chaise LaDousa,
Hamilton College
cladousa@hamilton.edu

Charles H. Hillman,
Northeastern University
c.hillman@northeastern.edu

Cheryl Teelucksingh,
Ryerson University
teeluck@ryerson.ca

Chi Zhang
The Hong Kong University
of Science and Technology
(Guangzhou)
chizhang@ust.hkn

Christine Horn,
RMIT University
christine.horn@rmit.edu.au

Christopher Boyle,
The University of Adelaide
chris.boyle@adelaide.edu.au

Clancy Blair,
New York University,
clancy.blair@nyu.edu

Clarence W. Joldersma,
Calvin University
cjolders@calvin.edu

Colter Mitchell,
University of Michigan
cmsm@umich.edu

Crystal Day-Hess,
University of Denver
crystal.day-hess@du.edu

Dale L. Albanese,
National Sun Yat-sen University
albanese@mail.nsysu.edu.tw

Daniel Johnson Mardones,
Universidad de Chile
djohnson@educarchile.cl

Daphne Oluwasen Martschenko,
Stanford University
daphnem@stanford.edu

Dave L. Edyburn,
University of Wisconsin –
Milwaukee
edyburn@uwm.edu

David A.G. Clarke,
University of Edinburgh
david.clarke@ed.ac.uk

David Stovall,
University of Illinois at Chicago
dostoval@uic.edu

Denis Le Bihan,
French Alternative Energies and
Atomic Energy Commission
denis.lebihan@gmail.com

Doret de Ruyter,
University of Humanistic Studies
d.deruyter@uvh.nl

Douglas H. Clements,
University of Denver, Denver
douglas.clements@du.edu

Dusana Dorjee,
University of York
dusana.dorjee@york.ac.uk

Dzulkifli Abdul Razak,
International Islamic University
Malaysia (IIUM)
dzulrazak@iium.edu.my

Edward A. Vickers,
Kyushu University
edvickers08@googlemail.com

Efrat Eilam,
Victoria University
efrat.eilam@vu.edu.au

Elena L. Grigorenko,
University of Houston, Houston
elena.grigorenko@times.uh.edu

Eleni A. Kyza,
Cyprus University of Technology
eleni.kyza@cut.ac.cy

Elsje van Bergen,
VU University Amsterdam
e.van.bergen@vu.nl

AUTHORS

Emma Sian Dobson,
Durham University
e.s.dobson@durham.ac.uk

Eric Herring,
University of Bristol
eric.herring@bristol.ac.uk

Erica Southgate,
University of Newcastle
erica.southgate@newcastle.edu.au

Fikile Nxumalo,
University of Toronto
f.nxumalo@utoronto.ca

Franklin N. Glozah,
University of Ghana
fglozah@ug.edu.gh

Gabrielle Ivinson,
Manchester Metropolitan University
g.ivinson@mmu.ac.uk

Gail Rosenbaum,
New York University
gailrosenbaum@nyu.edu

Grégoire Borst,
LaPsyDÉ, Université de Paris
gregoire.borst@parisdescartes.fr

Gregory Mannion,
University of Stirling
greg.mannion@stir.ac.uk

Guillermo Rodriguez-Guzman,
Center for Homelessness Impact
guillermo@homelessnessimpact.org

Hannah Blausten,
Education Endowment Foundation
hannah.blausten@eefoundation.org.uk

Harry Daniels,
University of Oxford
harry.daniels@education.ox.ac.uk

Harry Madgwick,
Education Endowment Foundation
harry.madgwick@eefoundation.org.uk

Heather Aldersey,
Queen's University
hma@queensu.ca

Heila Lotz-Sisitka,
Rhodes University
h.lotz-sisitka@ru.ac.za

Hilary F. Emery,
University of Oxford,
Formerly CEO of National Children's Bureau
hilary.emery@btconnect.com

Hope Kent,
University of Exeter
hnk201@exeter.ac.uk

Hsu-Chan Kuo,
National Cheng Kung University
kentcre8@mail.ncku.edu.tw

Indrajit Gunasekara,
University of Hawaii-West Oahu
indrajit@hawaii.edu

Inny Accioly,
Fluminense Federal University
innya@id.uff.br

Iris Bourgault Bouthillier,
Université du Québec à Montréal (UQAM)
bourgault_bouthillier.iris@uqam.ca

Jacqueline Specht,
Western University
jspecht@uwo.ca

Jae Hyung Park,
The Education University of Hong Kong
jpark@eduhk.hk

James Tonks,
University of Exeter
j.tonks@exeter.ac.uk

Jamie Mcphie,
University of Cumbria
jamie.mcphie@cumbria.ac.uk

Jason Ritter,
Duquesne University
ritterj@duq.edu

Jelena Obradović,
Stanford University
jelena.obradovic@stanford.edu

Jenny Gibson,
University of Cambridge
jlg53@cam.ac.uk

Jessica Dubois,
Université de Paris
jessica.dubois@inserm.fr

Jessica Pykett,
University of Birmingham
j.pykett@bham.ac.uk

Jiaxian Zhou,
East China Normal University
jxzhou@psy.ecnu.edu.cn

Jill Blackmore,
Deakin University
jillian.blackmore@deakin.edu.au

Jing-Jyi Wu,
National Chengchi University
jjwu@nccu.edu.tw

Jo Hickman Dunne,
Lancaster University
jo.hickmandunne@youthimpact.uk

Jo Rey,
Macquarie University
jo.rey@mq.edu.au

Jo Van Herwegen,
University College London
j.vanherwegen@ucl.ac.uk

Joan Y. Chiao,
Northwestern University
jc@cngmh.org

Joanna A. Christodoulou,
Harvard University and MGH
Institute of Health Professions
jac765@mail.harvard.edu

Joanna Anderson,
University of New England
jander62@une.edu.au

Joanne Marieke Buil,
UNESCO MGIEP
j.m.buil@hr.nl

John Ehrenfeld,
MIT (Retired)
john.ehrenfeld@alum.mit.edu

John Lupinacci,
Washington State University
john.lupinacci@wsu.edu

John Sabatini,
University of Memphis
jpsbtini@memphis.edu

Jonathan Kay, Education
Endowment Foundation
jonathan.kay@eefoundation.org.uk

Julia Jansen-van Vuuren,
Queen's University
17jmjv@queensu.ca

Julian Sefton-Green,
Deakin University
julian.seftongreen@deakin.edu.au

Julie Sarama,
University of Denver,
julie.sarama@du.edu

Julien Mercier,
Université du Québec à Montréal
mercier.julien@uqam.ca

Jun Xie,
Kyushu University
xie.jun.679@m.kyushu-u.ac.jp

Kaja Jasińska,
University of Toronto
kaja.jasinska@utoronto.ca

Kalervo Gulson,
University of Sydney
kalervo.gulson@sydney.edu.au

Kaori H. Okano,
La Trobe University
k.okano@latrobe.edu.au

AUTHORS

Karina V. Korostelina,
George Mason University
ckoroste@gmu.edu

Karl Friston,
University College London
k.friston@ucl.ac.uk

Kate O'Brien,
Manchester Metropolitan
University
katherine.obrien@mmu.ac.uk

Kathryn Paige Harden,
University of Texas at Austin
harden@utexas.edu

Kenneth Pugh,
Yale University and Haskins
Laboratories
kenneth.pugh@yale.edu

Kim Allen,
University of Leeds
k.allen1@leeds.ac.uk

Kimberly A. Schonert-Reichl,
University of British Colombia
kimberly.schonert-reichl@ubc.ca

Kimberley G. Noble,
Teachers College, Columbia
University in the city of New York
noble2@tc.columbia.edu

Kriti Singh,
UNESCO MGIEP
k.singh@unesco.org

Latika Gupta,
University of Delhi
latikasgupta@gmail.com

Layne Kalbfleisch,
George Washington University;
Northern New Mexico College
layne.2e@gmail.com

Lelya Troncoso Pérez,
Universidad de Chile
lelyatroncoso@uchile.cl

Lesley LL Le Grange,
Stellenbosch University
llg@sun.ac.za

Lien Peters,
University of Western Ontario
lien.peters4@gmail.com

Lily Steyer,
Stanford University
lsteyer@stanford.edu

Linda S. Siegel,
The University of British
Columbia
linda.siegel@ubc.ca

Lindsay G. Oades,
University of Melbourne
lindsay.oades@unimelb.edu.au

Lisa Flook,
University of Wisconsin-Madison
Davis
flook@wisc.edu

Louise Gascoine,
Durham University
louise.gascoine@durham.ac.uk

Ludo Verhoeven,
Radboud University Nijmegen
ludo.verhoeven@ru.nl

Manoj Chandrasekharan,
University of Memphis
manoj.c@memphis.edu

Marc F. Buck,
Helmut Schmidt University
buckm@hsu-hh.de

Marc Joanisse,
The University of Western Ontario
marcj@uwo.ca

Marcia A. Barnes,
Vanderbilt University
marcia.barnes@vanderbilt.edu

Marcia McKenzie,
University of Melbourne
marcia.mckenzie@unimelb.edu.au

Margaret Somerville,
Western Sydney University
margaret.somerville@
westernsydney.edu.au

Maria Katsipataki,
Durham University
maria.katsipataki@durham.ac.uk

Maria Poulou,
University of Patras
mpoulou@upatras.gr

Marissa Willcox,
RMIT University
marissa.willcox@rmit.edu.au

Mark Bray,
East China Normal University and
University of Hong Kong
mbray@hku.hk

Martha Chaves,
Fundación Mentes en Transición
marthacecilia.chaves@gmail.com

Michael Tusiime,
University of Rwanda
krwibasira@yahoo.com

Milene Bonte,
Maastricht University
m.bonte@maastrichtuniversity.nl

Mindy Blaise,
Edith Cowan University
m.blaise@ecu.edu.au

Mohammad Zaman, Education
Endowment Foundation
Mohammad.Zaman@
eefoundation.org.uk

Mohsen Saadatmand,
UNESCO MGIEP
saadatmand.m@gmail.com

Moses Oladele Ogunniran,
UNESCO MGIEP
ogunniranmoses1985@yahoo.com

Nadine Gaab,
Harvard University,
nadine_gaab@gse.harvard.edu

Nandini Chatterjee Singh,
UNESCO MGIEP
n.chatterjee@unesco.org

Natalia Rojas,
NYU Langone Health,
natalia.rojas@nyulangone.org

Navjit Gaurav,
Queen's University
19ng7@queensu.ca

Neil Humphrey,
University of Manchester
neil.humphrey@manchester.ac.uk

Neil Selwyn,
University of Manchester
neil.selwyn@monash.edu

Nicole Patton Terry,
Florida Center for Reading Research
& Florida State University
npattonterry@fsu.edu

Nienke M. van Atteveldt,
Vrije Universiteit Amsterdam
n.m.van.atteveldt@vu.nl

Ola Uduku,
University of Liverpool
o.uduku@liverpool.ac.uk

Oren Ergas,
Beit Berl College
oren.ergas@beitberl.ac.il

Padma M. Sarangapani,
Tata Institute of Social Sciences
psarangapani@tiss.edu

Pallawi Sinha,
University of Suffolk
p.sinha@uos.ac.uk

Pamela Burnard,
The University of Cambridge
pab61@cam.ac.uk

AUTHORS

Patrice M. Bain
patrice@patricebain.com

Patricia-Anne Blanchet,
Université de Sherbrooke
patricia-anne.blanchet@
USherbrooke.ca

Paul Howard-Jones,
University of Bristol
paul.howard-jones@bristol.ac.uk

Paul K. Steinle,
American Institutes for Research
psteinle@air.org

Pauliina Rautio,
University of Oulu
pauliina.rautio@oulu.fi

Peter Goodyear,
University of Sydney
peter.goodyear@sydney.edu.au

Peter Kraftl,
University of Birmingham
p.kraftl@bham.ac.uk

Pierina Cheung,
National Institute of Education
pierina.cheung@nie.edu.sg

R. Malatesha Joshi,
Texas A&M University,
mjoshi@tamu.edu

Radhika Iyengar,
Columbia University
iyengar@ei.columbia.edu

Raihani Raihani,
Sultan Syarif Kasim State Islamic
University of Riau
raihani@uin-suska.ac.id

Raul Olmo Fregoso Bailón,
The University of Texas Rio
Grande Valley
raul.fregosobailon@utrgv.edu

Rebecca J. Collie,
University of New South Wales
rebecca.collie@unsw.edu.au

Rebecca J.M. Gotlieb,
University of California,
rgotlieb@ucla.edu

Rebecca Merkley,
UNESCO MGIEP
rebeccamerkley@cunet.carleton.ca

Reinhard Pekrun,
University of Essex
pekrun@lmu.de

Richard Tucker,
Deakin University
richard.tucker@deakin.edu.au

Robert Talbert,
Grand Valley State University
talbertr@gvsu.edu

Robert William Roeser,
Pennsylvania State University
rwr15@psu.edu

Rogier A. Kievit,
Radboud University Medical
Center
rogier.kievit@radboudumc.nl

Rongxiu Wu,
UNESCO MGIEP
rwu227@g.uky.edu

Ros McLellan,
University of Cambridge
rwm11@cam.ac.uk

Rosiana Lagi,
The University of the South Pacific
lagi_r@usp.ac.fj

Rosie Flewitt,
Manchester Metropolitan
University
r.flewitt@mmu.ac.uk

Rosie Parnell,
University of Newcastle
rosie.parnell@ncl.ac.uk

Ruben Dario Pardo Santamaria,
Universidad del Quindío
rdpardo@uniquindio.edu.co

Rupal Patel,
Education Endowment
Foundation
rupal.patel@eefoundation.org.uk;

Sara Hart,
Florida State University
shart@fcrr.org

Sarah Fishstrom,
The University of Texas at Austin
sarah.fishstrom@utexas.edu

Sarah Howard,
University of Wollongong
sahoward@uow.edu.au

Sarah Strader,
Two Rabbits
sarah.strader2@gmail.com

Sebastian J. Lipina,
Unidad de Neurobiología Aplicada
(UNA, CEMIC-CONICET)
lipina@gmail.com

Sharon Vaughn,
University of Texas at Austin
srvaughn@austin.utexas.edu

Sharon Wolf,
University of Pennsylvania
wolfs@upenn.edu

Shunsuke Managi,
Kyushu University
managi.s@gmail.com

Sidarta Ribeiro,
Brain Institute,
Federal University of Rio Grande
do Norte (UFRN)
sidartaribeiro@neuro.ufrn.br

Sigrid Hartong,
Helmut Schmidt University
hartongs@hsu-hh.de

Smita Kumar,
Independent Educator &
Researcher
smitkp108@gmail.com

Sonia Guerriero,
UNESCO
s.guerriero@unesco.org

Soren Christensen,
Aarhus University
socr@edu.au.dk

Stephanie Bugden,
University of Winnipeg
bugden@uwinnipeg.ca

Steve E. Higgins,
Durham University
s.e.higgins@durham.ac.uk

Steve Sider,
Wilfrid Laurier University
ssider@wlu.ca

Susan Germein,
Western Sydney University
s.germein@westernsydney.edu.au

Sybille Lammes,
Leiden University
s.lammes@hum.leidenuniv.nl

Taha Rajab,
Durham University
taha.rajab@durham.ac.uk

Tal Gilead,
The Hebrew University of
Jerusalem
tal.gilead@mail.huji.ac.il

Tejendra Pherali,
University College London
t.pherali@ucl.ac.uk

AUTHORS

Tenelle Porter,
University of California
tjporter@ucdavis.edu

Teresa Iuculano,
Université de Paris,
La Sorbonne and Centre National
de la Recherche Scientifique
teresa.iuculano@u-paris.fr

Thérèse Jay,
Paris Descartes University
therese.jay@inserm.fr

Thomas Macintyre,
UNESCO MGIEP
thomas.macintyre@gmail.com

Tina Grotzer,
Harvard University
tina_grotzer@harvard.edu

Valeria Cavioni,
University of Milano-Bicocca
valeria.cavioni@unimib.it

Vanessa Rodriguez,
NYU Grossman School of
Medicine vanessa.rodriguez@
nyulangone.org

Victoria Goodyear,
University of Birmingham
v.a.goodyear@bham.ac.uk

Vinay Kathotia,
The Open University
vinay.kathotia@open.ac.uk

Vlad Glaveanu,
Webster University Geneva
glaveanu@webster.ch

W. Huw Williams,
University of Exeter
w.h.williams@exeter.ac.uk

William Pinar,
The University of British
Columbia
william.pinar@ubc.ca

Xiaomin Li,
Beijing Normal University
lixiaomin@bnu.edu.cn

Yaacov Petscher,
Florida State University
ypetscher@fsu.edu

Yusef Waghid,
Stellenbosch University
yw@sun.ac.za

Yuto Kitamura,
University of Tokyo
yuto@p.u-Tokyo.ac.jp

Yuzhen Xu,
Capital Normal University
yuzhen6@hotmail.com

Zhang Wei,
East China Normal University
wzhang@ed.ecnu.edu.cn

REVIEWERS

REVIEW EDITORS

WG1:

Elaine Unterhalter,
University of London
e.unterhalter@ucl.ac.uk

Iroise Dumontheil,
Birkbeck, University of London
i.dumontheil@bbk.ac.uk

WG2:

Anne-Laura van Harmelen,
Leiden University
a.van.harmelen@fsw.leidenuniv.nl

Greg William Misiaszek,
Beijing Normal University
gmisiaszek@bnu.edu.cn

WG3:

Elaine Unterhalter,
University of London
e.unterhalter@ucl.ac.uk

Roland Grabner,
University of Graz
roland.grabner@uni-graz.at

WG4:

Drew H. Bailey,
University of California, Irvine
dhbailey@uci.edu

PEER REVIEWERS

Aaron Redman,
Arizona State University
aaron.redman@gmail.com

A. Brooks Bowden,
University of Pennsylvania
bbowden@upenn.edu

Adele Diamond,
University of British Columbia,
Vancouver
adele.diamond@ubc.ca

Adar Ben-Eliyahu,
University of Haifa
adarbe@edu.haifa.ac.il

Anton Kühberger,
University of Salzburg
anton.kuehberger@sbg.ac.at

Bob Adamson,
University of Nottingham Ningbo
China
bob.adamson@nottingham.edu.cn

Brian Haig,
University of Canterbury
brian.haig@canterbury.ac.nz

Brian L. Butterworth,
University College London
b.butterworth@ucl.ac.uk

Cathy Jane Rogers,
NPC, UK
cathyjanerogers01@gmail.com

Chris Dede,
Harvard University
USA chris_dede@gse.harvard.edu

Christopher Martin,
University of British Columbia
(Okanagan)
christopher.martin@ubc.ca

Constadina Charalambous,
European University Cyprus,
Nicosia, Cyprus
co.charalambous@euc.ac.cy

Cyril Owen Brandt,
University of Antwerp
cy.brandt@yahoo.com

Darcia Narvaez,
University of Notre Dame
dnarvaez@nd.edu

Denis Cousineau,
Université d'Ottawa, Ottawa
denis.cousineau@uottawa.ca

Gert Biesta,
University of Edinburgh
gert.biesta@ed.ac.uk

Hongyun Liu,
Beijing Normal University
hyliu@bnu.edu.cn

Hyemin Han,
The University of Alabama,
hhan19@ua.edu

Jandhyala B.G. Tilak,
Council for Social Development
jtilak2017@gmail.com

Jason Beech,
Monash University
jason.beech@monash.edu

Jeremy Rappleye,
Kyoto University
jrappleye108@gmail.com

Jing Lin,
University of Maryland
jinglin@umd.edu

REVIEWERS

John-Tyler Binfet,
University of British Columbia,
Vancouver, Canada johntyler.
binfet@ubc.ca

Jonathan Cohen,
Columbia University in the city of New York
jc273@tc.columbia.edu

Karen L. Bierman,
The Pennsylvania State University
kb2@psu.edu

Katarzyna Jednoróg,
Polish Academy of Science
k.jednorog@nencki.edu.pl

Leon Benade,
Auckland University of Technology
leon.w.benade@aut.ac.nz

Ly Thi Tran,
Deakin University
ly.tran@deakin.edu.au

Marinus H. van IJzendoorn,
Erasmus University Rotterdam
marinusvanijzendoorn@gmail.com

Michael J. Reiss,
University College London,
m.reiss@ucl.ac.uk

Peter Barrett,
University of Salford
peter.x.barrett@gmail.com

Randall Curren,
University of Rochester
randall.curren@rochester.edu

Robert VanWynsberghe,
University of British Columbia
robert.vanwynsberghe@ubc.ca

Rwitajit Majumdar,
Kyoto University
majumdar.rwitajit.4a@kyoto-u.ac.jp

Sabre Cherkowski,
University of British Columbia
sabre.cherkowski@ubc.ca

Sami Adwan,
PRIME
adwan.sami@gmail.com

Steven J. Klees,
University of Maryland
sklees@umd.edu

Tonya Huber,
Texas A&M International University
tonya.huber@tamiu.edu

Victoria Knowland,
Newcastle University
vic.knowland@newcastle.ac.uk

Thomas DeVere Wolsey,
The American University in Cairo
thomas.wolsey@aucegypt.edu

ACKNOWLEDGEMENTS:

The ISEE Assessment would like to acknowledge contributions from the following: Shelby Roberts, Sue Stirling, Thomas Abrams, Bing Han, Emily Binks-Cantrell, Karol Ann Moore, Brahm Norwich, Célia Demarchi, Elizabeth R. Losin, Jack Van Honk, Katherine D. Blizinsky, Manulani Aluli Meyer, Philipp Koellinger, Pooja Reddy Nakamura, Selena Nemorin, Tetsuya Iidaka, Zhang Li, Julia Oas, Anika Alam, Rajiv Ramaswamy, Bethany Davies, Eti Chordia, Matthew John DeLong, Kavya Bhardwaj, Juliette Diebold De Mellis, Tsering Yangdol, Clara Paulina Cardona-Vasco, Richa Shrivastava, and the ISEE Assessment Secretariat Yoko Mochizuki, Shailly Gupta and Kriti Singh.

WORKING GROUP

04

EDUCATION – *Data & Evidence*

●

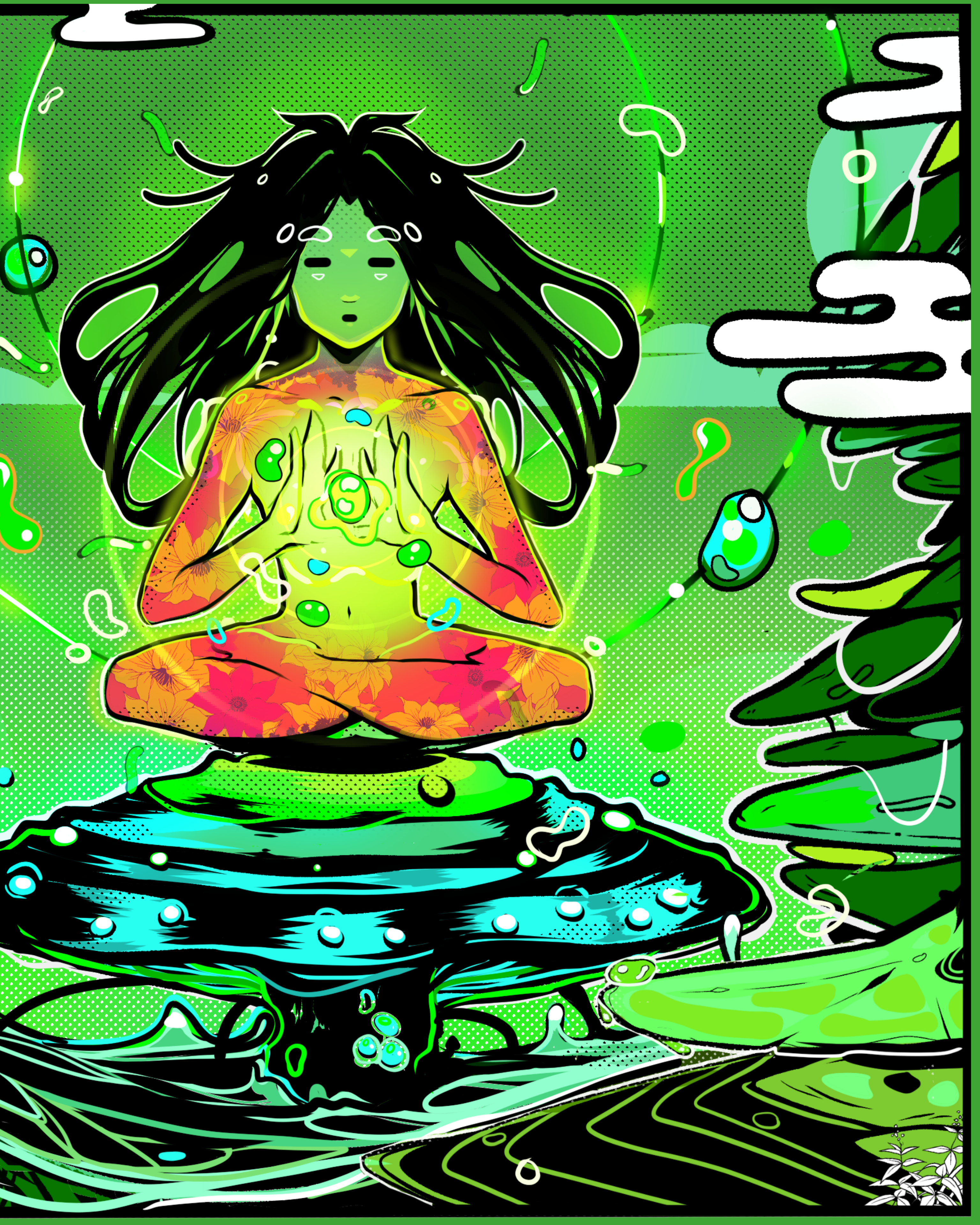
WORKING GROUP 4 CO-CHAIR:

JULIEN MERCIER

04

EDUCATION – *Data & Evidence*

Working Group 4 assesses traditional levels of evidence in evidence based education (EBE), proposes for levels of contextual fitting, providing implications for conducting future applied research for policy-making given levels of certainty in how well educational interventions work and the extent to which such interventions have been studied.



WORKING GROUP 04

CHAPTER

1

The EBE3 framework: extending evidence based education from causal ascriptions and effectiveness generalizations to relative effectiveness generalizations and local effectiveness predictions

DOI: https://doi.org/10.56383/QHEN7211

This chapter should be cited as:

Mercier, J. and Bourgault Bouthillier, I. (2022). ‘The EBE3 framework: E] extending evidence based education from causal ascriptions and effectiveness generalizations to relative effectiveness generalizations and local effectiveness predictions’ in Duraiappah, A.K., van Atteveldt, N.M., Borst, G., Bugden, S., Ergas, O., Gilead, T., Gupta, L., Mercier J., Pugh, K., Singh, N.C. and Vickers, E.A. (eds.) Reimagining Education: The International Science and Evidence Based Assessment. New Delhi: UNESCO MGIEP.

To more reliably achieve educational goals based on values and policies, quantitative and qualitative traditions should complement each other to strengthen the quality and impact of empirical research, under a broad banner of evidence based education (EBE). A different approach to EBE can help to solve questions relating to ‘what works’, by extending this question to ‘what is working best generally’ and ‘will a given intervention work here and now?’. This chapter proposes a more complete framework for EBE by delineating the information and reasoning needed to address a cascade of questions that jointly determine the best course of action for obtaining the best educational outcomes. The traditional levels of evidence are revised and complemented by a proposal for levels of contextual fitting, grounded in both theory building and theory testing. Implications for conducting future applied research, for policy-making and for improving educational practice are discussed.

On average, the temperature in Alaska is above freezing, but if I am planning a trip and hope to avoid the snow I must figure out when it is above freezing and when it is below. Similarly, the greater the impact varies among sites and students, the less we learn from an average treatment effect, even if it is accurate for the broad population (Joyce, 2019).

Coordinating Lead Authors

Julien Mercier

Lead Author

Iris Bourgault Bouthillier

WORKING GROUP 04

1

CHAPTER

1.1 Introduction: ‘what works’ is not enough

The goals of education are based on values and policies **(Brighouse et al., 2018)**. This public policy-making is a political process that requires conflict, negotiation, the use of power, bargaining and compromise **(Anderson, 2011)**. When it comes to the means to achieve those goals, the relative benefits of some approaches over others are assessed through empirical research, where quantitative and qualitative traditions have complementary roles **(National Research Council, 2002; Karrigan and Turner-Johnson, 2019)**. This research, which is usually conducted with samples of learners, involves, among other things, the observation of gains on target outcomes and processes. Since human learning and

From the perspective of evidence based education, decisions about which practices to use in a given learning context should ideally be based on evidence.

development are the cornerstones of educational goals (albeit reformulated through reforms), the domains contributing to educational research rest essentially on a vast number of fields in the learning sciences and cognitive science (psychology and neuroscience (behaviour and brain processes), computer science (computer based learning systems, learning analytics), and economics and social sciences (the learner and their broader context).

1.1.1

PROBLEM: THERE IS A NEED TO APPLY A HIGH MINIMUM STANDARD FOR WHAT COUNTS AS EVIDENCE OF IMPROVED LEARNING

Pertinent research relies on a variety of methods, which, in essence, focus on different aspects of theory building and validation. From the perspective of evidence-based education (EBE), decisions about which practices to use in a given learning context should ideally be based on evidence **(Slavin, 2020)**. Evidence starts with a demonstration of the effect of some treatment on a defined outcome **(Connolly, Keenan and Urbanska, 2018)** and, more broadly, of empirical support that a policy works generally or in a specific context **(Joyce and Cartwright, 2020)**. This essential foundation means that we should expect a higher standard: that is, to know whether an intervention works better than what we were already doing, compared to a control group and after eliminating as many possible sources of bias. To know this, a level of confidence in the inferences made from the empirical investigations need to be considered. The study of 'what works' is limited to causal ascriptions, that is, the estimated causal effect of an intervention on the targeted outcomes. These arise from a comparison of an experimental group with a control group. Causal ascriptions, combined with assumptions about generalizability and replicated across a few studies, lead to general effectiveness claims. These are the inferences that results obtained with samples will apply to the corresponding population(s) and context(s). Thus, the demonstration simply indicates that a given intervention is better than the normal practice (which has proven difficult to define) **(see Kornell, Rabelo and Klein (2012) for an example)**. Even state-of-the-art experiments carried out at the cluster level (e.g. forty to fifty schools or classrooms) as advocated by Slavin **(2020)**, use designs limited to testing causal ascriptions, just like traditional comparisons between experimental and control groups. These complex experiments often use hierarchical linear modelling to take into account the similarity of the participants within a school or classroom. This only improves what Shadish, Cook and Campbell **(2002)** call statistical conclusion validity (by getting the standard errors right) but not internal validity (the potential to establish the unbiased effect of an intervention) or external validity (notably the potential for generalization). The results of a collection of high-quality studies (unbiased sampling, randomized treatment/group assignment, well-defined intervention, valid and reliable measures, statistical analyses with power, effect size and significance tests) comparing an experimental group given a target intervention with a control group has been the cornerstone of EBE for decades under the label 'what works'. It is the main, but not sufficient, building block of EBE, because such studies provide relatively isolated indications of the effectiveness of interventions, which remain to be further compared and rank-ordered empirically. Thus, there is a need for a higher minimum standard for what counts as evidence of improved learning. This chapter proposes an evolution of previous efforts and capitalizes on the EBE building block 'what works' to develop further rationales for establishing the efficacy of interventions.

In a logic of cumulative generalization and abstraction of claims of effectiveness, the best evidence is available when every possible intervention for a specific goal and target population - including the context of that population - has been tested with equally valid studies (ideally replicated) and then rank-ordered with respect to its established effect.

In a majority of cases, the evidence is scattered, emerging and incomplete, or based on a multiplicity of research designs, methods and conceptual frameworks.

In a logic of cumulative generalization and abstraction of claims of effectiveness, the best evidence is available when every possible intervention for a specific goal and target population – including the context of that population – has been tested with equally valid studies (ideally replicated) and then rank-ordered with respect to its established effect. In such cases, choosing the best intervention and which to try first, second or third in terms of specific outcomes is straightforward, at least in terms of efficacy **(Goldacre, 2013)**. Unfortunately, educational issues tested this way are scarce but have been increasing during the last decade **(Connolly, Keenan and Urbanska, 2018)**. In a majority of cases, the evidence is scattered, emerging and incomplete, or based on a multiplicity of research designs, methods and conceptual frameworks. The common denominator is the level of trust in the inferences made from empirical investigations. It is important to consider that when alternatives exist, effectiveness of available interventions is always relative to the effectiveness of some other intervention(s). We call these inferences 'general relative effectiveness claims', because they stem directly from the comparison of effectiveness generalizations.

1.1 .2

PROBLEM: BEFORE THE NEED FOR ADDITIONAL EVIDENCE IN THE FORM OF NEW TESTS OF INTERVENTIONS, THERE IS A NEED FOR 'RELATIVE EVIDENCE'

We define relative evidence as the result of thorough comparisons of extant interventions, under the assumption **(see the Australian Society for Evidence Based Teaching)** that combined results coming from meta-analyses or systematic reviews are much more

Relative evidence arises from the combined results of multiple studies, using meta-analysis and made possible by thorough comparisons of effect sizes of multiple extant interventions.

informative than single – albeit excellent – studies when necessary precautions are taken **(Simpson, 2018)**. These necessary precautions consider that effect size (the indication of the impact of a given intervention) is due not only to the intervention, but also may be part of the whole study (e.g. sample size, test characteristics and comparison treatment). Relative evidence arises from the combined results of multiple studies, using meta-analysis and made possible by thorough comparisons of effect sizes of multiple extant interventions. The consistency or variability of effect sizes across studies of similar interventions is critical to support assertions regarding their general effectiveness. In addition, the consistency of effect sizes across studies is critical to empirically support assertions about what we have termed relative effectiveness generalizations, that is, claims that the relative effectiveness of interventions, tested with samples, will apply to the corresponding populations and contexts. However, there is a lack of relative evidence in extant literature regarding most educational issues: new interventions are tested against a control group (business as usual) and well-documented interventions rarely get rank-ordered through a proper meta-analytic approach.

Aside from scientific challenges, the lack of relative evidence may unfortunately be explained, at least in part, by policies governing research. Indeed, the neoliberal model underlying the funding of research and educational institutions 'has forced academic researchers to dismiss methodological limitations of social science research … and overestimate the impact of their research in order to obtain highly competitive, and scarce, research money … fueling a replication controversy in published research' **(Karrigan and Turner-Johnson, 2019, p. 290)**. Moreover, Chubb and Watermeyer **(2017)** synthesize a drift from traditional and still desirable norms in academia including communism, universalism, disinterestedness and organized scepticism; and the defence of critical, objective truth. This drift pulls academics toward professional pragmatism and sponsorism as a survival response in the face of demands and directives of academic capitalism and 'managerial' governmentality, seen as hegemonic and inescapable. In the end, these forces rewarding short-term and shallow productivity do not encourage the undertaking of thorough synthesis work.

An effectiveness prediction is the prediction that a given intervention, abstracted through causal ascriptions, effectiveness claims and relative effectiveness generalizations will work concretely within the specific constellation of variables of a given application context.

1.1 .3

PROBLEM: IF RELATIVE EVIDENCE IS AVAILABLE AND GENERAL RELATIVE EFFECTIVENESS CLAIMS ARE SUPPORTED BY APPROPRIATE EVIDENCE, THERE IS A NEED FOR STRONG ASSERTIONS ABOUT HOW THE LOCAL CONTEXT IN WHICH THE EVIDENCE IS TO BE APPLIED OUGHT TO AFFECT OUR EXPECTATIONS OF IMPACT

An effectiveness prediction **(Joyce and Cartwright, 2020)** is the prediction that a given intervention, abstracted through causal ascriptions, effectiveness claims and relative effectiveness generalizations will work concretely within the specific constellation of variables of a given application context. Such comparisons can enable assessment of the effectiveness against specific outcomes of all pertinent interventions, allowing practitioners to answer the question: given all the possible interventions available to me, which is most likely to succeed in my specific context? Ultimately, this context concerns a specific

Local effectiveness predictions are generally either absent from implementation efforts, or tackled through biased, non-scientific reasoning, such as beliefs, peer pressure, marketing, and so on.

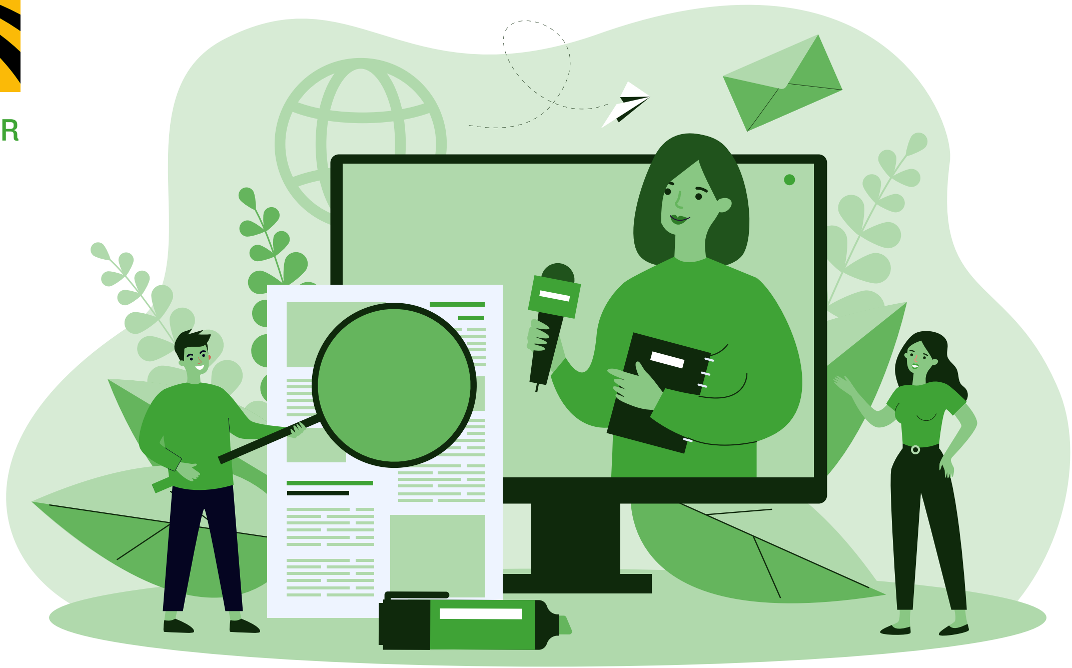
teacher and a specific classroom at a specific moment in time, this specificity being the opposite of the potential for generalization sought by quantitative research. In other words, evidence is needed to support the prediction that a given intervention abstracted across causal ascriptions and general effectiveness claims will work concretely within the specific constellation of variables of a given context of application. These assertions are what Joyce and Cartwright **(2020)** have termed 'local effectiveness predictions'.

These local effectiveness predictions have proven elusive in the traditional view of EBE. Reasoning about how causal claims related to a given intervention will yield documented outcomes in a target concrete and specific context (a given school for example), evidenced by the right information, has not been clear or available. Consequently, EBE at this step has consisted of merely applying research-based practices, that is, causal ascriptions and general effectiveness claims. This applicationist stance is accompanied by concerns about teacher training, teachers as technicians rather than professionals, educational leadership, accountability and scaling up of interventions. Local effectiveness predictions are generally either absent from implementation efforts, or tackled through biased, non-scientific reasoning, such as beliefs, peer pressure, marketing, and so on. It would be possible, in education, to be a lot more efficient in implementing best practices by applying a rationale increasingly used in other fields **(Pawson et al., 2005; Pawson, 2006)** that explicitly concerns how contextual elements facilitate the release of active ingredients in interventions documented as the most effective.

What constitutes a fully operational EBE has not yet been framed as a coherent cascade of questions related to the decision-making involved in implementing the best interventions and driving the production/consideration of the necessary information. Nor have these questions been

This chapter aims to provide an overview of the nature of scientific evidence in education and to suggest a framework that encompasses all current types of efforts related to the development of educational knowledge, and posits the overall progress of educational research as a compromise between theory building and validation.

operationalized in terms of required evidence paired with the necessary empirical work. This leaves the vast majority of educational research, synthesis work and application endeavours subject to gaps that need to be satisfactorily resolved in a specific sequence. Globally, in line with Joyce and Cartwright, **(2020)**, we are concerned with the information and reasoning needed to address a cascade of questions that jointly determine the best course of action for obtaining the best educational outcomes: what works? What is working best generally? Will it work here (tomorrow, in my classroom)? This chapter aims to provide an overview of the nature of scientific evidence in education and to suggest a framework that, firstly, encompasses all current types of efforts related to the development of educational knowledge, and, secondly, posits the overall progress of educational research as a compromise between theory building and validation. It is expected that this integrated framework is both practical and useful for stakeholders (researchers, policy-makers and practitioners) in educational systems. Hence, the first section of this chapter discusses the importance of theory building and theory testing in educational research. The second section discusses the levels of evidence, their usefulness and their limits. The third section presents an original framework aiming at the application of evidence in specific contexts, which to date has been underspecified. Finally, the usefulness of this new framework for stakeholders is discussed. An appendix outlines a procedure for obtaining the necessary information and making the necessary inferences from it to answer key questions in a process of EBE: after determining the most important educational goals, identifying the means to attain these goals by using or fostering necessary results from pertinent empirical work. The application of this procedure can ultimately be used as a practical tool for conducting literature reviews and implementation work as well as policy-making.



1.2 Theory building and theory testing in educational research: divide, compromise or synergy?

Besides the emphasis on empirical developments in EBE, another essential aspect of educational research is the development of theory. The National Research Council (NRC) **(2002)** briefly defines theory as follows: scientific theories are conceptual models used to explain phenomena. In the social sciences and humanities (including education) the nature of theories has been largely discussed. The NRC recognizes a continuum between 'grand' theories, that aim at generalizing theoretical understanding, and research that seeks to achieve deep understanding of particular events or circumstances. In between these two extremes are mid-range theories attempting to account for social aspects and particular elements of situations. All theories, wherever they are located on this continuum, consist of representations or abstractions of some aspect of reality that can only be approximated by such models. We place limited emphasis on the 'grand' theories that aim at generalizing theoretical understanding and focus on mid-range theories attempting to account for social aspects and particularities of situations. Mid-range theories consist of representations or abstractions of aspects of reality that can be approximated by conceptual models, which can be subjected to empirical tests. According to Maciver et al. **(2019 pp. 13–14)**:

> ***The term "middle range" theory refers to the level of abstraction at which useful theory for realist work is written: detailed enough and "close enough to the data" that testable hypotheses can be derived from it, but abstracted enough to apply to other situations as well ... Middle range theorization is useful because it offers an analytical approach to linking findings from different situations.***

According to the NRC, one of the main principles of scientific inquiry is to link empirical research to relevant theory. Empirical research can be linked in many ways to theory. Depending on the underlying epistemology and the advancement of knowledge in the field, theory can either be what guides a study or what emerges from it. In many cases, theory can be linked to research in both ways when a study is based on theory and at the same time enriches it. In short, theory is what 'drives the research question, the use of methods, and the interpretation of results' **(National Research**

Besides the emphasis on empirical developments in EBE, another essential aspect of educational research is the development of theory.

Council, 2002). Thus, theory has an undeniable importance in applied science.

In the learning sciences, theory is notably what allows researchers, decision-makers and practitioners to support the use of interventions in specific contexts and understand the underlying mechanisms **(Joyce, 2019)**. When reviewing the scientific literature about a topic, stakeholders in education should therefore be able to determine the contribution of a study or group of studies to the advancement of theory. Research can contribute to theory in two main ways: theory building and theory testing (validation). These two types of contribution are not mutually exclusive. Research has shown in some fields that the more an article contributes to theory in one or both ways, the more it will be cited **(Colquitt and Zapata-Phelan, 2007)**. While the citation rate is not the only way to measure the importance of a scientific publication **(Sugimoto and Larivière, 2018)**, it can be considered a general indicator of the impact of research. Hence, in some fields, the more an article is contributing to theory, whether by building it, testing it, or both, the more impactful this piece of research tends to be for the scientific community, as reflected by its citation rate.

The next paragraphs describe a taxonomy created by Colquitt and Zapata-Phelan **(2007)** that can be used to capture many facets of the theoretical contributions of an empirical study. Although their article is focused on the field of management, it can easily be transferred to the field of education, given these fields share many similarities. For example, they are both social sciences based on values and policies and the research methods and nature of theories used in both fields are mostly the same. The taxonomy is built on two orthogonal axes, theory building and theory testing, which are both divided into five ordinal levels. A given empirical study is situated on both axes. Qualifying a corpus of studies in a given field this way may help assess the maturity of the research on a given educational issue

In the learning sciences, theory is notably what allows researchers, decision-makers and practitioners to support the use of interventions in specific contexts and understand the underlying mechanisms

and may help in extracting the information needed to address the main questions of the framework proposed.

One axis presents five levels of theory building. The first two levels of theory building on the axis are considered low-level contributions. The first level represents attempts to replicate results that already support existing theories. Replication studies are very important to science because they offer substantial protection to the quality and credibility of empirical scientific work; specifically, issues linked to false positives results, null results and questionable research practices **(Frias-Navarro et al., 2020)**. Despite their importance, they are considered the lowest level in terms of contributing to building new theories. Level 2 attempts to examine effects that have already been the subject of prior theorization. Level 3 includes studies that introduce new variables (e.g. mediators or moderators) to existing theories on relationships or processes. Level 4 studies explore new relationships or processes. Finally, level 5 includes studies that propose entirely new theories, models or concepts, or that significantly reconceptualize existing ones.

The other axis illustrates five levels of theory testing. Studies from the first level are either inductive or ground their predictions within logical speculation. In this level, one may find exploratory studies that are not necessarily based on prior theory or concepts. Level two studies ground their predictions with references to past findings. This means that the results are put in relation to other findings but are not explicitly based on prior theory or concepts. Level three includes studies that ground their predictions with existing conceptual arguments, while level four studies' predictions are grounded within existing models, diagrams or figures. Finally, level five studies explicitly ground their predictions on existing theory.

The interaction between the two axes enables us to distinguish five discrete article types in terms of

their theoretical contribution: the reporters, the testers, the qualifiers, the builders, and the expanders. For specific examples of articles that fit into each of these categories, see the article by Colquitt and Zapata-Phelan **(2007)**.

The reporters category includes empirical articles that score low on both axes. For example, an article that aims at replicating a previous study **(level 1 of theory building)** with hypotheses based on findings of several prior other studies on the topic **(level 2 of theory testing)** would be classified in this category. Even when studies are considered to be low on both axes, it is important to stress that they can still be constructive and useful for science. Testers includes articles that show high levels of theory testing and low levels of theory building. This category includes articles that aim primarily at testing existing theories empirically without incorporating new constructs or variables. The qualifiers category is composed of articles that contain moderate levels on both axes. They can be articles that push previously demonstrated relationships a little further. For example, articles in this category can be based on previously demonstrated relationships between concepts and try to add a new mediator to qualify this relationship. Builders are articles that score high on the theory building axis and low on the theory testing axis. This category includes, amongst others, inductive studies that elaborate new constructs, relationships or processes. Finally, the expanders are articles that are high on both theory testing and theory building axes. Like builders, they focus on new constructs, relationships and processes that have not already been theorized, but they do it while also testing existing theory.

While the taxonomy of theoretical contributions for empirical articles that allows classification of articles according to their level of theory building and theory testing contribution can be very informative, it only depicts empirical studies intended theoretical development, not how well it is done **(Colquitt and Zapata-Phelan, 2007)**. As the authors themselves argue, many other important underlying factors could be added to their taxonomy: how interesting is a new construct, how much a new relationship adds to the relevant literature, how rigorously a theory is tested, and so on. This taxonomy conveys a profound message: theory is at the heart of the advancement of science and the value of empirical observations is contingent on their contribution to theory building and theory testing. As will be discussed in the next sections, theory is central to progress in the hierarchies of the EBE3 framework. To answer the question of what is working best generally, theory defines and isolates the active ingredients in interventions. This is critical for classification of interventions in meta-analytic work so that the comparisons are warranted and interpretable. To answer the question about replicating the efficacy of a given intervention in a specific context, pertinent theory defines experimental and observational elements to take into account and mechanisms and processes not to take into account for the purposes of predicting efficacy **(Pearl and Bareinboim, 2014)**.

While the taxonomy of theoretical contributions for empirical articles that allows classification of articles according to their level of theory building and theory testing contribution can be very informative, it only depicts empirical studies intended theoretical development, not how well it is done

1.3 What is working best generally: levels of evidence

Along with Joyce (**2019**), we consider causal ascriptions, on which the so-called ‘what works’ approach hinges, to be extremely limited in informing the implementation of interventions in EBE. Consequently, we begin our discussion of the necessary ingredients of an empirical demonstration of effectiveness with the notion of general effectiveness claims. General effectiveness claims build upon causal ascriptions and consist of a further empirical demonstration of: (1) the relative effectiveness of available intervention; and (2) the variations in effect across studies, contexts and populations. This empirical demonstration requires a meta-analytic approach, conducted with state-of-the-art procedures to avoid common, published mistakes **(Borenstein, 2019).**

Insofar as applied research improves professional practices in education, and given the impact of these practices on learners, it seems desirable to be able to judge the relative value of available research results relevant to practice, following a set of considerations pioneered by Cochrane **(1972)**. For each aspect of the role of the teacher or professional, it must be possible to determine either an absence of research, the presence of poor-quality research, the presence of quality research and possibly the accumulation of relevant and converging research. From an interventionist perspective that follows a basic premise, namely that the best information for practice is of an applied and causal nature **(Joyce, 2019)**, it is necessary to formulate unambiguous inferences between an intervention and its effect on the learner. In this regard, consensual criteria on which these causal inferences can be established, taken up across a majority of applied fields emanating from the human sciences, are brought together through the notion of levels of evidence.

In light of the cumulative nature of empirical evidence, the levels of evidence are operationalized domain by domain, from a gradation of internal and external validity of the available evidence. Also, considering a standard benchmark of effectiveness, the most common being effect size, is essential in merging evidence about relative effectiveness across increasingly broad educational areas of intervention in order to prioritize intervention in these areas.

In light of the cumulative nature of empirical evidence, the levels of evidence are operationalized domain by domain, from a gradation of internal and external validity of the available evidence.

In areas related to learning, different types of research questions are needed to design and document the effectiveness of practices empirically. These types of questions are accompanied by different methods: manipulation of experimental groups, correlational studies, single-case designs and qualitative methods. Several authors have suggested hierarchies allowing classification of scientific evidence according to the level of confidence that can be attributed to the inferences drawn from them (see the literature review on the subject) **(Nutley, Powell and Davies, 2013)**. Most of these classifications are generally similar to one another in content. In Table 1, we propose such a classification of the pseudoscientific and scientific evidence applied to educational research. This proposal of criteria for the efficacy of intervention seeks to extend prevalent hierarchies of evidence to encompass the various types of evidence created and disseminated, including inadequate, pseudoscientific evidence, **(e.g. Evans, 2003; Burns, Rohrich and Chung, 2011)**. It allows the distinction of: (1) information of pseudoscientific or non-scientific nature; 2) the results emanating from a scientific approach; and 3) probative evidence concerning the relative convergence and divergence of the integrality of available research results. The terms probative, scientific and pseudo-scientific/non-scientific are used for clarity in relationship with the common language of researchers, practitioners and policy-makers in education. They are used to provide clear benchmarks to classify sources of evidence and should not be seen as exclusive or unrelated. Hansson **(2009)** defines a pseudoscientific assertion using three criteria: (1) it pertains to an issue within the domains of science (in the wide sense); (2) it is not epistemically warranted; (3) it is part of a doctrine creating the impression that it is epistemically warranted. Scientific, in the context of applied educational research, is meant to provide limited empirical indications about the efficacy of a given intervention. Probative is understood as the ability of

The terms probative, scientific and pseudo-scientific/non-scientific are used for clarity in relationship with the common language of researchers, practitioners and policy-makers in education evidence to make an assertion true, in this case the assertion pertaining to 'effectiveness'. The pseudo-scientific category comes from belief, biased observation, and so on. The scientific category, on the other hand, comes from rigorous research answering valid research questions. The probative nature of research results refers to the best level of confidence that can be placed in the results of scientific studies aimed at establishing the effectiveness of interventions.

Each level of evidence is described, in descending order of potential to empirically answer the question of what is working best generally. Contrary to Goldacre's **(2013)** claim that students are 'similar enough that research can find out which interventions will work best overall' **(p. 7)**, it is essential to stress the importance of carefully analysing the circumstances of practice that we want to support scientifically **(Joyce, 2019)**. Thus, the learning object, the learner's particularities, as well as the context of intervention are among the elements to be considered to establish the correspondence between the educational act and the available scientific literature. Any discrepancy between the circumstances of 'real' practice and the circumstances of practice as studied in the scientific literature decreases the level of scientific evidence. It can be suggested that the 'real' practice circumstances prevail, and that this will establish the level of scientific evidence that applies, rather than implementing practices supported by the best scientific evidence that would prove unrelated to the current practical needs. Although this is tangential to this chapter, it should be noted that proper training and expertise of the educational professional are necessary for the analysis outlined above.

The only probative sources of evidence are grouped at level 1. Probative qualifies evidence that fully proves a given assertion about the relative effectiveness of interventions. Levels 2, 3, 4 5 and 6 constitute the scientific range because they support causal

TABLE 1 LEVELS OF EVIDENCE APPLIED TO EDUCATION RESEARCH TOWARD EFFECTIVENESS GENERALIZATIONS

RANGE	LEVEL	SOURCES OF EVIDENCE	MAIN LIMITATIONS
Probative: provide effectiveness generalizations	1	Mega-analysis, meta-analysis, narrative literature review, evidence-based review	Abstracted, decontextualized recommendations
Scientific: provide causal ascriptions	2	Experimental studies	Do not provide relative effectiveness generalizations
	3	Quasi-experimental studies	Internal validity
	4	Correlational studies, quantitative case studies	Impossible to verify causality
	5	Experts committees, clinical experience from experts (teamwork reports)	Opinions subject to political or personal influences
	6	Qualitative research, single case protocols	Lack of generalizability
Pseudo-scientific and non-scientific: beliefs not related to solid observation or reasoning	7	Bad quality research (qualitative or quantitative)	Improper methodology
	8	Absence of research, practice reports, trends	Lack of systematic empirical observations

inference, generalizability and replication to varying degrees. The pseudo/non-scientific range is included last, with levels 7 and 8 as red flags, because practitioners in education are frequently exposed to information pertaining to these levels. Levels 1 and 2 are discussed in more detail below, level 1 because although it represents the best sources of general effectiveness claims, it is not exempt from issues in improving educational intervention, and level 2 because it has been seen as the gold standard for EBE for decades despite significant strengths and limitations. Solutions to the limitations of level 1 are suggested later in this chapter.

Level 1 shows the relative effectiveness and variability in outcomes of all the interventions tested experimentally. Mega-

Any discrepancy between the circumstances of 'real' practice and the circumstances of practice as studied in the scientific literature decreases the level of scientific evidence.

analyses (the meta-analysis of meta-analyses, also called meta-meta-analysis) and meta-analyses are preferred because they provide relatively unbiased empirical results. Narrative literature reviews (a discussion of important topics on a theoretical point of view **(Jahan et al., 2016)** and evidence-based reviews (also called systematic reviews) also qualify as probative because they concern available interventions and their relative effectiveness, although it must be noted that they are much weaker than the meta-analytic approach; they are more subjective and may lack the sensitivity to extremes and combination of factors that is characteristic of meta-analyses. The major limitation of this level is that it provides abstracted, decontextualized recommendations. Indeed, the increasing level of aggregation of results needed for probative evidence implies a gradual dissociation with the contexts of the experiments. It is important to point out that evidence at this level is absolutely necessary to qualify research results as probative for any given issue, but the quality of evidence at this level depends on the quality of the primary studies in the scientific range, which get aggregated in the probative range. Also, the demonstration in this chapter that there is no substitute for properly aggregated results at the probative level indicates that interventions implemented should be properly documented at the probative level. If educational goals in policy-making involve means not documented at the probative level, then the implementation of these means in practice should be deferred until the necessary evidence is available. In fact, these goals should drive the production of this evidence.

Level 2 contains the best experimental evidence to support causal ascriptions and effectiveness generalizations. Experimental studies, the gold standard being randomized-controlled trials, allows adaptation of the design to specific target populations and the intervention context. As stated earlier, the more an experimental design is closely related to the

Because of the complexity of educational issues, a conservative position seems warranted and it appears that, all things considered, opinions remain weaker than scientific observations.

real context of practice, the more confidence one can have in the interpretations drawn from the evidence in their own context of practice. The main caveat, as Joyce **(2019)** describes, is the difficulty of determining which characteristics of the populations and the intervention contexts must be considered salient for educational decision-making. Selected characteristics are used as evidence of the representativeness of sampling without supporting their relevance for educational outcomes with evidence. Applying them indiscriminately will not help educators find studies that are appropriately representative and may even lead them astray.

Level 3 shows that quasi-experimental studies have documented the effect of an intervention. However, sampling and assigning to different conditions does not guarantee the equivalence of groups. Also, the internal validity is compromised and does not unequivocally link a difference between the groups with the effect of the intervention tested.

Level 4 indicates the presence of correlational studies or quantitative case studies that do not allow establishing the causality between an intervention and its effect. As intervention involves causal reasoning supported by indications showing that such intervention produces such results ('if I do this, then the student should progress'), studies that do not show a directional link explaining the learning gains contribute very little to the orientation of the interventions. Note that in cases where variables of interest cannot be manipulated, such as gender for example, correlational studies are entirely adequate or even decisive.

Level 5 refers to various reports, think tanks and recommendations from the judgement of expert researchers or clinicians on a predefined question presumably in the absence of higher-level scientific evidence. Because of the complexity of educational issues, a conservative position seems warranted and it appears that, all things considered, opinions remain weaker than scientific observations. It should be noted at the outset that relying on experts' good reputation does not overcome inherent weaknesses in this type of consensus exercise **(DellaVigna and Pope, 2018)**. In addition, DellaVigna and Pope **(2018)** show that groups always perform better at predicting a rank-ordering of the efficacy of treatments than single individuals, even when these individuals are recognized experts. The judgements formulated by groups of experts take the form of projections, hypotheses and extensions of the available data, which are subject to a large number of biases, including, to begin with, the choice of experts consulted. However, groups of experts may be used very productively to answer another type of questions. DellaVigna, Pope and Vivalt **(2019)** propose a methodology to use expert judgement in novel ways in the conduct and dissemination of research results that may improve the use of evidence at higher levels.

Level 6 refers to the exclusive reliance on qualitative studies. Their interpretative nature shows what is possible but not necessarily probable in terms of the effect of given interventions. By nature, a qualitative study does not aim to generalize results, but instead explain a specific situation in its context. This can lead to the identification of pertinent variables to study experimentally **(Slavin, 2020)**. Alternatively, single-case designs are available, which demonstrate the effect of an intervention experimentally and clearly, but are not generalizable, unless a large number of single-case studies are available to submit to a meta-analytical approach, in which case they will lack representativeness.

Level 7 implies the availability of relevant empirical observations that can serve to instigate future research but with problematic methodological origins. It should be noted that the levels of scientific evidence beyond this level apply only to well-conducted scientific studies.

By nature, a qualitative study does not aim to generalize results, but instead explain a specific situation in its context.

Level 8 indicates the absence of systematic empirical observations. Thus, this level includes professional success stories, principled positions, trending and hot topics, or media attention to strategically selected research results.

The applied and professional fields, which rely on scientific knowledge, can view research results according to the levels of scientific evidence presented. Levels of scientific evidence help establish a level of confidence in research results, which seems essential given that many professionals in education (including special education teachers) work with vulnerable populations of learners. It should be noted that a professional stance based on knowing and applying research-based practices reinforces, rather than diminishes, the importance of professional judgement. Professionals become responsible for knowing the aspects of their role that can be oriented by evidence and those that cannot. They also become responsible for applying evidence in their practice, an application that requires great expertise to match interventions with given needs, rank them according to their likely effect and contextualize the best intervention without threatening its active ingredients. The levels of evidence are also useful in helping researchers classify evidence that supports their own research processes and results. Finally, they are instrumental in guiding policy-makers in their decision process regarding educational practice and the appropriateness of interventions.

THEORY BUILDING AND THEORY TESTING, AND THE NEED TO MOVE UP ACROSS LEVELS OF SCIENTIFIC EVIDENCE IN EDUCATIONAL RESEARCH.

Going back to the need to know the likely effect of an intervention and, importantly, its mechanism as a condition for its implementation (i.e. general effectiveness claims), it is therefore possible to conclude that potential best practices based on evidence will initially be drawn from cumulative and converging evidence originating from experimental research (participants randomly assigned between groups), quasi-experimental research, and single-case studies on more or less proximal target outcomes. Such evidence is currently expressed in terms of effect size, a notion originally used to design better replications of a study in terms of statistical power **(Cohen, 1962)**, and recuperated following the need to establish practical significance **(Kirk, 1996)**. Technically, an effect size is the mean difference (standardized or in natural units) in outcome scores between a study's intervention and comparison groups **(Simpson, 2018)**. It is generally considered the best estimation of the effectiveness of an intervention, which can be compared across studies and interventions. However, Simpson argues that an effect size is mostly a measure of the clarity of the results of a study because it is also influenced by the psychometric characteristics of the outcome measures and characteristics of the samples, in addition to the effect of the intervention. Therefore, the effect size is the best solution to date, but technical improvements are warranted. It should be noted that the publication process likely inflates effect sizes because of a scarcity of reporting confidence interval and statistical power in the context of dichotomous statistical decision-making **(Fritz, Scherndl and Kuhberger, 2012)**. In other words, if statistically significant findings tend to get published more, then conditional on being published, effect sizes will be larger than they are in all of the studies (or, more importantly, statistical tests) undertaken. While traditional thinking underscored that confidence in research relied on it being of a high-quality standard (e.g. correct and faithful implementation) with solid psychometric measures and with little or no subject attrition, the present reasoning implies a major reconsideration of the veracity of research findings. Ioannidis **(2005)** has boldly demonstrated that, in principle, more than 50 per cent or research findings are very likely to be false as a result of bias such as research design, nature of the data, analysis strategy and reporting. Consequently, he concludes that confidence in research should arise from larger samples, larger effect sizes, more uniformity in research designs, definitions, outcome measures and analytical strategies. Because of the difficulty of conducting experimental studies in a school environment, we take a realistic stance to insist on the accumulation of quasi-experimental studies. Thus, within these constraints, the convincing

Although the evidence-based trend is widespread in education, its application by practitioners has been the subject of widespread criticism targeting in turn internal and external validity.

Often seen as a hierarchy of scientific methodological quality because of its grounding in internal validity, the applicability of EBE according to a policy (decision-maker) perspective is frequently overlooked.

nature of an intervention will typically be demonstrated by a large number of study results that demonstrate significant results or few studies that demonstrate mixed effects, with many studies demonstrating positive effects and no or few studies demonstrating negative effects.

Although the evidence-based trend is widespread in education, its application by practitioners has been the subject of widespread criticism targeting in turn internal and external validity. Internal validity is the extent to which an empirical study establishes and univocally explains a relationship between an intervention and its outcome; external validity refers to the possibility of applying the conclusions of an empirical study outside the context of the study.

Even in the case of the higher levels of evidence, the construct validity of studies regarding a given issue may be less than ideal: the definition of a given tested intervention may vary significantly across studies **(Davis, 2018; Simpson, 2018)** even if they stem from the same theoretical background. Thus, the cumulative evidence of desirable effects may be misleading in failing to capture the active ingredients in the approach as implemented in studies, departing from the apparently homogeneous theoretical definitions and further confounding the variability of impact across populations and contexts.

Often seen as a hierarchy of scientific methodological quality because of its grounding in internal validity, the applicability of EBE according to a policy (decision-maker) perspective is frequently overlooked **(Parkhurst and Abeysinghe, 2016)**. Indeed, evidence-based practice and evidence-based policy do not face the same challenges. Regarding evidence for policy-making, one may prefer to use the term evidence-informed because not only higher-level evidence is useful in the policy-making process. Higher-level evidence may be very useful to determine the effects of an intervention at the practical level **(Slavin, 2020)**, but evidence of a different nature

Another limitation of the hierarchy of scientific evidence is the external validity of the evidence.

is needed from a policy-making perspective depending on the context. Particularly in a field like education, where practice is based on policies, aspects such as popular opinion of practices, social determinants of target groups and other contextual variables are important to take into account **(Parkhurst and Abeysinghe, 2016)**. These aspects may therefore also hinge on high-quality evidence, but with respect to a different criterion corresponding to a different type of assertions. As will be discussed in the next section, assertions related to a particular context have to be seen as complementing previous levels of evidence that support relative effectiveness generalizations. Doing so will contribute to developing the educational policies on which practices are ultimately based.

Another limitation of the hierarchy of scientific evidence is the external validity of the evidence **(Joyce, 2019)**. Higher-level evidence aims at increasing the internal validity of studies to better demonstrate the effect of an intervention, but the external validity of these studies remains limited **(Orr, 2015)**. In the biomedical field, for example, there is an expectation that one entity will be similar to another (e.g. one human body is similar to another). This allows extrapolation of the results obtained in the laboratory to other contexts. In psychosocial fields (e.g. education), these similarities between entities are harder to demonstrate. Hence, interventions are more likely to produce different results in different groups, contexts, and so on. In such cases, results from experimental studies are not always isomorphically transposable or transportable to 'real-life' contexts **(Schmuckler, 2001)**. Even meta-analyses are susceptible to introducing biases regarding the external validity of a body of research since they pool studies conducted in several contexts that are not necessarily comparable **(Parkhurst and Abeysinghe, 2016).**

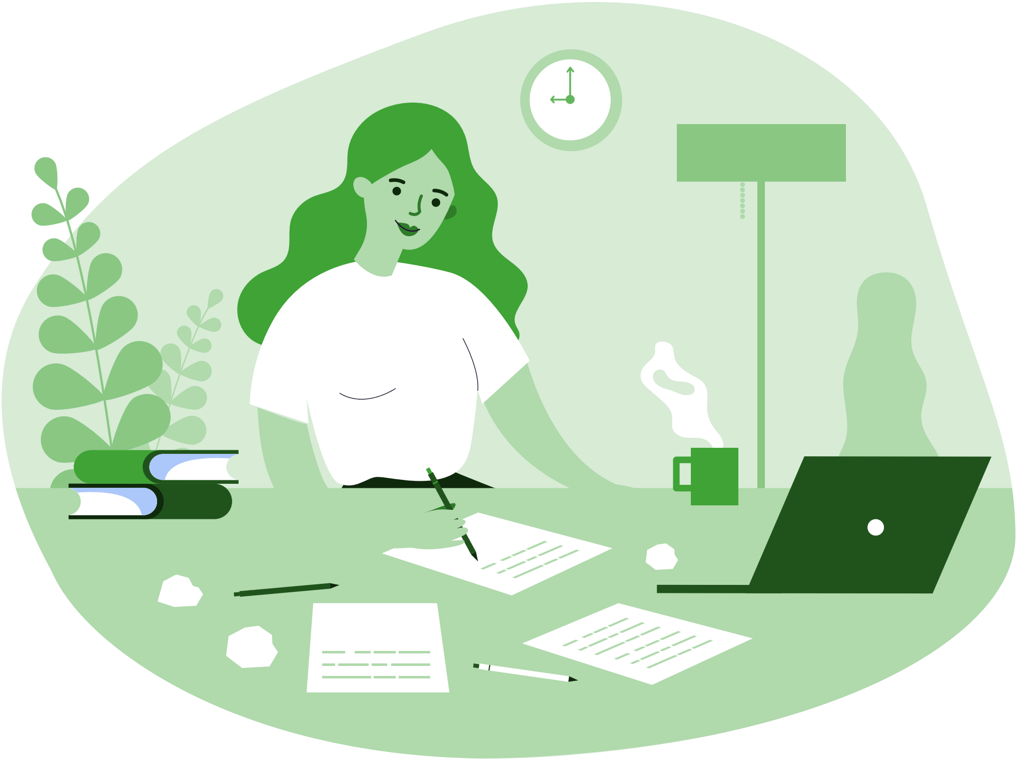
Another aspect that can affect the external validity of meta-analyses is the publication bias from the articles they include **(Gage, Cook and Reichow, 2017)**. Publication

Thus, cumulative evidence of desirable effects may be misleading by not capturing the active ingredients in a given approach as implemented in studies that deviate from seemingly homogeneous theoretical definitions, thereby further confusing the variability of the impact between populations and contexts.

bias is defined as the fact that articles with greater effect sizes or statistical significance are more likely to be published, with articles with mixed results or statistically in significant results less likely to be published. Although both scientific and probative levels of evidence are affected by publication bias, the meta-analytic process can be particularly affected by it because, without rigorous pre-specification and inclusion of grey literature, it can carry this bias by selecting articles from among an already biased pool of published articles. By doing so, meta-analytic results can boost the effect size tainted by the publication bias **(Fritz, Scherndl and Kuhberger, 2012)**.

Thus, cumulative evidence of desirable effects may be misleading by not capturing the active ingredients in a given approach as implemented in studies that deviate from seemingly homogeneous theoretical definitions, thereby further confusing the variability of the impact between populations and contexts. All the previous caveats can, in principle, be alleviated by recourse to relevant theory. Indeed, these caveats stem at least in part from definitional issues related to critical aspects of empirical work, such as population characteristics, interventions, outcomes, control variables and contexts.

In sum, the first aspect of next-generation EBE is the provision of general relative effectiveness claims (which takes the form of a new, more stringent, probative level in the framework), indicating that an intervention has a stable causal capacity relative to all other comparable interventions. This is a significant improvement over traditional EBE based on 'what works', which culminated with a miscellaneous collection of interventions essentially shown to be better than nothing. What is needed to complement these general relative effectiveness claims are credible assertions about how a local context affords a causal pathway through which the most effective intervention can make a positive contribution.



1.4 Will it work here? How the local context affords a causal pathway through which the intervention can make a positive contribution

Ultimately, we don't just want to know if an intervention works, we want to know if it will work in the specific context in which it is intended to be used.

Ultimately, we don't just want to know if an intervention works, we want to know if it will work in the specific context in which it is intended to be used. This question implies a shift toward a context-focused approach to EBE **(Joyce and Cartwright, 2020)**, which, in our proposed framework, is the necessary complement to the general relative effectiveness claims discussed earlier. Answering the question 'will it work here and now?' amounts to demonstrating, by means of empirical data or literature, how the local context affords a causal pathway through which an intervention documented as effective can make a positive contribution. The inferences made through this reasoning have been termed local effectiveness predictions by Joyce and Cartwright **(2020)**. While local effectiveness predictions will never be certain, incorporating this information in the reasoning supporting the implementation of evidence-based practices can improve them **(Joyce and Cartwright, 2020)**.

Proponents of EBE generally attribute the gap between research and practice results to shortcomings in the way tasks are performed in either knowledge production or knowledge use in practice **(Joyce and Cartwright, 2020)**. However, we argue that a major part of the necessary reasoning in EBE, formulating local effectiveness predictions, has been overlooked. With this in mind, qualitative research, which appears to be lower-level evidence in the context of establishing what works best **(see Table 1)** becomes mandatory in our proposed framework to attain higher levels of evidence in the context of establishing a fit with local context **(see Table 2)**. For example, ethnographic approaches or local surveys are also needed in order to assemble a body of evidence supporting the utility of an intervention in a specific context **(Parkhurst and Abeysinghe, 2016).**

What kind of reasons can support projectability and transportability of extant research in educational contexts? Results from a sample representing a given population permits generalizing results to that population, but not transporting

… the argument theory of evidence specifies that ‘a fact counts as evidence for a specified claim when it speaks to the truth of that claim’…

results to specific targets within it **(Pearl and Bareinboim, 2014)**. To this end, and because a progression to higher levels of evidence does not provide effectiveness predictions (transportability is a causal, not statistical notion) **(see Pearl and Bareinboim, 2014)**, a complementary, mostly inductive rationale is needed. As discussed by Joyce and Cartwright **(2020)**, the argument theory of evidence specifies that ‘a fact counts as evidence for a specified claim when it speaks to the truth of that claim’ **(p. 1051)**. Additionally, the material theory of induction underscores the importance of empirical work; observations are encoded in substantive claims that connect the evidence with the hypothesis **(Norton, 2003)**. Considered in this light, a research result is evidence relative to a target hypothesis and to a set of additional claims describing material facts about the world **(Joyce and Cartwright, 2020)**. In considering local effectiveness predictions, the hypothesis to be evidenced is: the outcomes specified in claims about relative effectiveness generalizations will occur within a local context. As illustrated next within the discussion of the realist approach, the evidence needed to test this hypothesis may come from empirical research, observations and credible theory. A formal graph-based procedure may also be used to logically encode and analyse differences between contexts **(Pearl and Bareinboim, 2014)**. Given the state of the research in education, in which mechanisms and processes are generally not sufficiently understood, this procedure may be best used for the moment to foster the necessary types of research, rather than to warrant the transportability of results across contexts.

A realist approach to the review and synthesis of evidence from the literature and to the evaluation of implementation of a given intervention seems particularly productive to answer the question ‘will it work here?’ The goal of a realist review is to explore the contexts that trigger certain mechanisms and the resultant ‘outcomes of interventions’ **(Defever and Jones, 2021, p. 9)**. Moreover, in light of the need for evidence of contextual fitting in EBE, the realist review appears to be a mandatory analysis following systematic review and meta-analysis in our proposed framework. In that sense, coupling systematic reviews and meta-analyses with realist reviews is the only way to be fully probative in EBE. The approach underlying a realist review focuses on the same key aspect as the levels of evidence, that is, causality between interventions and outcomes. Indeed, mechanisms, in the realist approach, represent causal processes **(Caswell et al., 2020)** in the form of structure, culture and agency **(De Souza, 2016)**. According to De Souza **(2016)**, these pre-existing conditions establish boundaries that contribute to constraining or enabling the effectiveness of different aspects of a complex programme. To strengthen the impact of EBE, these conditions need to be reported as evidence in research findings. 'Gaining insights about the contexts within which programmes are implemented can point to the conditions needed to help trigger its potential successful workings. It also enables explanations about the conditions existing that might be hindering the intended integration, uptake, or outcome of the program.' **(De Souza, 2016, pp. 226-227)**. In our view, it is the process of looking beyond variables that are studied, compared or controlled in quantitative work.

A realist synthesis is a narrative summary focused on interpretive theory that applies a realist philosophy to the synthesis of primary study results that affect a single applied research question. Realist review and classic systematic reviews procedures are relatively similar. An essential difference, however, is an insistence on the notion that experimental results are always context-dependent and that interventions are never implemented in the same context **(Smets and Struyven, 2018)**. A realist review uses an interpretive inter-case comparison to understand and explain, how and why the observed results occurred in the studies included in a literature review **(Wong et al., 2012)**. Realist

evaluation provides a framework for understanding how the context and underlying mechanisms affect the outcomes of an intervention **(Ericson et al., 2017)**. In trying to understand why policy programmes are usually not implemented as designed, Verger, Bonal and Zancajo **(2016)** emphasize one aspect of the realist approach, the agency of actors. These authors insist on the notion that the application of policy programmes is mediated by the previous experiences, values and interests of the subjects, and by the ways in which they interpret the rules of the programme.

These methods were originally developed by Pawson and Tilley to evaluate complex intervention policies in health and social services **(Pawson and Tilley, 1997; ;**

... the success of an intervention depends on how participants interact with it in local contexts, and a realist approach should uncover these processes.

Pawson et al., 2005; Pawson, 2006). In a realist approach, data is collected and analyzed in order to determine context–mechanism–process effect configurations **(Haynes et al., 2017)**. An explanation and understanding of the interaction between the context, the mechanism and the impact of the intervention is then produced **(Wong et al., 2012)**. This joint focus on context, mechanism and process effect should overcome one crucial limitation of quantitative research: authors have argued that traditional study designs such as randomized controlled trials, and non-randomized and prospective cohort studies, although useful, depending on the objective of the evaluation, overlook a key element, namely being able to identify contextual information that is useful when replicating the results in another context **(Graham and McAleer, 2018)**.

In other words, the success of an intervention depends on how participants interact with it in local contexts **(Haynes et al. 2017)**, and a realist approach should uncover these processes.

The working hypothesis behind a realistic synthesis is that a particular intervention (or class of interventions) will trigger particular mechanisms somewhat differently in different contexts. In realism, it is the mechanisms that trigger change rather than the interventions themselves, and realistic reviews therefore focus on 'families of mechanisms' rather than 'families of interventions' **(Wong et al., 2012)**.

1.4 .1

LEVELS OF CONTEXTUAL FITTING APPLIED TO EDUCATIONAL RESEARCH TOWARD LOCAL EFFECTIVENESS PREDICTIONS

In the same way that levels of evidence establish the information

TABLE 2 LEVELS OF CONTEXTUAL FITTING APPLIED TO EDUCATIONAL RESEARCH

RANGE	LEVEL	EVIDENCE REQUIRED	MAIN LIMITATIONS
Probative	1	Realist review	
Scientific	2	Qualitative research during implementation work	Correspondence between studied population/context established for the target population, but without taking into account all contextualized elements from the literature
	3	Qualitative research during experimental work	Correspondence between studied population/context established only from the population studied
	4	Exclusive reliance on relative effectiveness generalizations	Correspondence between studied population/context unestablished
Pseudo-scientific/ non-scientific	5	Exclusive reliance on causal ascriptions and general effectiveness claims	Based on arbitrary[1] choices among 'what works'

[1] Arbitrary is meant to include, but is not restricted to epistemological biases, personal preferences, emphasizing the latest research or more globally acting without the required information.

needed to make relative effectiveness generalizations, **Table 2** proposes a classification of the contextual fitting of effective interventions based on scientific evidence. Akin to the previous levels of evidence, this proposal of criteria allows us to distinguish between: (1) information of pseudoscientific/non-scientific

The qualitative work in this level is very similar to that in level 3, with the important difference that the observations are conducted in the context of application.

nature; (2) the results emanating from a scientific approach; and (3) the probative level in which the relative convergence and divergence of results is uncovered based on a thorough literature review. The facts needed to improve the level of contextual fitting come from empirical research, observations and credible theory.

As shown in **Table 2**, level 5 is considered pseudo/non-scientific, whereas levels 2 to 4 are deemed scientific. The probative range is limited to level 1.

Level 1. This level is the only one to provide probative information necessary to test the hypothesis that the outcomes specified in claims about relative effectiveness generalizations will occur within a local context. The information is probative because it is based on a review of the literature, and can be considered the best way to identify, define and establish the salience of the variables involved in effectiveness predictions.

Level 2. The qualitative work in this level is very similar to that in Level 3, with the important difference that the observations are conducted in the context of application.

Level 3. Level 3 involves qualitative research during quantitative experimental work, a strategy underlying mixed-methods research. While the quantitative approach provides causal ascriptions, qualitative work establishes, inductively, a complementary model to explain the results. The limitation, especially in comparison with level 2, is that this explanation is part of an ‘external’ study, the results of which have to be transported to the context of application.

Level 4. In level 4, the reliance on relative effectiveness generalizations established from meta-analytic work and syntheses does not provide evidence of the transportability of a relatively effective intervention to a new context, beyond a collection of sampling variables that may not be salient in making effectiveness predictions.

Level 5. Given the limitations of causal ascriptions and general effectiveness claims presented earlier, especially with respect to a lack of information about how a given intervention compares to others (and not just to business-as-usual teaching), level 5 posits that choosing an intervention to replace the one currently implemented in this context is so likely to be suboptimal that the status quo is probably better. As such, the hypothesis that the outcomes specified in claims about relative effectiveness generalizations will occur within a local context cannot even be tested.

As this proposal for levels of contextual fitting aims to demonstrate, for credible evidence-based policy or practice, the assumption that populations are alike must be supported **(Joyce, 2019)** by theory and other empirical results. Judging when generalized results from studies and specific applied settings are similar enough and in the right ways requires theory – lots of it and of very different kinds. Key aspects of the realist approach are linked to the use of theory in the form of context–mechanism–process effect configurations **(Haynes et al., 2017)**.

As this proposal for levels of contextual fitting aims to demonstrate, for credible evidence-based policy or practice, the assumption that populations are alike must be supported.

1.4 .2

THEORY BUILDING AND THEORY TESTING, AND THE NEED TO MOVE UP ACROSS LEVELS OF CONTEXTUAL FITTING IN EDUCATIONAL RESEARCH

Effectiveness predictions are obtained through the identification of contextual influences **(Joyce and Cartwright, 2020)**. Because we contend that contextual fitting necessarily occurs after obtaining the best level of evidence for relative

While local effectiveness predictions will never be certain, we propose that the sources of information used to formulate them can inform us about their accuracy and potential for transportability.

effectiveness generalizations, we specify the identification levels of contextual fitting as a process of disaggregation of contextual influences. This takes place through cumulative abstraction, in which relative effectiveness generalizations are 'reverse-engineered' once the target best intervention has been determined. The disaggregation of contextual influences through a realist review involves analyzing intervention characteristics that generate observed changes (i.e. mechanisms) and can inform the development or refinement of a conceptual framework **(Defever and Jones, 2021)**.

Also, we suggest that this process of disaggregation cumulatively leads to an increase in what we call levels of contextual fitting. Incorporating this information into the reasoning that supports the implementation of evidence-based practices will, in principle, improve the likelihood of replicating documented outcomes **(Joyce and Cartwright, 2020)**. While local effectiveness predictions will never be certain, we propose that the sources of information used to formulate them can inform us about their accuracy and potential for transportability. This increase in levels of contextual fitting hinges on theory building in the sense that identifying the causal mechanisms behind the effectiveness of an intervention constitutes the main asset for transporting (from one context to another by re-examining the variables, different from generalizing across contexts) effectiveness predictions. An increase in levels of contextual fitting signifies more reliable predictions about what might work in a given school or district, and with targeted students and, as Joyce and Cartwright **(2020)** insist, how it might work.

1.5 Conclusion: the framework and its implications

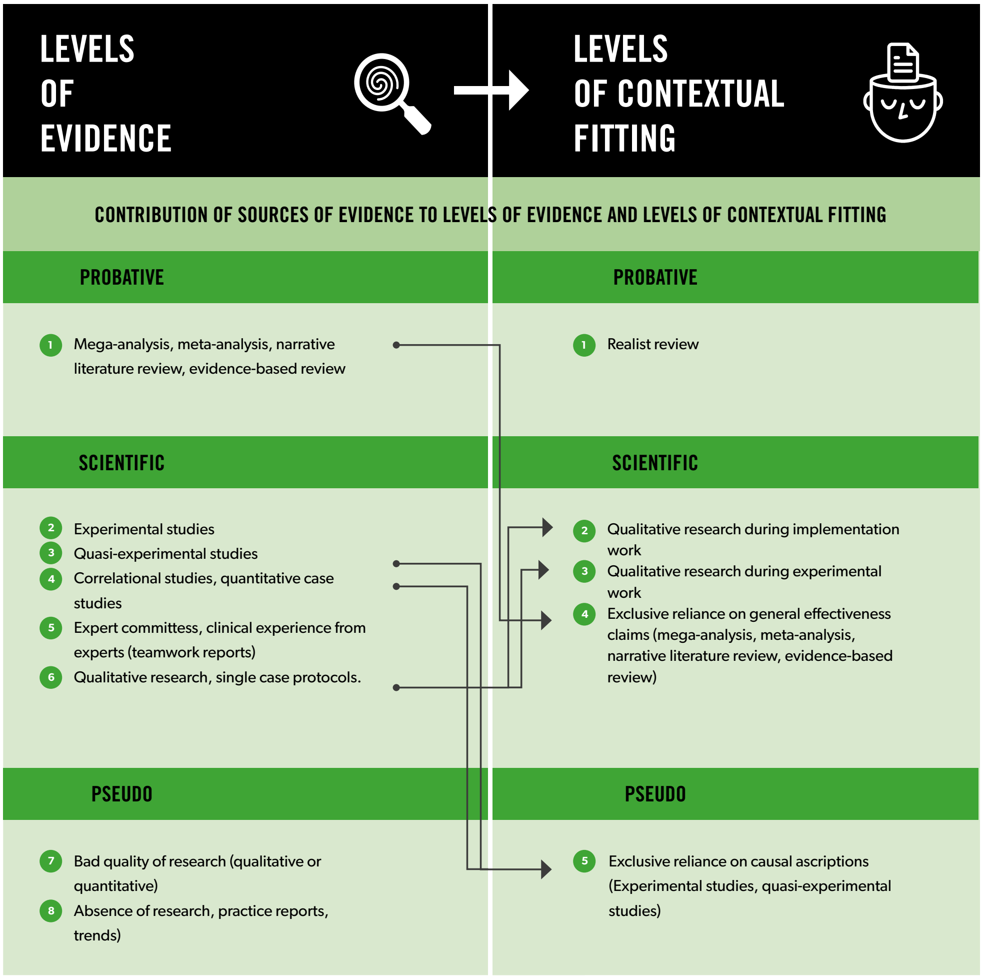
This chapter tackles the issue that while reviewing the scientific literature, it is sometimes difficult stakeholders such as policy-makers and practitioners to apply the evidence for the best possible effects in specific contexts, given the plethora of studies available. This chapter considered the importance of scientific theory in

... the question ‘will it work here?’ is now posited as absolutely necessary to complement the information and reasoning pertaining to ‘what works best’.

explaining phenomena, and the contribution of empirical research to theory building and theory testing. It then examined the levels of scientific evidence and the need to accumulate appropriate evidence across these levels in order to support specific inferences in education. Finally, it discussed the need to fit general relative effectiveness claims to specific contexts of application.

Articulating the two main ingredients of next-generation EBE posited in this paper – general effectiveness claims and effectiveness predictions – in an effort to go beyond ‘what works’ leads to a new articulation of applied empirical research within a given educational field, as seen in **Figure 1**. A few notable proposals emerge from the current work. Within the traditional view of levels of evidence, the probative level now concerns only relative effectiveness generalizations (i.e., a rank-ordering (generalizable to a population) of the effectiveness of all pertinent interventions), and not effectiveness generalizations (how a given intervention compares to a control). This places the meta-analytic approach as key to the provision of the required information to answer the most important question: what works best? Consequently, the gold standard of EBE, the randomized controlled trial, is no longer in the probative range. In addition, the conceptualization and operationalization of the levels of contextual fitting, in response to the need for local effectiveness predictions, can be seen as the most important contribution of the current work. Its most constructive implication is that the synergy between quantitative and qualitative approaches in applied research is more apparent. Also, the question ‘will it work here?’ is now posited as absolutely necessary to complement the information and reasoning pertaining to ‘what works best’.

The proposed articulation of causal ascriptions, relative effectiveness generalizations and local effectiveness predictions generated by empirical research in education in the form of the EBE3 framework has implications



LEVELS OF EVIDENCE	LEVELS OF CONTEXTUAL FITTING
CONTRIBUTION OF SOURCES OF EVIDENCE TO LEVELS OF EVIDENCE AND LEVELS OF CONTEXTUAL FITTING	
PROBATIVE	**PROBATIVE**
1 Mega-analysis, meta-analysis, narrative literature review, evidence-based review	1 Realist review
SCIENTIFIC	**SCIENTIFIC**
2 Experimental studies 3 Quasi-experimental studies 4 Correlational studies, quantitative case studies 5 Expert committess, clinical experience from experts (teamwork reports) 6 Qualitative research, single case protocols.	2 Qualitative research during implementation work 3 Qualitative research during experimental work 4 Exclusive reliance on general effectiveness claims (mega-analysis, meta-analysis, narrative literature review, evidence-based review)
PSEUDO	**PSEUDO**
7 Bad quality of research (qualitative or quantitative) 8 Absence of research, practice reports, trends)	5 Exclusive reliance on causal ascriptions (Experimental studies, quasi-experimental studies)

Figure 1

for conducting future research, for policy-making and for improving educational practice.

Concerning the orientation of applied scientific research, the framework in **Figure 1** may shed light on the need for specific

By reviewing and integrating the state of the art in EBE, it becomes clear that quantitative and qualitative research leverage each other in achieving the cumulative steps necessary for better intervention in a given domain.

kinds of quantitative studies, meta-analyses and synthesis of work, as well as qualitative implementation work. Thus, it helps in bridging the perceived divide between quantitative and qualitative research in education by suggesting a sound integration of quantitative and qualitative methodologies around a common applied goal: providing the necessary information for the improvement of educational intervention. By reviewing and integrating the state of the art in EBE, it becomes clear that quantitative and qualitative research leverage each other in achieving the cumulative steps necessary for better intervention in a given domain. As De Souza **(2016)** notes, methodologies for realist evaluation and review are still in development and are likely to make increasing contributions to the application of empirical research.

In light of the importance of meta-analyses and systematic reviews underlined when discussing the need for effectiveness generalizations, it should be noted that the realist review process presented as a method for establishing effectiveness predictions can be reused to facilitate the automation of meta-analyses and enable living reviews of evidence. The realist approach has provided a consistent rationale for synthesizing evidence across forms and types of interventions **(Pearson et al., 2015)**. Indeed, realist reviews can be key in standardizing coding frameworks for studies, with common coding of cohorts, intervention delivery mechanisms and core components. In addition, the framework presented in **Table 2** helps in focusing research efforts directly on a frequently overlooked issue, that is, how to build local effectiveness predictions. It outlines various kinds of information that can improve predictions and encourages using appropriate methods for acquiring that information.

With respect to policy-making, the framework presented in **Tables 1** and **2** may feed into the mechanisms identified by Langer, Tripney and Gough

Finally, practice should be greatly improved by a widened view of the necessary evidence in the implementation of so called best practices, especially regarding effectiveness predictions.

(2016) as facilitating research use by policy-makers, beyond the preconditions regarding enhancing decision- makers' opportunity, capability and motivation to use evidence. By insisting on a more complete scientific demonstration of efficacy, from causal ascriptions to effectiveness generalizations and effectiveness predictions, the framework may provide the materials for interventions facilitating access to research evidence and for interventions building decision-makers' skills to access and make sense of evidence.

At the level of organizations and systems, this more complete scientific demonstration of efficacy outlined in **Table 2** may help identify the right information for the right people that can be used in the design of interventions that foster changes to decision-making structures and processes. Notably, an increased focus on core components, that is, mechanisms that represent active ingredients in interventions, can help policy-makers avoid biases toward scientific disciplines that may seem compelling but do not provide the best explanations about how interventions work and why. The consequences of evidence-based reform refined operationally in this paper could be profound. If educational policies begin to favour programmes with clear evidence, publishers, software developers, university researchers and entrepreneurs will have an incentive to engage in serious development and evaluation efforts. Governments, seeing the cumulative impact of such research and development, might provide substantially greater funding for these activities in education.

Finally, practice should be greatly improved by a widened view of the necessary evidence in the implementation of so-called best practices, especially regarding effectiveness predictions. Effectiveness predictions help frame practitioners' reasoning concerning the match between general, abstracted evidence and their own specific and idiosyncratic context around a specific kind of inference that is amenable to analysis and testing in

In sum, the EBE3 framework presented in this paper may be one of the most integrative in terms of research traditions and with respect to the different roles (teachers, researchers, policy-makers) involved in EBE.

the context of day-to-day practice.

Evidence brokerage is also crucial to bridge the gaps between research and practice **(Langer, Tripney and Gough, 2016)**. Because the EBE3 framework identifies the reasoning and the supporting information for next-generation EBE, it could be used in information design, to enhance the structure of evidence repositories and other resources. Langer, Tripney and Gough **(2016)** also conclude that interaction among professionals can build a professional identity with common practices and standards of conduct fostering EBE. Using social influence and peer-to-peer interaction as catalysts, districts may be able to use support specialists (e.g. curriculum specialists, programme specialists) and schools may be able to use onsite personnel, including literacy facilitators or highly effective general or special education teachers (peers) as coaches. The focus could then be on those teachers who need follow-up support instead of providing the same support for all teachers across all professional development activities.

In sum, the EBE3 framework presented in this paper may be one of the most integrative in terms of research traditions and with respect to the different roles (teachers, researchers, policy-makers) involved in EBE. Future work should evaluate the implications of such an integration in terms of its conceptual, operational and organizational aspects.

1.6 Key messages

The results of a collection of high-quality studies comparing an experimental group given a target intervention with a control group (usually receiving business-as-usual teaching) has been the cornerstone of EBE for decades under the label 'what works'. It is the main, but not sufficient, building block of EBE, and there is a need for a higher minimum standard for what counts as evidence of improved learning.

For a given educational issue, what is needed is a complete inventory of available interventions, rank-ordered in terms of relative efficacy to answer the question 'what works best generally?'.

An EBE initiative is not complete without solid indications that a specific application context will enable the 'working best in general' intervention to yield the expected benefits. This will answer the question 'will it work here'. Concretely, a realist review should be seen as complementary to a systematic review and meta-analysis and therefore should be conducted in tandem.

1.7

Key recommendations

The potential of the EBE3 framework to go beyond 'what works' will be fully realized by:

emphasizing effectiveness generalizations by the production of meta-analytic work as soon as there are enough published experimental studies on a given issue; and

emphasizing effectiveness predictions by undertaking qualitative work relating to effectiveness predictions in given contexts as soon as meta-analytic results are available.

The potential of the EBE3 framework to provide greater cohesion to applied empirical work on a given issue will be fulfilled by:

- ***focusing on theory building and theory testing*** in conducting empirical studies, despite the applied nature of educational research.

- ***aligning the goals/research questions of quantitative and qualitative research*** with the maturity of a field to optimize the outcomes when applied to educational interventions.

REFERENCES

Anderson, J.E. (2011) Public policymaking. Boston: Wadsworth Cengage Learning.

Borenstein, M. (2019) Common mistakes in meta-analysis and how to avoid them. Englewood: Biostat.

Brighouse, H., Ladd, H.F., Loeb, S. and Swift, A. (2018) Educational goods: values, evidence, and decision-making. Chicago: University of Chicago Press.

Burns, P.B., Rohrich, R.J. and Chung, K.C. (2011) 'The levels of evidence and their role in evidence-based medicine', Plastic Reconstruction Surgery, 128(1). https://doi.org/:10.1097/PRS.0b013e318219c171.

Caswell, R.J., Maidment, I., Ross, J.D.C. and Bradbury-Jones, C. (2020) 'How, why, for whom and in what context, do sexual health clinics provide an environment for safe and supported disclosure of sexual violence: protocol for a realist review', BMJ Open, 10. https://doi.org/10.1136/bmjopen-2020-037599.

Chubb, J. and Watermeyer, R. (2017) 'Artifice or integrity in the marketization of research impact? Investigating the moral economy of (pathways to) impact statements within research funding proposals in the UK and Australia', Studies in Higher Education, 42. https://doi.org/10.1080/03075079.2016.1144 182.

Cochrane, A.L. (1972) Effectiveness and efficiency: random reflections on health services. London: Nuffield Trust.

Cohen, J. (1962) 'The statistical power of abnormal-social psychological research: a review', Journal of Abnormal and Social Psychology, 65(3), pp. 145–153.

Colquitt, J.A. and Zapata-Phelan, C.P. (2007) 'Trends in theory building and theory testing : a five-decade study of the Academy of Management Journal', Academy of Management Journal, 50(6). https://doi.org/10.5465/amj.2007.28165855.

Connolly, P., Keenan, C. and Urbanska, K. (2018) 'The trials of evidence-based practice in education: a systematic review of randomised controlled trials in education research 1980–2016', Educational Research, 60(3). https://doi.org/10.1080/00131881.2018.1493353.

Davis, A. (2018) 'Evidence-based approaches to education: direct instruction, anyone?', Management in Education, 32(3), pp. 135–138.

De Souza, D. (2016) 'Critical realism and realist review: analyzing complexity in educational restructuring and the limits of generalizing program theories across borders', American Journal of Evaluation, 37(2), pp. 216−237.

Defever, E. and Jones, M. (2021) 'Rapid realist review of school-based physical activity interventions in 7 - to 11 - year-old children', Children, 8. https://doi.org/10.3390/children8010052.

DellaVigna, S. and Pope, D. (2018) 'Predicting experimental results: who knows what?', Journal of Political Economy, 126(6). https://doi.org/10.1086/699976.

DellaVigna, S., Pope, D. and Vivalt, E. (2019) 'Predict science to improve science', Science, 366(6464). https://doi.org/10.1126/science.aaz1704.

Ericson, A., Löfgren, S., Bolinder, G., Reeves, S., Kitto, S. and Masiello, I. (2017) 'Interprofessional education in a student-led emergency department: a realist evaluation', Journal of Interprofessional Care, 31(2). https://doi.org/10.1080/13561820.2016.1250726.

Evans, D. (2003) 'Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions', Journal of Clinical Nursing, 12, pp. 77–84.

Frias-Navarro, D., Llobell, J., Pascual-Soler, M., Perez-Gonzalez, J. and Berrios-Riquelme, J. (2020) 'Replication crisis or an opportunity to improve scientific production?', European Journal of Education, 55. https://doi.org/10.1111/ejed.12417.

Fritz, A., Scherndl, T. and Kühberger, A. (2012) 'A comprehensive review of reporting practices in psychological journals: are effect sizes really enough?', Theory & Psychology, 23(1), pp. 98–122.

Gage, N.A., Cook, B.G. and Reichow, B. (2017) 'Publication bias in special education meta-analyses', Exceptional Children, 83(4), pp. 428−445.

Goldacre, B. (2013) 'Building evidence into education'. Available at: https://www.gov.uk/government/news/building-evidence-into-education (Accessed: 30 June 2021).

Graham, A.C. and McAleer, S. (2018) 'An overview of realist evaluation for simulation-based education', Advances in Simulation, 3(13). https://doi.org/10.1186/s41077-018-0073-6.

Hansson, S.O. (2009) 'Cutting the gordian knot of demarcation', International Studies in the Philosophy of Science, 23(3). https://doi.org/10.1080/02698590903196007.

Haynes, A., Brennan, S., Redman, S., Williamson, A., Makkar, S.R., ... and Butow, P. (2017) 'Policymakers' experience of a capacity-building intervention designed to increase their use of research: a realist process evaluation', Health Research Policy and Systems, 15(99). https://doi.org/10.1186/s12961-017-0234-4.

Ioannidis, J.P.A. (2005) 'Why most published research findings are false', PLOS Med, 2(8). https://doi.org/10.1371/journal.pmed.0020124.

Jahan, N., Naveed, S., Zeshan, M. and Tahir, M.A. (2016) 'How to conduct a systematic review: a narrative literature review', Cureus, 8(11). https://doi.org/10.7759/cureus.864.

Joyce, K. and Cartwright, N. (2020) 'Bridging the gap between research and practice: predicting what will work locally', American Educational Research Journal, 57(3). https://doi.org/10.3102/0002831219866687.

Joyce, K.E. (2019) 'The key role of representativeness in evidence-based education', Educational Research and Evaluation, 25(1–2). https://doi.org/10.1080/13803611.2019.1617989.

Karrigan, M.R. and Turner-Johnson, A. (2019) 'Qualitative approaches to policy research in education: contesting the evidence-based, neoliberal regime', American Behavioral Scientist, 63(3). https://doi.org/10.1177/0002764218819693.

Kirk, R.E. (1996) 'Practical significance: a concept whose time has come', Educational Psychological Measurement, 56(5), pp. 746−759.

Kornell, N., Rabelo, V. C. and Klein, P. J. (2012) 'Tests enhance learning: compared to what?', Journal of Applied Research in Memory & Cognition, 1(4), pp. 257–259.

Langer, L., Tripney, J. and Gough, D. (2016) The science of using science: researching the use of research evidence in decision-making. London: EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University College London.

Maciver, D., Rutherford, M., Arakelyan, S., Kramer, J.M., Richmond, J. and Todorova, L. (2019) 'Participation of children with disabilities in school: a realist systematic review of psychosocial and environmental factors', PLOS ONE, 14(1). https://doi.org/10.1371/journal.pone.0210511.

National Research Council (2002) Scientific research in education. Washington, DC: National Academies Press.

Norton, J.D. (2003) 'A material theory of induction', Philosophy of Science, 70. https://doi.org/10.1086/378858.

Nutley, S., Powell, A. and Davies, H. (2013) What counts as good evidence? Provocation paper for the alliance for useful evidence. Available at: https://www.alliance4usefulevidence.org/assets/What-Counts-as-Good-Evidence-WEB.pdf (Accessed: 30 June 2021).

REFERENCES

Orr, L.L. (2015) '2014 Rossi Award Lecture: beyond internal validity', Evaluation Review, 39(2). https://doi.org/10.1177/0193841X15573659.

Parkhurst, J.O. and Abeysinghe, S. (2016) 'What constitutes "good" evidence for public health and social policy-making? From hierarchies to appropriateness', Social Epistemology, 30(5-6). https://doi.org/10.1080/02691728.2016.1172365.

Pawson, R. (2006) Evidence-based policy: a realist perspective. London: Sage.

Pawson, R. and Tilley, N. (1997) Realistic evaluation. London: Sage.

Pawson, R., Greenhalgh, T., Harvey, G. and Walshe, K. (2005) 'Realist review: a new method of systematic review designed for complex policy interventions', Journal of Health Services Research & Policy, 10, pp. 21–34.

Pearl, J. and Bareinboim, E. (2014) 'External validity: from do-calculus to transportability across populations', Statistical Science, 29(4). https://doi.org/10.1214/14-STS486.

Pearson, M., Chilton, R., Wyatt, K., Abraham, C., Ford, T., Woods, H.B. and Anderson, R. (2015) 'Implementing health promotion programmes in schools: a realist systematic review of research and experience in the United Kingdom', Implementation Science, 10(1). https://doi.org/10.1186/s13012-015-0338-6.

Schmuckler, M.A. (2001) 'What is ecological validity? A dimensional analysis', Infancy, 2(4), pp. 419−436.

Shadish, W.R., Cook, T.D. and Campbell, D.T. (2002) Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin.

Simpson, A. (2018) 'Princesses are bigger than elephants: effect size as a category error in evidence-based education', British Educational Research Journal, 44(5). https://doi.org/10.1002/berj.3474.

Slavin, R.E. (2020) 'How evidence-based reform will transform research and practice in education', Educational Psychologist, 55(1). https://doi.org/10.1080/00461520.2019.1611432.

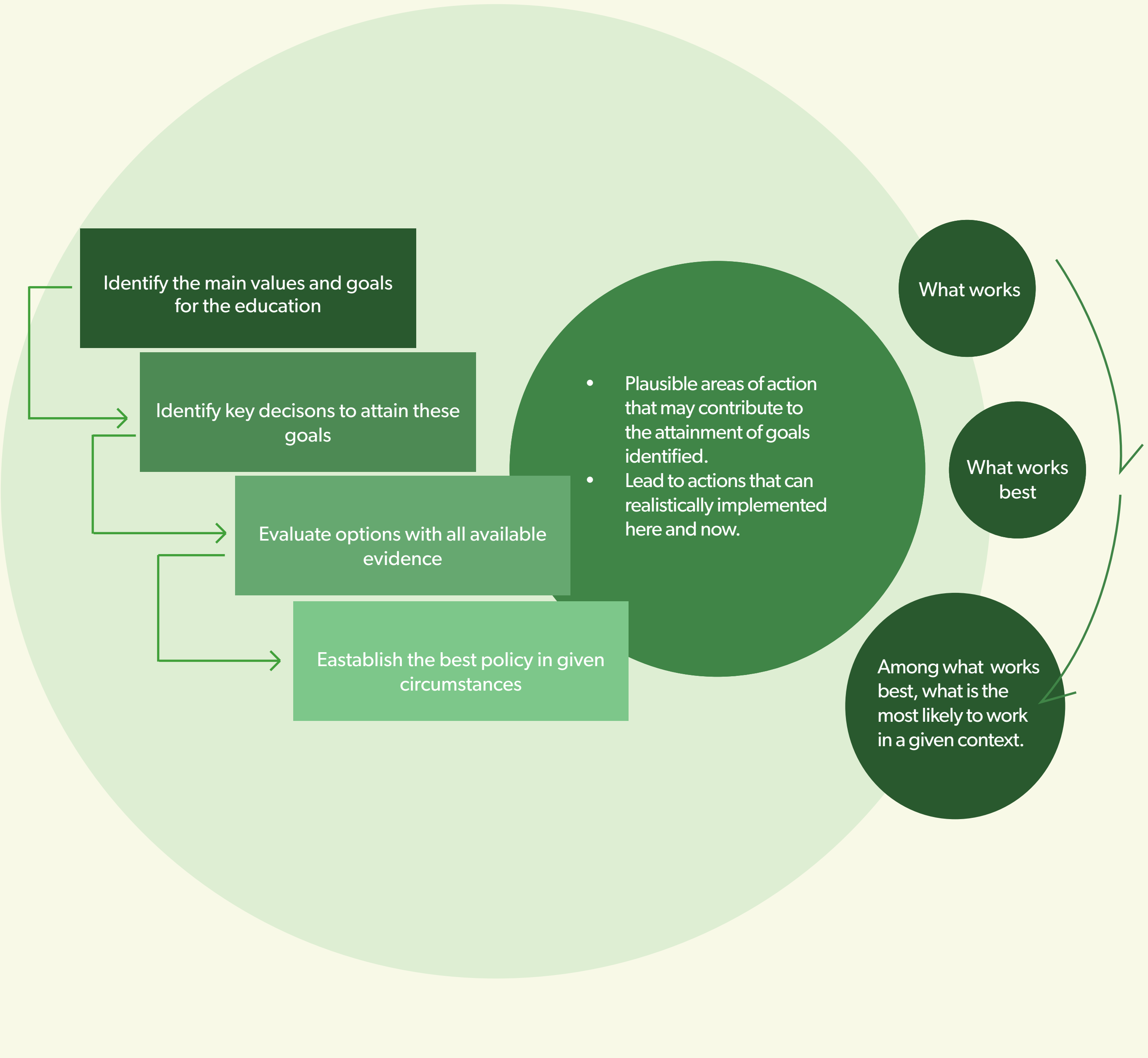
Smets, W. and Struyven, K. (2018) 'Realist review of literature on catering for different instructional needs', Educational Science, 8. https://doi.org/10.3390/edusci8030113.

Sugimoto, C. R. and Larivière, V. (2018) Measuring research: what everyone needs to know. Oxford: Oxford University Press.

Verger, A., Bonal, X. and Zancajo, A. (2016) 'What are the role and impact of public-private partnerships in education? A realist evaluation of the Chilean education quasi-market', Comparative Education Review, 60(2), pp. 223–248.

Wong, G., Greenhalgh, T., Westhorp, G. and Pawson, R. (2012) 'Realist methods in medical education research: what are they and what can they contribute?', Medical Education, 46. https://doi.org/10.1111/j.1365-2923.2011.04045.x.

A procedure for next generation evidence based education


Identify the main values and goals for the education
Identify key decisons to attain these goals
Evaluate options with all available evidence
Eastablish the best policy in given circumstances
• Plausible areas of action that may contribute to the attainment of goals identified.
• Lead to actions that can realistically implemented here and now.
What works
What works best
Among what works best, what is the most likely to work in a given context.

WORKING
GROUP 04

CHAPTER

2

How well does this intervention work? Statistical significance, uncertainty and some concepts to interpret the findings of evaluations of educational interventions[1]

DOI: https://doi.org/10.56383/MOBT3765

This chapter should be cited as:

Rodriguez-Guzman, G. (2022) 'How well does this intervention work? Statistical significance, uncertainty and some concepts to interpret the findings of evaluations of educational interventions' in, Duraiappah, A.K., van Atteveldt, N.M., Borst, G., Bugden, S., Ergas, O., Gilead, T., Gupta, L., Mercier J., Pugh, K., Singh, N.C. and Vickers, E.A. (eds.) Reimagining Education: The International Science and Evidence Based Assessment. New Delhi: UNESCO MGIEP.

Coordinating Lead Authors

Guillermo Rodriguez-Guzman

[1] This document discusses the role of uncertainty in the interpretation of empirical evaluations in education. This is an expanded and revised version of a statement on statistical significance first published by the Education Endowment Foundation (EFE). The EEF is an independent charity based in England, dedicated to breaking the link between family income and educational achievement.

The author and the EEF are grateful to Ronald Wasserstein, Nicole Lazar, Allen Schirm, and Steve Higgins for the invaluable comments and suggestions made to the original statement published by the EEF.

The authors are also grateful to Julien Mercier and Drew Bailey for outstanding recommendations to multiple versions of this paper; and to the three anonymous peer reviewers for other thoughtful suggestions.

2.1 Introduction: the importance of understanding and interpreting uncertainty

For anyone interested in how evidence can support more effective decision-making in education, the term 'statistical significance' will be a familiar one – and yet one probably shrouded in confusion. Despite the claims one might hear circulating in the media, policy circles and from different pundits, no study will give the ultimate and unquestionable truth about whether a programme or intervention will achieve a specific impact.

Policy decisions and prescriptions for action are often made on the basis of incomplete and imperfect information and the uncertainty around quantitative results is one of the key factors at play. As the eventual implementation of interventions may have positive

Despite the claims one might hear circulating in the media, policy circles and from different pundits, no study will give the ultimate and unquestionable truth about whether a programme or intervention will achieve a specific impact.

or negative impacts on learners, understanding uncertainty of impact estimates is integral to educational practice and policy-making. In principle, not considering this uncertainty means that policies and changes in practice, despite being based on research evidence, overlook relevant scenarios. This can lead to overly cautious decision-making in some cases or risk detrimental effects to learners in others.

Reflecting this complexity and uncertainty, researchers have been using 'statistical significance' to attempt to deal with uncertain, incomplete answers. But the use of statistical significance divides the research community in a range of disciplines, from statistics to social policy, including education. Some consider statistical significance an essential part of impact evaluation, just one aspect of a broader picture, while others regard it as a meaningless and misleading concept that should be abolished altogether **(Shrout, 1997; Ziliak and McCloskey, 2008; Trafimov and Marks, 2015; Gorard, 2016; Hubbard, 2016; Wasserstein and Lazar, 2016; Amrhein, Greenland and McShane, 2019; McShane et al., 2019; Wasserstein, Schirm and Lazar, 2019)**.

For the average classroom teacher, school leader or policy-maker, this lack of consensus among educational researchers is highly problematic, making it difficult to answer the very reasonable question: 'how well does this intervention work?'

This chapter outlines some key concepts underpinning notions of uncertainty, and proposes a way forward, which is then adopted in the subsequent chapter that presents estimates of impact, costs and certainty for a range of common education interventions and approaches. The key proposal is that impacts should be reported as effect sizes, and interpreted alongside internal validity and uncertainty when making a decision about a programme. We summarise relevant scholarship in this topic, which proposes moving away from a dichotomous interpretation of *p*-values and significance testing as the means to gauge the effectiveness of a programme.

2.2 How well did this intervention work? Some building blocks and an example

2.2 .1

KEY CONCEPT 1 – EFFECT SIZE

An effect size is a number that conveys the strength of the relationship between two variable factors. This number is obtained, for any given dependent variable, by scaling the difference between group means by the dispersion of the observations (the standard deviation).

In education, factors manipulated experimentally usually are subject to a specific intervention to measure the outcomes achieved by learners (e.g. educational attainment). In an experimental setting, this would usually compare the average in the intervention group and the average in the control group, scaled by how dispersed the results are (i.e. the standard deviation). The larger the effect size, the larger the difference between the two groups and the stronger the relationship between the intervention and the outcomes being measured.

Effect sizes are an important and useful metric because they enable us to move away from the simplistic question of whether something works or not (further complicated by the reliance on a dichotomous interpretation of statistical significance – more on this below). Instead, effect sizes help to answer the more relevant question 'how well did this work?' **(Coe, 2002; Major and Higgins, 2019; Higgins, 2021)**. Effect sizes are also useful as they provide a common metric to compare the relative effectiveness (see Chapter 1) of different interventions, which is more meaningful for decision-makers choosing between competing alternatives.

A key challenge regarding the use of effect sizes is that they describe differences in terms of standard deviations rather than measures that are more readily understood by the very audience who should be able to make the most of research results: policy-makers and teachers.

This is why, when communicating evidence of impact, it can be helpful to translate outcomes into other more meaningful measures while trying to introduce them into the common parlance of decision-makers.

... when communicating evidence of impact, it can be helpful to translate outcomes into other more meaningful measures while trying to introduce them into the common parlance of decision-makers.

2.2 .2

KEY CONCEPT 2 – MONTHS OF (STANDARD) PROGRESS AS A PRACTICE-ORIENTED TRANSFORMATION OF EFFECT SIZE

To overcome this communication challenge, the Education Endowment Foundation's (EEF) toolkit **(Major and Higgins, 2019; Higgins, 2021)** transforms effect size into a single scale of school progress: months of progress.

This transformation is done by dividing effect size, which is a measure of progress in terms of standard deviations, by the progress that could be expected in a school year for a given group of learners, also measured in standard deviations. The result is the amount of progress that would have been made in comparison to the average progress made in a year. That is, a standardized benchmark that allows drawing comparisons between multiple interventions in a metric that is easier to understand for teachers and decision-makers.

The average progress in a year is estimated to be around one standard deviation; and while this is likely to be a conservative estimate which may vary for different ages and types of tests, a crude measure is preferred to ensure findings remained more accessible and meaningful **(Major and Higgins, 2019; Higgins, 2021)**.

Other transformations and metrics have been proposed and reviewed by **(Bloom et al., 2008; Lipsey et al., 2012; Baird and Pane, 2019; Evans and Yuan, 2019)**. These include months of progress measures that account for differences across tests and the speed at which pupils learn over time, as well as alternatives like percentile ranges. These alternatives have their merits, as they address some of the methodological shortcomings of the simpler months of progress measure used by the EEF. However, this can also result in more complex interpretation, which is the problem these alternatives are trying to address. Stakeholders may decide to use one or several of these transformations, depending on the levels of literacy and exposure of the decision-makers they are seeking to inform or influence. For example, using months of progress as a metric, researchers can explain that an intervention that had an impact of 0.3 standard deviations could be represented as achieving the equivalent of three months' progress – a measure that is likely to be easily understood by practitioners and decision-makers.

Stakeholders may decide to use one or several of these transformations, depending on the levels of literacy and exposure of the decision-makers they are seeking to inform or influence.

A confidence interval is a range that is often used to measure uncertainty around an estimated value, such as an effect size or the mean of a distribution.

In addition to the 'mean' effect identified by an evaluation, quantitative researchers need to clearly express the uncertainty around those results – that is, other results that would be plausible under the statistical model being used and considering characteristics of the data.

2.2 .3

KEY CONCEPT 3 – 'CONFIDENCE' INTERVALS (OR 'COMPATIBILITY' INTERVALS)

A confidence interval is a range that is often used to measure uncertainty around an estimated value, such as an effect size or the mean of a distribution. This range of values is bounded above and below the statistic's mean. A 95 per cent 'confidence interval' includes a range of values for which 95 per cent of the confidence intervals computed from many hypothetical studies would contain the unknown population parameter if all the conditions under which the intervals are built hold. The interpretation of confidence intervals can be challenging and has been extensively criticized **(Greenland et al., 2016; Morey et al., 2016)** for reasons akin to the problems with *p*-values (see below).

2.2 .4

KEY CONCEPT 4 – P-VALUES AND STATISTICAL SIGNIFICANCE

Another standard way of assessing this uncertainty is using a *p*-value. These are measures of the compatibility between the observed data and a particular model of the data and are closely related to the idea of a 'confidence interval'. Both concepts are probabilities computed for many hypothetical studies under a set of conditions. We define these terms in greater detail in the section 2.4.

P-values are difficult to interpret for researchers and practitioners alike and have been widely criticized for misleading decision-making and biasing the literature, particularly given the tendency to interpret them in a dichotomous way due to a reliance on the idea of 'statistical significance' **(Greenland et al., 2016; Wasserstein and Lazar, 2016; Amrhein et al., 2019; Wasserstein, Schirm and Lazar, 2019).**

A result is deemed 'statistically significant' if the 95 per cent confidence interval does not include zero or if a p-value is below a given threshold, often 0.05, which is symmetrical to the 95% confidence

A result is deemed 'statistically significant' if the 95 per cent confidence interval does not include zero or if a *p*-value is below a given threshold, often 0.05, which is symmetrical to the 95 per cent confidence interval. When a result is 'statistically significant' it is often interpreted as meaning that the intervention 'had an effect'. As explained in section 2.4, this is not true. This dichotomous interpretation is at the heart of the problems with *p*-values, confidence intervals and significance testing.

Nonetheless, the interpretation of *p*-values could be seen as more heinous than confidence intervals because a range of values is more likely to be interpreted with caution. A range of values is more plausible than imprinting a false sense of certainty for decision-makers who observe a result that is 'statistically significant' and believe it to be the 'true' effect. This has been reflected in the preference of a growing number of journals to report confidence intervals instead of *p*-values **(Greenland et al., 2016).**

2.2.5

KEY CONCEPT 5 – INTERNAL VALIDITY

To evaluate the impact of a programme or intervention, researchers would like to compare the 'treatment' outcomes those without the 'treatment' or intervention. This scenario is called the counterfactual. Clearly, it is not possible to observe both scenarios in the real world, which requires researchers to compare

Most EEF-funded evaluations use a randomized controlled trial (RCT) design to estimate the impact of a programme; this is one of the most robust ways to identify a valid counterfactual.

the results of the group that was treated with those of a group identified as a suitable comparison (i.e. a valid counterfactual). The differences in outcomes between the treatment and the comparison groups, considering the mean outcome and its variability in each group, is interpreted as the estimate of impact and measured as an 'effect size'.

Most EEF-funded evaluations use a randomized controlled trial (RCT) design to estimate the impact of a programme; this is one of the most robust ways to identify a valid counterfactual. The evaluation design, in this case a RCT, is one of the crucial factors defining how confident we can be that the findings are a good representation of the impact of the intervention. However, to make this assessment, it is also important to consider other dimensions including:

- the overall size of the study;[2]

- whether the relevant information from participants is present, and, if not, understanding why (outcome attrition);

- whether appropriate and reliable outcome measures were used to track progress;

- whether those in the control group received the intervention being tested or experienced any other changes that could affect their behaviour and progress, such as non-compliance or experimental effects, among others.

Taken together, these may be understood as the internal validity of a study. EEF-funded studies are assigned a 'padlock rating' using the EEF's classification of the security of the findings. This systematically summarises the

[2] Sample sizes are intrinsically linked to the level of 'uncertainty' in a study, but they are also related to its internal validity. While one can obtain an unbiased (yet imprecise) treatment impact estimate from a small study, a larger study is less likely to suffer internal validity problems such as randomization failure whereby the two groups are substantially different. The effectiveness of randomization relies on the law of large numbers and the central limit theorem, which are compromised in smaller samples.

The EEF’s classification system for single studies summarizes relevant aspects of the internal validity of findings and considers the professional judgement of the peer reviewers assigning them.

characteristics that define the internal validity and whether these make an estimate of impact from a given study more or less credible.

These dimensions cumulatively affect how much credence we give to a study. For instance, a study that succeeds to capture information on every participant would be more credible than one where only 60 per cent sat the relevant exam (all else being equal). Failing to include every learner in the follow up (called outcome attrition) can be a problem because those who did not sit the exam could have been different from those who did in a way that is related to the intervention.

The EEF’s classification system for single studies summarizes relevant aspects of the internal validity of findings and considers the professional judgement of the peer reviewers assigning them. These ratings should not be understood in a definite manner either, but as providing useful information to interpret findings. However, there are many other tools and resources used to gauge the robustness of a single study: from relatively simple approaches focusing on study design such as the Maryland Scientific Methods Scale **(Farrington et al., 2002)**, to others that consider multiple sources of bias and external validity problems depending on the type of design being considered **(Higgins et al., 2016; Sterne et al., 2017)**.

2.2.6

AN EXAMPLE

Now, using the key concepts described above, imagine you have three studies in the same domain, each with the goal of establishing the impact of an intervention:

- the evaluation of programme A was well-designed and well-conducted and found an effect size (ES) of 0.10; compatibility interval (CI): −0.10, 0.3; not statistically significant;

- the evaluation of programme B, also well-designed and well-conducted, found an ES of 0.10; CI: −0.01, 0.21; not statistically significant;

- the evaluation of programme Z1 was fraught with problems of internal validity that reduced its credibility; it found an ES of 0.20; CI: −0.20, 0.4; not statistically significant.

- the evaluation of programme Z2 was fraught with problems of internal validity that reduced its credibility; it found an ES of 0.20; CI: 0.10, 0.3; statistically significant.

The evaluation of programmes Z1 and Z2 suffered from important internal validity limitations[3] and thus the results are more likely to be called into question. One additional difficulty is that these problems with the design and implementation of a study are not always measurable and might be operating in different directions. This means that we might be overstating or underestimating the impact of an intervention, but the magnitude and direction in which this is happening is both difficult to ascertain and quantify. On these grounds, researchers are unlikely to recommend the use of Z as the evidence is not credible enough to claim that Z might be effective at improving outcomes. The findings could be understood as tentative at best and additional evidence of the effectiveness of Z would be necessary, by means of a better study.

The findings could be understood as tentative at best and additional evidence of the effectiveness of Z would be necessary, by means of a better study.

Studies for programmes A and B were well-conducted and methodologically robust[4] and had

[3] Using the EEF's classification system for single studies, these studies would be awarded a very low rating – probably one or two padlocks. For example, this could be an observational study designed to compare outcomes before and after without a control group. As it would not be possible to distinguish the effects of the intervention and the natural progress of pupils, we are unable to confidently conclude the intervention can improve pupil outcomes.

[4] Using the EEF's classification system for single studies, these studies would be awarded the maximum of five padlocks.

Quantitative studies in education and other applied domains provide a range of possible answers that need to be analysed, considering multiple sources of uncertainty.

the same estimate of impact: an ES of 0.10, which is equivalent to +2 months' additional progress.[5] However, as we stated above, studies do not give a single, unequivocal and definitive answer. The CI associated with both studies indicates that the data for programme B were also compatible with a range of effects from no impact to moderate impact, whereas the data for programme A were also compatible with a range of effects from a small negative impact to high impact.

Using statistical significance as the only criteria, researchers would have concluded that programme Z2 had statistically significant results (which is often understood as 'having an impact') while both programmes A and B had non-significant results (which is often understood as 'not having an impact').

This dichotomous interpretation of statistical significance is at the core of its problems and the source of contention around its use **(Wasserstein and Lazar, 2016)**. The advancement and use of scientific knowledge in the quantitative approach is not as simple as concluding that something works, and something does not. This example illustrates how the exclusive reliance on statistical significance could be very misleading as it obscures a much more nuanced picture: one where we are interested in understanding how well something works and which are the plausible scenarios that we can expect – that is, the uncertainty around the results.

Quantitative studies in education and other applied domains provide a range of possible answers that need to be analysed, considering multiple sources of uncertainty. Otherwise, decision-making is severely impeded. In the context of the example, no sound decision can be made exclusively on the basis of statistical significance because the uncertainty highlighted by coupling the effect size with confidence intervals

[5] The estimate of months of progress is based on EEF Guidance.

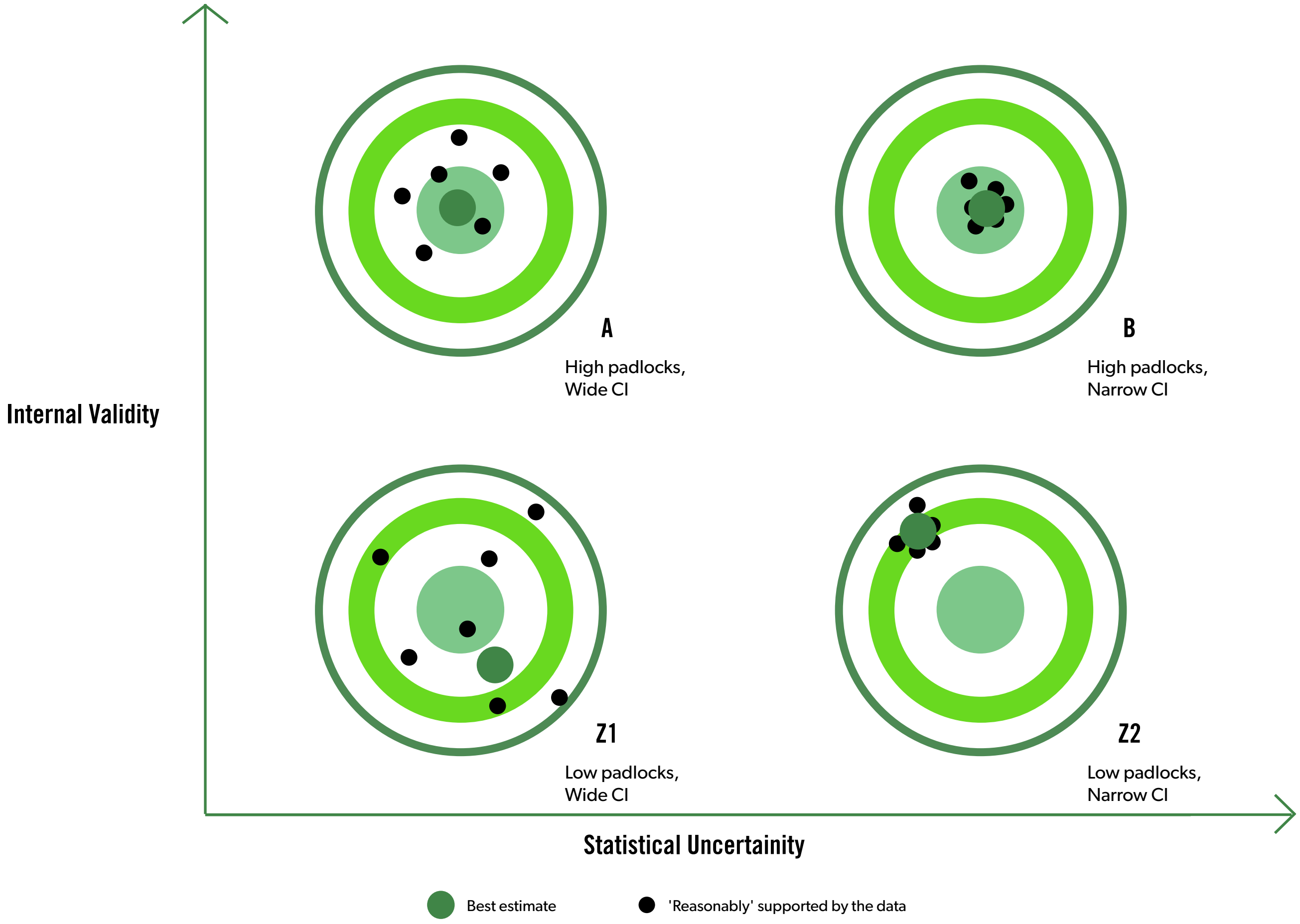


Figure 1. Schematic representation of internal validity and uncertainty

(an aspect commonly neglected) means that the findings in A are also compatible with a negative impact (−0.1) or a larger positive impact (0.3) while those of B are compatible with an educationally-very-small negative effect (−0.01) or a larger impact (0.3). Note that these are not the only values that are compatible with the data because confidence intervals should not be interpreted in a dichotomous way either, see **section 2.5**.

For a teacher or policy-maker deciding which of two similar programmes to invest in, both pieces of information are important and are represented conceptually in **Figure 1**.

Comparing Z with A or B would be like a vertical comparison in

Figure 1: between not-so-well-designed, tentative studies, and well-conducted, more credible studies. This comparison could be interpreted as the internal validity of the finding.

However, to discern between programmes A and B it is also relevant to consider other aspects. Even if both have the same estimate of impact (effect size), the findings of programme A are compatible with more variability (confidence intervals): from negative effects to larger positive impacts included in the intervals. In contrast, the findings of programme B show less variability being only compatible with a very small negative effect or a larger positive effect. This compares the uncertainty of the findings.

Making this distinction—between internal validity and uncertainty—accessible to decision-makers is fundamental: while the best estimate of A suggests a positive impact, the variability around it suggests more caution as the model of the data is compatible with the programme being harmful; however, the best estimate of B found the same positive impact, but at worst the model of the data was less compatible with the programme being harmful. Thus, with this information, a decision-maker may be more confident to implement B.

Decision-makers also need to consider a series of aspects when deciding which programme to implement; these include costs and resources, for example, which is why each EEF evaluation report provides an estimate of the required investment. For more information (**see EEF Cost Evaluation Guidance**). Other aspects include the programme's acceptability, its relevance to the problems faced by a particular school and the quality of programme implementation, among others. EEF evaluations strive to cover such topics as part of the Implementation and Process Evaluation component of all EEF-funded studies. For more information, (**see EEF IPE Guidance**).

Decision-makers also need to consider a series of aspects when deciding which programme to implement; these include costs and resources, for example, which is why each EEF evaluation report provides an estimate of the required investment. For more information see EEF Cost Evaluation Guidance.

2.3 Where does uncertainty come from?

There are multiple sources of uncertainty; but in the context of evaluations, two types are particularly relevant: sampling uncertainty and allocation uncertainty.

Even in a well-designed and well-conducted study with good internal validity, there are at least two steps in a RCT that introduce uncertainty.

1. When a group of schools or pupils is selected to take part in a study, random sampling leads to sampling uncertainty. This uncertainty is accepted because it is not practically feasible or economically viable to include

every school in every single study. Even if a random sample from the population is selected, such schools or pupils might be different from the population at large for reasons we might not be able to identify. Note that in most cases, samples of participants taking part in a RCT are not drawn at random from the population.

2. When these schools or pupils are subsequently randomly allocated to the intervention or control group, random assignment leads to allocation uncertainty. Even if these are randomly assigned, there might be differences between the two groups for reasons we might not be able to identify.

When these schools or pupils are subsequently randomly allocated to the intervention or control group, random assignment leads to allocation uncertainty.

These two processes thus introduce sampling uncertainty and allocation uncertainty, respectively.

Even if the same experiment is repeated a large number of times, these sources of uncertainty imply that the observed differences between groups could differ under each of these identical hypothetical experiments. These types of uncertainty are closely linked with the heterogeneity between units in the population and the sample.

When individuals in the population are very different from each other, it is more likely that a random sample would end up with a group with very different characteristics for which the estimate of impact could be different from the 'true' population effect (1). Likewise, even within a given sample, the random allocation might lead to a treatment group with very different characteristics for which the estimate of impact could also be different from the impact estimate that would be obtained with a different random configuration of the treatment and control groups (2).

This means that it is always possible that the true effect size[6] observed in an RCT will differ from the true average effect size in the sample because, even for two identical experiments, the observed effect size is likely to differ a bit, and will occasionally differ a lot, as a result of this statistical uncertainty.

Likewise, the observed effect size may also be different from that on the population. In addition to the problems related to inferences in a sample, to make broader claims around the external validity of the findings to a population it is necessary to consider many other aspects beyond statistical uncertainty, which are more likely to influence whether the results observed in a sample can be expected to be replicated for the population **(Deaton and Cartwright, 2018)**.

However, these are not the only sources of statistical uncertainty. For instance, to focus on one of the most common, the accuracy and reliability of an outcome test may also introduce measurement uncertainty from the selected instruments. This relates to the margin of doubt that exists for the result of any measurement that could be due both to the instrument being used (e.g. a test, a timer) and how this translates the relevant behaviour into a quantitative value (e.g. a score). This can also be affected by the construct being measured (e.g. algebra, self-efficacy). Hence every measurement differs from the 'true' value that it is trying to capture. This difference is the error, while measurement uncertainty is the quantification of those expected errors and is often expressed as a confidence interval around a measurement. The measurement uncertainty introduced by using a specific outcome measure could be considered an internal validity problem but it also adds to the variability of the results observed.

This means that it is not possible to isolate the multiple sources of uncertainty from some aspects of internal validity.

[6] It is not possible to know the 'true average effect size' as that would require pre-test and post-test outcomes for each member of the sample/population both with and without the intervention, which is not possible.

2.4 The problems with statistical significance

To assess uncertainty, many researchers consider a hypothetical situation where:

1. a (random) sample is drawn from the population of interest[7] (which would be related to sampling uncertainty);

[7] RCTs are hardly ever a random sample from the population. EEF-funded studies are not random samples. This means that the interpretation of the *p*-values should not be considered as making claims about the external validity of the study (inferences on the impact on the population) but only as relating to the sample at hand (inferences on the internal validity of the study on the sample).

One of the reasons for misinterpretation is that p-values give the right answer to the wrong question.

2. the same experiment is conducted a large number of times on samples drawn from the same population (which would be related to allocation uncertainty, and other sources of uncertainty related to the internal validity of the study); and

3. the intervention has no true impact on the population (i.e., the real impact of the intervention is zero).

Then, researchers estimate how likely it would be, in this hypothetical situation, to observe a difference at least as big as the difference they observed due to the statistical uncertainty.

This probability to observe a difference at least as big as the difference they observed is called the *p*-value.

This statistic has been strongly criticized because frequent misuse and misinterpretation lead to distortions in scientific enquiry **(Wasserstein and Lazar, 2016; Amrhein, Greenland and McShane, 2019; Wasserstein et al., 2019)**. One of the reasons for misinterpretation is that *p*-values give the right answer to the wrong question. In practice, the question we want to answer is, 'does this intervention work?' Instead, *p*-values explain, 'how rare would these results be in a world where the intervention had no effect?' (i.e. the hypothetical situation, which also requires fulfilling the other assumptions mentioned above)'. For example, imagine you want to identify whether a programme improves pupil outcomes and you found a difference equivalent to three months of progress. The question we want to answer is: given that we observed a difference of three months of progress, how likely is it that this programme had no effect? This is not what a *p*-value tells us. The *p*-value shows the probability that you would observe a difference of three months or more given that the intervention had no impact (the hypothetical situation, which also includes the other relevant assumptions described above).

P-values give neither an indication of the likelihood that the intervention had an effect nor give the probability that the observed result was produced by random chance alone **(Greenland et al., 2016; Wasserstein and Lazar, 2016; Amrhein et al., 2019; Wasserstein, Schirm and Lazar, 2019)**. *P*-values give a very indirect answer to the question we are truly interested in. The smaller the *p*-value, the more unusual the results if all the assumptions under the hypothetical situation are true. However, a very small *p*-value does not tell us which of the assumptions might be incorrect even if we are only truly interested in the question of whether this intervention worked – closely related with the third assumption above **(Greenland et al., 2016)**.

... the most salient problem with p-values (and also similar statistics such as confidence intervals, discussed below) is the convention to treat them in a dichotomous way around a 0.05 threshold - a 'bright-line'

However, the most salient problem with *p*-values (and also similar statistics such as confidence intervals, discussed below) is the convention to treat them in a dichotomous way around a 0.05 threshold – a 'bright-line' where on one side an impact is inferred to exist, while on the other, the possibility of an impact is entirely disregarded as inconsistent with the data.

This simplification is a caricature of the necessary complexity to make inferences to advance scientific knowledge and violates the spirit of how *p*-values are supposed to be interpreted. Originally, the 0.05 threshold was chosen as a way to limit the risk of false positives. It means that if you were to repeat the experiment 100 times under the hypothetical situation (i.e., the programme has no effect), in five of them, you would see results as extreme or more extreme than yours. The original proponent of the *p*-value, Ronald Fisher, argued that a statistically significant finding was worthy of further investigation. Alas, in a gross misrepresentation of that spirit, this threshold became the value to consider a finding 'true', which is not true **(Wasserstein and Lazar, 2016)**.

Rather than a 'bright-line' where effectiveness can be decided, *p*-values provide a continuum of how compatible the data are with the hypothetical situation. Values at either side of the threshold should not be treated as definitive answers but as different tonalities

A finding might be of educational/practical significance (represented as a large effect size) even if it is not deemed 'statistically significant' by reaching the arbitrary 0.05 cut-off point.

of grey – data that are more or less compatible with the estimate of impact. Even if actionable recommendations may require an affirmative answer, making inferences on the basis of an arbitrary threshold is incorrect and has distorted decision-making **(Wasserstein, Schirm and Lazar, 2019).**

This dichotomy at each side of the threshold also conflates practical and statistical relevance. A finding might be of educational/practical significance (represented as a large effect size) even if it is not deemed 'statistically significant' by reaching the arbitrary 0.05 cut-off point. This problem is particularly heinous because when a study is large, even small violations of the assumptions can lead to a 'statistically significant' result that affects how decisions are made.[8] Contrariwise, even an educationally relevant difference could fail to be 'statistically significant' if the sample is not large enough. Sometimes a statistically significant result simply means that a very large sample was used.[9]

The most common alternative is to report confidence intervals or compatibility intervals (CI). As is the case with *p*-values, confidence intervals are also prone to misinterpretation **(Greenland et al., 2016; Morey et al., 2016).** These estimate that if the same experiment were conducted a large number of times and interval estimates are made on each

[8] For example, (Sullivan and Feinn (2012) mention an example of a study for aspirin. In the study, more than 22,000 subjects used aspirin over five years and the authors identified a statistically significant reduction in heart disease even if the reduction in risk was very small – and clinically negligible – for most patients. However, aspirin was recommended for general prevention for years. More recent studies confirm aspirin should be taken only for those who have suffered heart disease or a stroke and medical guidelines have been adapted accordingly.

[9] This also highlights the importance of relying on bodies of evidence, instead of single studies. By combining the information from multiple studies, systematic reviews (and statistical methods such as meta-analyses that combine different findings into a single metric) help to use information across all observations, which can help mitigate some of the problems related to single studies relying on statistical significance. However, it is important that the interpretation of these analyses is not subject to the same dichotomous interpretation of statistical significance.

P-values and CI are calculated based on similar hypothetical situations, and suffer from similar problems; including the erroneous dichotmous interpretation

occasion, the resulting intervals would bracket the true population parameter in approximately 95 per cent of the cases if the hypothetical situation is true.

P-values and CI are calculated based on similar hypothetical situations, and suffer from similar problems; including the erroneous dichotmous interpretation. CI are often interpreted as 'not crossing zero' to suggest that a result is 'statistically significant' and thus, 'true'. This is untrue. Symmetrically to *p*-values, a CI can only help to conclude how compatible the results are with a given statistical model. Just because a value lies outside of the specific CI, it does not mean that this value can be refuted or excluded from the data – just that it is less compatible with the assumptions used.

However, as argued above, using CI is seen as superior to *p*-values because presenting a range of values that is consistent with a given model of the data is more likely to be interpreted with caution rather than a single value that is often understood as evidence that an effect 'exists' or not **(Greenland et al., 2016)**.

In short, the issue around the interpretation and use of *p*-values, CI and statistical significance has less to do with the assumptions upon which they are constructed than with the obsession with a clear decision rule (i.e. a threshold) to conclude whether something is 'true' or not. This shows a naïve interpretation of the statistical assumptions underpinning these concepts but, more importantly, it steers decision-makers and practitioners away from key pieces of information needed to formulate new policies and introduce changes.

2.5 The way forward: bringing together internal validity and uncertainty to make the best use of evidence in educational decision-making

Effect sizes provide a better indication of the magnitude of impact and thus should be reported for all estimates.

Internal validity and uncertainty should be considered in tandem when making a decision about a programme, as illustrated in the discussion above. Internal validity measures the suitability of the design of the study to produce estimates close to the true estimate of impact, that is, how close one is to the bull's eye or the bias of the estimate. Uncertainty measures how likely it is that the same experiment, repeated under the same conditions, would find a similar effect, that is, how close are different estimates of impact to each other or to the spread of the estimate. This was represented conceptually in **Figure 1**.

Ideally, a study should be well-designed and well-implemented (good internal validity) and likely to find a similar effect if replicated under the same conditions (low uncertainty). However, studies are hardly ever definitive and both aspects need to be factored into any interpretation of the results.

To address the criticisms above we propose that findings should be discussed in terms of effect sizes, with a thorough description of their internal validity using well-regarded tools; and importantly, emphasizing the role that uncertainty plays in decision-making and moving away from a dichotomous interpretation of statistical significance. Commissioners and researchers may also consider translating these measures into other, more readily understood, measures such as months of progress.

To aid the effective communication of findings for educational interventions, we propose the following principles, which distill work by Wasserstein and Lazar **(2016)**, Wasserstein, Schirm and Lazar **(2019)**, and Amrhein, Greenland and McShane **(2019)**.

1. Use effect sizes to focus on the practical/scientific significance of a finding rather than relying on whether the finding was statistically significant.

The arbitrary 0.05 cut-off conflates practical and statistical relevance. However, statistical significance does not explain whether a finding is practically/scientifically/educationally interesting. Effect sizes provide a better indication of the magnitude of impact and thus should be reported for all estimates. These may be considered alongside other transformations to aid interpretation such as measures of months of progress that might be more accessible for decision-makers.

2. Include assessments of internal validity.

Results should be accompanied by a thorough description of the different elements that affect the internal validity of the study. This could be reported either using standardized tools such as Robins I or Risk of Bias Assessments, or bespoke tools such as EEF's Padlocks Rating. Threats to internal validity should always be reported transparently, even if the magnitude and direction of biases are difficult to quantify.

Results should be accompanied by a thorough description of the different elements that affect the internal validity of the study. This could be reported either using standardized tools such as Robins I or Risk of Bias Assessments, or bespoke tools such as EEF's Padlocks Rating

3. Accept uncertainty in findings and always present a measure of this uncertainty.

Statistical modelling should not be interpreted as providing unique and definitive answers, or what Gelman **(2016)** calls 'a sort of alchemy that transmutes randomness into certainty'. Instead, it is paramount to understand that, in real-world situations, statistical modelling only attempts to identify 'signals' in noisy data with considerable variability. Therefore, we should acknowledge that statistical models only provide incomplete and uncertain – yet potentially useful – answers to scientific questions. Abandoning a dichotomous interpretation of *p*-values and other statistics, including 'CI', advances in this direction moving us away from the detrimental simplification of findings as 'true' or not. Thus, researchers must present a measure of the uncertainty around all effect sizes, recognizing that uncertainty is an integral part of statistical modelling and scientific enquiry.

... in real-world situations, statistical modelling only attempts to identify 'signals' in noisy data with considerable variability.

4. Use precise language and clearly consider assumptions behind the statistics used to represent uncertainty.

P-values do not measure the probability that 'the studied hypothesis is true' nor the probability that the 'data were produced by random chance alone' **(Wasserstein and Lazar, 2016)**. Similar misinterpretations are common when describing confidence intervals **(Greenland et al., 2016; Morey et al, 2016)**. To a large extent, the problem with *p*-values is that they offer an answer to a question we are not necessarily seeking to answer – that of the hypothetical scenario. However, ignoring the assumptions upon which *p*-values are calculated goes a long way toward explaining why they have become contentious and potentially misleading. Thus, researchers must be accurate in the interpretation of *p*-values (or any other statistic used), what they are and what they are not, carefully considering the assumptions upon which these are constructed.

5. Report continuous *p*-values (or other measures of statistical uncertainty), interpreting them as varying degrees of statistical uncertainty and avoiding dichotomization of decisions around the arbitrary cut-off of p = 0.05.

P-values are the probability, under a specified statistical model (the hypothetical scenario), that the mean difference between two groups would be equal or more extreme than the observed value in the study **(Wasserstein and Lazar, 2016)**. As a continuous probability, *p*-values are a measure of the degree of compatibility of the data with the hypothetical model imposed on that data. Claiming a finding as 'statistically significant' suggests a dichotomous interpretation that contravenes **Recommendation 1**. Therefore, abandon the dichotomous interpretation of *p*-values, recognizing that different *p*-values suggest different levels of strength of the evidence and thus should be reported as a value and interpreted as a continuum. Findings should be interpreted neutrally, irrespective of whether results are 'positive' (positive effect size, not statistically significant) or not. Other statements that suggest a dichotomous interpretation around the 0.05 should also be shunned. For example, phrases such as 'no evidence of impact', 'there is no difference', and 'nearly statistically significant' should be discontinued entirely.

To report statistical uncertainty around the point estimate, discuss the educational/ scientific relevance of the point estimate and also the extremes of the compatibility intervals.

6. Discuss the practical relevance of 'CI'.

Avoid referring to 'confidence' intervals as the word confidence suggests ungranted certainty **(Amrhein, Greenland and McShane, 2019; Greenland, 2019; Wasserstein et al., 2019)**. To report statistical uncertainty around the point estimate, discuss the educational/ scientific relevance of the point estimate and also the extremes of the compatibility intervals. Note that these compatibility intervals reflect other values, under the hypothetical statistical model used, that are also compatible with the data. Even if intervals are estimated based on a predetermined threshold – conventionally 95 per cent aligned with a *p* of 0.05 – they should also not be interpreted in a dichotomous way as outlined in **Recommendation 5**: values closer to the point estimate (the best estimate of impact) are better supported by the data, while those farther away are less compatible with it. Values outside these intervals are less compatible with the data, not inconsistent with it.

7. Consider accompanying *p*-values and 'CI' with other statistics.

Explore other statistics that could help interpretation, rather than interpreting them in a dichotomous way regardless of which statistic is chosen. Researchers may, for instance, consider permuted *p*-values that do not rely on the assumption of random sampling and thus do not intend to make generalizations beyond the sample, or other statistics like Bayesian CI, which rely on other assumptions. The

American Statistical Association's (ASA) special issue, *Statistical inference in the 21st century: A world beyond* $p<0.05$, offers some suggestions. Researchers may also want to present alternatives to test the sensitivity of the statistical uncertainty captured by different models.

8. Discuss practical and scientific significance considering all relevant information.

Interpret the findings considering internal validity, statistical uncertainty, the strength of the existing evidence, the plausibility of the causal mechanism, the evidence of the quality

...we propose that findings should be discussed in terms of effect sizes, with a statement about the internal validity of the finding and representing the statistical uncertainty of the finding as a continuous p-value, 'CI', and/or alternative statistics.

of the implementation, and considerations of the context. Also consider the process through which the statistics were obtained: For example, if the design and analysis were pre-registered, the effect size is more likely to approximate the true effect of interest than if the effect was observed only after exploring a range of subgroup, outcomes, and/or treatment variations, and selected on the basis of its magnitude or associated *p*-value. If a design and analysis are not pre-registered, or if the analytic process is not transparently described, a promising effect should be appropriately discounted.

Furthermore, researchers should be thoughtful in describing how the finding shifts the evidence-base and existing priors. This is important because these statistics should be understood in the context of the processes that generated them, and thus, bringing additional information is crucial to decision-making.

In sum, we propose that findings should be discussed in terms of effect sizes, with a statement about the internal validity of the finding and representing the statistical uncertainty of the finding as a continuous *p*-value, 'CI', and/or alternative statistics.

Advancing scientific knowledge in education is a complex endeavour. But it is also a laudable one - it has the potential to improve people's lives by fostering learners' strengths and, if needed, providing scaffolding to move past difficulties. We hope that these principles will help researchers move closer to that goal by providing decision-makers with the necessary information to make the right decisions about educational interventions grounded in evidence of what works, and eventually, what works best **(WG4-ch1)**.

REFERENCES

Amrhein, V., Greenland, S. and McShane, B. (2019) 'Retire statistical significance', Nature, 567, pp. 305–307.

Baird, M. and Pane, J. (2019) 'Translating standardized effects of education programs into more interpretable metrics', Educational Researcher, 48(4), pp. 217–228.

Bloom, H.S., Hill, C.J., Black, A.B., and Lipsey, M.W. (2008) 'Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions', Journal of Research on Educational Effectiveness, 1(4), pp. 289-328.

Coe, R. (2002) 'It's the effect size, stupid. What effect size is and why it is important', British Educational Research Association Annual Conference, Exeter.

Deaton, A. and Cartwright, N. (2018) 'Understanding and misunderstanding randomized controlled trials', Social Science & Medicine, 210, pp. 2–21.

Evans, D. and Yuan, F. (2019) Equivalent years of schooling. a metric to communicate learning gains in concrete terms. Washington DC: World Bank Policy Research Working Paper 8752.

Farrington, D.P., Gottfredson, D.C., Sherman, L.W. and Welsh, B.C. (2002) The Maryland Scientific Methods Scale. Milton Park: Routledge.

Gelman, A. (2016) 'The problems with p-values are not just with p-values', The American Statistician, 70, pp. 1–2.

Gorard, S. (2016) 'Damaging real lives through obstinacy: re-emphasising why significance testing is wrong', Sociological Research Online, 21(1), pp. 102–115.

Greenland, S. (2019) 'Valid p-values behave exactly as they should: some misleading criticisms of p-values and their resolution with s-values', The American Statistician, 73(Sup1), pp. 106–114.

Greenland, S., Senn, S.J., Rothman, K.J., Poole, C., Goodman, S.N. and Altman, D.G. (2016) 'Statistical tests, P values, confidence intervals and power: a guide to misinterpretations', European Journal of Epidemiology, 31, pp. 337–350.

Higgins, J., Savovic, J., Page, M.J. and Sterne, J.A. (2016) Revised Cochane Risk of Bias Tool for Randomized Trials (RoB 2.0).

Higgins, S. (2021) Improving learning: meta-analysis of intervention research in education. Cambridge: Cambridge University Press.

Hubbard, R. (2016) Corrupt research: the case for reconceptualizing empirical management and social science. London: SAGE Publications.

Lipsey, M., Puzio, K., Yun, C., Hebert, M., Steinka-Fry, K., Cole, M., ... and Busick, M. (2012) Translating the statistical representation of the effects of education interventions into more readily interpretable forms. Washington DC: National Center for Special Education Research.

Major, L.E. and Higgins, S. (2019) What works? Research and evidence for successful teaching. London: Bloomsbury.

McShane, B.B., Gal, D., Gelman, A., Robert, C. and Tackett, J.L. (2019) 'Abandon statistical significance', The American Statistician, 73(1), pp. 235–245.

Morey, R., Hoekstra, R., Rouder, J., Lee, M. and Wagenmakers, E. (2016) 'The fallacy of placing confidence in confidence intervals', Psychonomic Bulletin & Review, 23(1), pp. 103–123.

Shrout, P.E. (1997) 'Should significance tests be banned? Introduction to a special section exploring the pros and cons', Psychological Science, 8(1), pp. 1–2.

Sterne, J., Hernán, M.A., Reeves, B.C., Savović, J., Berkman, N.D., ... and Higgins, J.P.T. (2017) 'ROBINS-I: a tool for assessing risk of bias in non-randomised studies if interventions', British Medical Journal, 335.

Sullivan, G.M. and Feinn, R. (2012) 'Using effect size-or why the P value is not enough', Journal of Graduate Medical Education, 4(3), pp. 279–282.

Trafimov, D. and Marks, M. (2015) 'Editorial', Basic and Applied Social Psychology, 37(1), pp. 1–2.

Wasserstein, R.L. and Lazar, N.A. (2016) 'The ASA statement on p-values: context, process and purpose', The American Statistician, 70(2), pp. 129–133.

Wasserstein, R.L., Schirm, A.L. and Lazar, N.A. (2019) ‘Moving to a world beyond “p<0.05”’, The American Statistician, 73(Sup1), pp. 1–19.

Ziliak, S. and McCloskey, D.N. (2008) The cult of statistical significance: how the standard error is costing us jobs, justice and lives. Ann Arbor: University of Michigan Press.

WORKING
GROUP 04

CHAPTER

3

The extent to which education interventions have been studied and the range of effects typically observed

DOI: https://doi.org/10.56383/HFEY6916

This chapter should be cited as:

Kay, J., Higgins, S.E., Villanueva Aguilera, A.B., Dobson, E.S., Gascoine, L., Katsipataki, M., Rajab, T., Zaman, M., Ellis-Thompson, A., Madgwick, H., Patel, R. and Blausten, H. (2022) 'The extent to which education interventions have been studied and the range of effects typically observed' in Mercier, J. (ed.) Data and Evidence in, Duraiappah, A.K., van Atteveldt, N.M., Borst, G., Bugden, S., Ergas, O., Gilead, T., Gupta, L., Mercier J., Pugh, K., Singh, N.C. and Vickers, E.A. (eds.) Reimagining Education: The International Science and Evidence Based Assessment. New Delhi: UNESCO MGIEP.

Coordinating Lead Authors

Jonathan Kay

Steve E. Higgins

Alaidde Berenice Villanueva Aguilera

Emma Sian Dobson

Louise Gascoine

Maria Katsipataki

Taha Rajab

Mohammad Zaman

Amy Ellis-Thompson

Harry Madgwick

Rupal Patel

Hannah Blausten

WORKING GROUP 04

3

CHAPTER

3.1 Introduction

This chapter has been compiled from the findings of the Education Endowment Foundation's (EEF) education database project, a joint study conducted with Durham University.

The EEF's education database is comprised of thousands of education research studies from across the globe, all focused on measuring the impact of education interventions on students' outcomes.

The EEF's education database is comprised of thousands of education research studies from across the globe, all focused on measuring the impact of education interventions on students' outcomes. The studies in the database have been coded to enable analysis and searching across a range of factors, including country, pupil age and type of intervention.

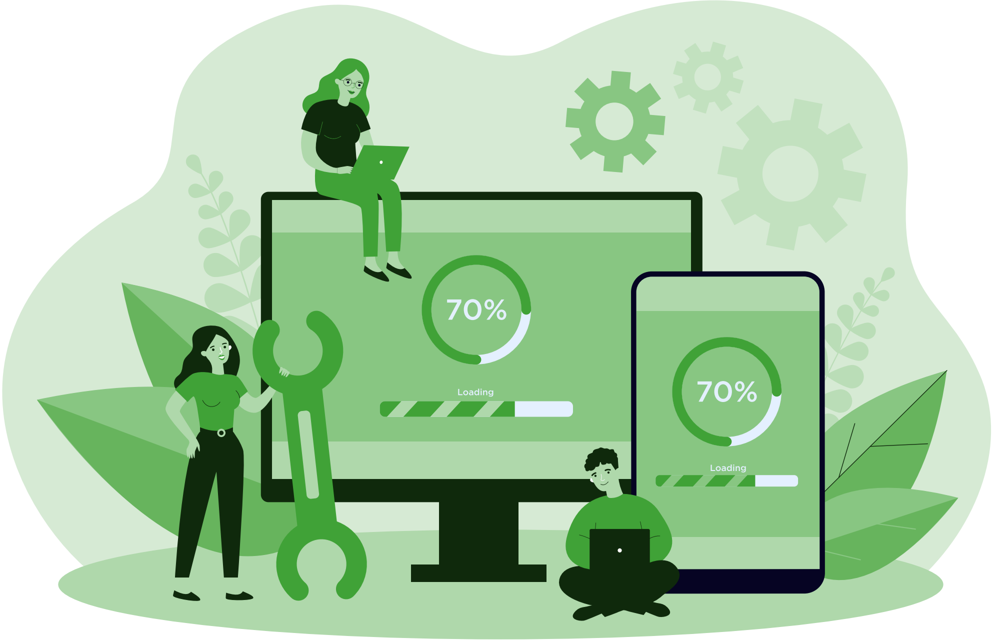
Rather than simply focusing on the impact of interventions, the database also records information about the delivery of interventions (such as the frequency and intensity of the intervention) and detailed quantitative impact data, such as variations in effects based on subject or delivery mechanism (such as whether an intervention is delivered by a qualified teacher or a classroom assistant). Impact is translated from standardized effect sizes to 'months of learning' for ease of communication and to aid discussion around the impact of interventions. Months of learning, communicated as a headline figure for each approach, however, can hide important variation caused by duration of intervention, group size and the test measures used. Building the database containing all of this data allows researchers to examine which factors are driving the impact behind the overall average to find the signal amongst the noise. It is this detailed data which makes this education database unique. It will significantly reduce the time and effort needed to review the impact of different types of interventions, and to analyse the factors that increase or reduce effectiveness.

3.2 The database and the EEF Teaching and Learning Toolkit

The database has been designed to underpin updated versions of the EEF's Teaching and Learning Toolkit and Early Years Toolkit.

The EEF Toolkits are accessible summaries of education research for teachers and decision-makers. With over forty approaches for improving teaching and learning, each is summarized in terms of its average impact on attainment, its cost and the strength of the evidence supporting it.

The database and the Toolkit are living reviews of the evidence. They are updated whenever new studies become accessible and coded. This document – the International Science and Evidence based Education (ISEE) Assessment – cannot replicate the living nature of the reviews, and so readers should consult the live versions which are available at https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/, and via any of the EEF's global partner organizations, listed at https://educationendowmentfoundation.org.uk/about/international-work/eefs-international-partnerships/.



3.3 The future of the database

Much like the Cochrane Library and other living reviews of evidence, the database is a long-term project that will continue to grow and develop over time. Among the goals for the database over the next five to ten years are:

- inclusion of non-English - language studies by EEF partner organizations around the world, allowing the Toolkit to be further tailored to different contexts;

- national and international partnerships and fellowships, enabling external researchers to use the database for research and analysis thus contributing to the global education evidence base;

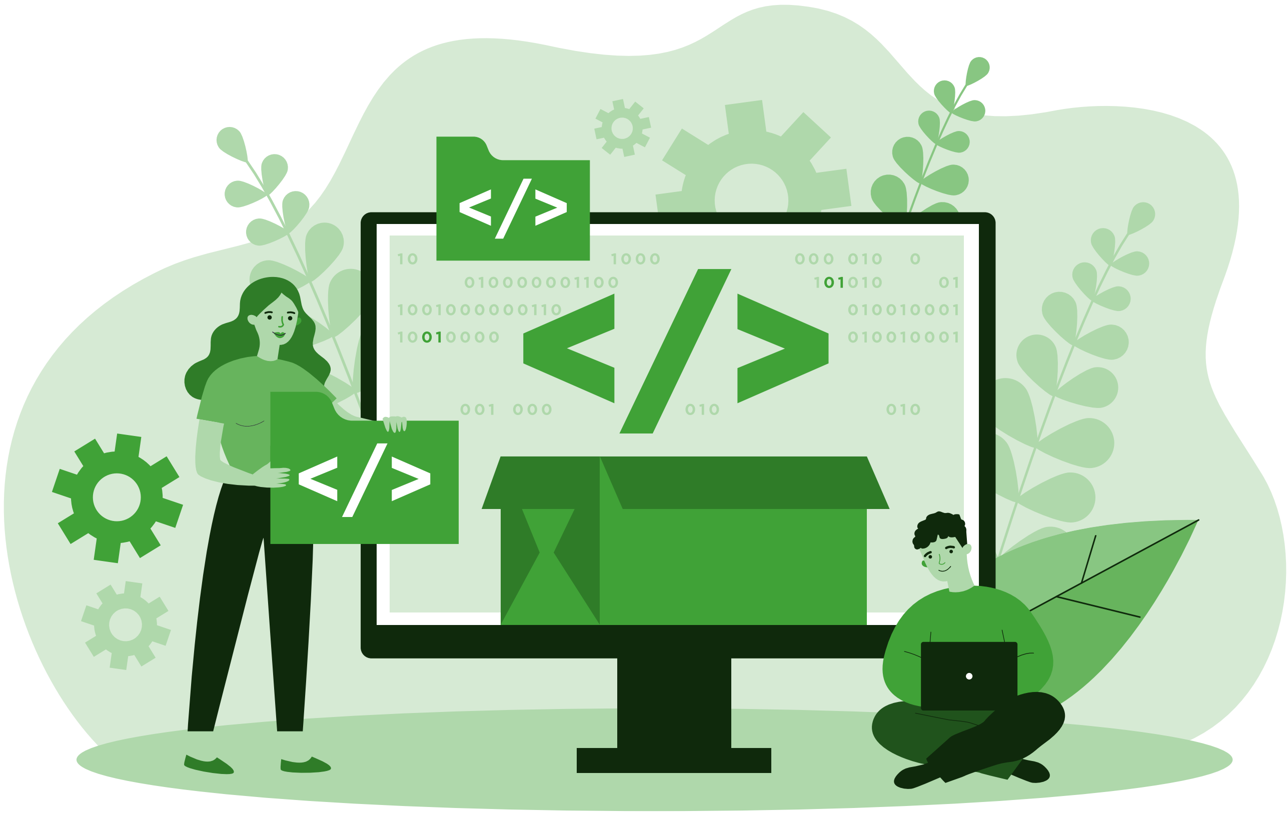
- use of machine learning and artificial intelligence to automatically search for, identify and extract data from new studies, reducing the time and cost of reviewing education research;

- automatic live updating of the EEF Toolkit from the database, allowing for the 'living' systematic review to be updated quickly with the most recently published studies.

WORKING GROUP 04

3

CHAPTER



3.4 Methods

The first set of studies for the database has been identified from the current version of the Sutton Trust–EEF Teaching and

Learning Toolkit. Nearly all of the strands are based on meta-analyses and systematic reviews which have been identified through a systematic updating process **(EEF, 2018[1])** since the initial version of the Toolkit was published by the Sutton Trust in 2011 **(Higgins, Kokotsaki and Coe, 2011)**.

These meta-analyses have been systematically 'unzipped' so that the included studies which contribute to the overall pooled effect are identified and screened (a two-stage process of title and abstract and then full text screening) for inclusion in the database **(Higgins et al., 2022)**.

These meta-analyses have been systematically 'unzipped' so that the included studies which contribute to the overall pooled effect are identified and screened … for inclusion in the database.

3.4 .1

INCLUSION CRITERIA FOR THE EEF EVIDENCE DATABASE

The inclusion criteria aim to identify relevant educational evidence for schools and policy-makers interested in school-based education, consistent with the mission of the EEF, which is dedicated to breaking the link between family income and educational achievement. Specifically, the EEF aims to:

- raise the attainment of three- to eighteen-year-olds, particularly those facing disadvantage;

- develop their essential life skills; and

- prepare young people for the world of work and further study.

PICOS and SPIDER analyses **(Methley et al., 2014)** were used to define the scope of the database:

[1] https://educationendowmentfoundation.org.uk/public/files/Toolkit/Toolkit_Manual_2018.pdf

3.4 .1 .1

PICOS SPIDER DATABASE SCOPE: EXPLANATION AND EXAMPLES

POPULATION	Sample	Early years and school-age learners from three to eighteen learning in their first language.	The focus is on educational settings. This can include out-of-school interventions, such as summer schools or after-school clubs, where the aim is to improve academic learning; or where the impact of the activity is evaluated in terms of its educational benefit (e.g. Scouts or Guides or an Outward Bound course). Higher education settings (degree-level) are excluded. Studies of second-language learners (L2) studying subjects other than an additional language are excluded.[2]
INTERVENTION	Phenomenon of interest	Educational intervention or approaches, including named or clearly defined programmes and recognizable approaches that are classifiable according to the Toolkit strand definitions (e.g. peer tutoring or small group teaching). The intervention or approach is undertaken in a normal educational setting or environment, such as a nursery or school or a typical setting (e.g. an outdoor field centre or museum).	The focus is on the ecological validity of the research. The intervention or approach should last for at least one week or a minimum of five hours of activity time in terms of learners' experience. This excludes laboratory studies or atypical environments used to test theoretical rather than educational questions.
COMPARISON	Design	A valid comparison between those receiving the educational intervention or approach and those not receiving it.[3]	The aim is to provide an estimate of impact based on a counterfactual comparison. Studies where this is no control for maturation (e.g. single subject studies or single cohort designs with pre- and post-tests only for the intervention or approach) would be excluded.

[2] A study of Spanish-speaking students learning mathematics in English would be excluded. A study of Spanish-speaking students learning French in a Spanish medium school would be included.

[3] Specific design features are identified through coding so that these can be investigated as moderators.

OUTCOME(S)	Evaluation	Assessment of educational or cognitive achievement which reports quantitative results from testing of attainment or learning outcomes such as via standardized tests or other appropriate curriculum assessments or school examinations or appropriate cognitive measures (**Higgins et al., 2022**).	The focus is on educational achievement in schools or other educational settings. The availability of non-cognitive outcomes is recorded, but these are not extracted because of the challenge of commensurability.
STUDY DESIGN	Research type	Designs where a quantitative estimate of the impact of the intervention or approach on the educational attainment of the sample can be calculated or estimated in the form of an effect size (standardized mean difference) based on a counterfactual comparison.	A standardized mean difference of the impact of the intervention or approach must be reported or must be calculable,[4] such as from randomized controlled trials, quasi-experimental studies, regression discontinuity designs and natural experiments with a valid comparison. In addition, the standard error of this effect must be reported, calculable or estimable.

This analysis was used to create specific inclusion and exclusion criteria.[5]

[4] This includes other measures of impact such as correlational and categorical effect sizes where these result from a counterfactual comparison and where they can meaningfully be converted to a standardized mean difference (**Borenstein et al., 2009**).

[5] Sample size is not included in these criteria. This is because we intend to undertake an analysis of the relationship between sample size and effect size based on the existing evidence of an inverse relationship in education (**e.g. Slavin and Smith, 2009**) and other fields (**e.g. Button et al., 2013; Kühberger, Fritz and Scherndl, 2014**). This has implications for meta-analysis as methods for publication bias and the use of a random effects model assume sample size and effect size are independent.

3.4.2

INCLUSION CRITERIA EXCLUDED

The majority of the sample (greater than 50 per cent) on which the analysis is based are learners or pupils aged between three and eighteen (further education or junior college students are to be included where their study is for school-level qualifications).	The majority of the sample are: those are post-secondary education; in higher education; adults; infants under three; other students over 18.
The intervention or approach evaluates the impact of an educational intervention or approach, including named or clearly defined programmes and recognizable approaches classifiable according to the Toolkit strand definitions (see the statistical analysis plan here "https://educationendowmentfoundation.org.uk/public/files/Toolkit/EEF_Evidence_Database_Protocol_and_Analysis_Plan_June2019.pdf").	The intervention or approach is not classifiable with regard to the current Toolkit strand definitions (see the statistical analysis plan here "https://educationendowmentfoundation.org.uk/public/files/Toolkit/EEF_Evidence_Database_Protocol_and_Analysis_Plan_June2019.pdf").
The intervention or approach is undertaken in a normal educational setting or environment, such as a nursery or school or a typical setting (e.g. an outdoor field centre or museum).	Laboratory studies; specially created environments (both physical and virtual) designed for theoretical research questions, rather than educational benefit[6].

[6] For example, by using the conversions available in programs like Comprehensive Meta-Analysis or David B. Wilson's online conversion tool: https://campbellcollaboration.org/escalc/html/EffectSizeCalculator-Home.php.

3.4 .3

SEARCH STRATEGY FOR IDENTIFICATION OF RELEVANT SINGLE STUDIES

Where there were no existing meta-analyses or systematic reviews with quantitative data in the existing Toolkit strands, a new systematic search was undertaken for primary studies to update the existing single studies identified for the Toolkit. The following sources (gateways and databases) were used:

- First search
  - Article First
  - ECO
  - Papers First
  - World Cat Dissertations
- EBSCO
  - BEI
  - Education Abstracts
  - Education Administration Abstracts
  - ERIC
  - PsycArticles
  - PsycINFO
- Taylor and Francis
  - Educational Research Abstracts Online
- ProQuest
  - ProQuest Dissertations & Theses (Global)
- Elsevier
  - Science Direct
- Thomson Reuters
  - Web of Science

In addition, informal searching for 'grey' literature (reports and unpublished studies) was undertaken using Google, Google Scholar and Microsoft Academic.

We did not use citation searching, 'pearl growing' **(Schlosser et al., 2006)** or expert nomination, although we did use these techniques to ensure the adequacy of search terms **(Papaioannou et al.,**

2010). Our rationale for this is that the use of such approaches on their own, without subsequently adapting the search criteria, is likely to increase the risk of publication bias **(Higgins, 2018)**. Where we identified relevant studies from non-systematic approaches we aimed to refine our search criteria and to run additional searches to find other similar studies retrieved with the amended search strings

3.4 .4 RELEVANCE FOR TEACHERS, SCHOOL LEADERS AND POLICY-MAKERS

The database allows users to review the impact of different approaches to improving outcomes for children and young people by understanding not just the average impact of an intervention, but how that impact varies depending on subject, age of pupils and country. It will give teachers, school leaders and policy-makers a much better idea of whether an approach that has worked elsewhere can also work in their particular context.

3.4 .5 LIMITATIONS

The evidence summaries that follow describe the average impact and some of the reasons for variation that have been identified. They cannot predict the impact of an approach in any classroom.

In particular, a search of the global evidence reveals gaps in research on pedagogical strategies in particular contexts. While randomized control trials (RCTs) have proliferated in the context of development, few of these studies look at pedagogical approaches. Rather, they frequently focus on structural approaches or efforts to increase access to education. While these efforts are critical to

In order to improve accessibility to teachers – the meta-analytic results, average cost and evidence security are communicated using months progress, a five-point cost scale and padlocks. The details of these headline estimates can be found below.

Summary of headline figure data. More information can be found here.

Evidence strength	Evidence strength is communicated in padlocks. The evidence is awarded an initial padlock based on the number of studies that meet the inclusion criteria (e.g. 70 or more for 5 padlocks). Padlocks are then reduced where threats to validity are identified: - Small percentage of recent studies - Small percentage of randomised controlled trials - Threats to ecological validity (e.g. delivered by researchers) - Large percentage of studies are not independently evaluated - High unexplained heterogeneity
Cost	The cost estimates are based on the average cost of delivering an intervention. A components based approach is used to measure the per pupil cost, which is then reported on a 5 point scale. For example, one £ is awarded for approaches that cost less than £80 per pupil per year. Full details on cost calculations can be found here.
Months progress	The months progress estimates are based on the effect size calculated using a random-effects meta-analysis. This meta-analytic result is translated into months progress to support accessibility. It uses the assumption that pupils make approximately 1 standard deviation of progress over a year (e.g. an effect size of 0.5 will be 6 months progress.

improving education, they will need to happen alongside efforts to improve pedagogy.

Many of the topics that follow draw heavily on evidence from high-income countries – in particular the United States (USA). It is therefore crucial to carefully consider the contextual barriers to implementation before adopting any strategy to improve learning.

REGION	INCLUDED (N)	TOOLKIT (%)
North America	1924	76%
Europe & Central Asia	349	14%
East Asia & Pacific	83	3%
Middle East & North Africa	82	3%
Sub-Sahara Africa	21	1%
Latin America & Caribbean	9	<1%
South Asia	5	<1%
No Code	62	2%

Table 3.1 Frequency of studies in Toolkit by region overall

3.4 .6

FURTHER INFORMATION

The statistical analysis plan for the database can be found here.

Data extraction for the database is undertaken with three data extraction tools:

- EEF main data extraction, used for all studies;

- EEF Toolkit effect size data extraction, used for all studies;

- strand-specific data extraction (additional codes for each Toolkit strand, such as information about tutors and tutees in peer tutoring, or groups size in small group – used for studies in each strand).

Strand title	Arts participation
Update date	28 June 2021
Number of studies	80
Summary	Moderate impact for very low cost, based on moderate evidence
Cost	Very low
Padlocks	3
Impact	+3 months

What is it?

Global	Arts participation is defined as involvement in artistic and creative activities, such as dance, drama, music, painting or sculpture. It can occur either as part of the curriculum or as an extra-curricular activity. Arts-based approaches may be used in other areas of the curriculum (e.g. the use of drama to develop engagement and oral language before a writing task). Participation may be via regular weekly or monthly activities, or more intensive programmes such as summer schools or residential courses. Whilst these activities have important educational value in themselves, this Toolkit entry focuses on the benefits of arts participation for core academic attainment in other areas of the curriculum, particularly literacy and mathematics.

Key Findings

Global	Arts participation approaches can have a positive impact on academic outcomes in other areas of the curriculum.
Global	The research here summarizes the impact of arts participation on academic outcomes. It is important to remember that arts engagement is valuable in and of itself and its value should be considered beyond mathematics or English outcomes.
Global	If the aim of the arts approach is to improve academic attainment it is important to identify the link between the chosen arts intervention and the outcomes that need to be improved.
Global	Arts-based approaches may offer a route to re-engage older pupils in learning, though this does not always translate into better attainment. It is important to consider how increased engagement will be used to improve teaching and learning for these pupils.

How effective is the approach?

Global	Overall, the average impact of arts participation on other areas of academic learning appears to be positive but moderate, about an additional three months' progress. Improved outcomes have been identified in English, mathematics and science. Benefits have been found in both primary and secondary schools. Some arts activities have been linked with improvements in specific outcomes. For example, there is some evidence of the impact of drama on writing and a potential link between music and spatial awareness. Wider benefits such as more positive attitudes to learning and increased well-being have consistently been reported.

Behind the Average

Global

The impact is similar for both primary and secondary school pupils.

Global

The effects tend to be higher for writing and mathematics than reading.

Closing the disadvantage gap

Local

There is intrinsic value in teaching pupils creative and performance skills and ensuring disadvantaged pupils have access to a rich and stimulating arts education. Arts participation may be delivered within the core curriculum, or through extra-curricular or cultural trips but the latter can be subject to financial barriers for pupils from deprived backgrounds.

There is some evidence to suggest a causal link between arts education and the use of arts-based approaches with overall educational attainment. Where the arts are being taught as a means to boost academic achievement for those eligible for the pupil premium, schools should carefully monitor whether this aim is being achieved.

How could you implement it in your setting?

Global

Arts participation relates to a broad range of subjects including traditional fine arts, theatre, dance, poetry and creative writing. It also includes teaching strategies that explicitly include arts elements, such as drama-based pedagogy.

Some components of arts education approaches might include:

- explicit teaching of creative skills and techniques;
- opportunities for pupils to practise, reflect on their strengths and identify areas for improvement;
- access to materials, equipment, extra-curricular activities and cultural experiences.

Global

Arts education may take the form of regular lessons or monthly activities, after school clubs, small group or one-on-one tuition, or whole school programmes. Activities can also be delivered through more intensive programmes such as summer schools or residential courses.

Local

The average cost of arts education is expected to be very low, with costs ranging from very low to high depending on the type of provision. Costs to schools are largely based on teacher professional development and resources. Costs are greater where activities fall outside of the school day or involve small group or one-on-one tuition from specialist teachers.

Implementing arts education will require a small amount of additional staff time compared with other approaches as it is part of the core curriculum. Arts activities may also involve professional artists, and certified drama or music teachers.

In addition to time and cost, school leaders should consider how to maximize the professional development needs of staff to effectively integrate arts activities (such as drama, visual arts or music) in the classroom and evaluate their impact on pupil outcomes.

When introducing new approaches, schools should consider implementation. For more information see Putting Evidence to Work – A School's Guide to Implementation.

How secure is the evidence?

Global

The security of the evidence around arts participation is rated as moderate. Eighty studies were identified. The topic lost a padlock because a large percentage of the studies were not independently evaluated. Evaluations conducted by organizations connected with the approach, for example, commercial providers, typically have a larger impact, which may influence the overall impact.

As with any evidence review, the Toolkit summarizes the average impact of approaches as researched in academic studies. It is important to consider the context and apply professional judgement when implementing an approach.

Strand title	Behaviour interventions
Update date	18 June 2021
Number of studies	89
Summary	Moderate impact for very low cost, based on moderate evidence
Cost	Very low
Padlocks	2
Impact	+4 months

What is it?

Global	Behaviour interventions seek to improve attainment by reducing challenging behaviour in school. This entry covers interventions aimed at reducing a variety of behaviours, from low-level disruption to aggression, violence, bullying, substance abuse and general anti-social activities. The interventions themselves can be split into three broad categories: 1. Approaches to developing a positive school ethos or improving discipline across the whole school which also aims to support greater engagement in learning. 2. Universal programmes that generally take place in the classroom and seek to improve behaviour. 3. More specialized programmes that are targeted at students with specific behavioural issues. Other approaches, such as parental engagement and SEL programmes, are often associated with reported improvements in school ethos or discipline, but are not included in this summary, which is limited to interventions that focus directly on behaviour.

Key Findings

Global	Both targeted interventions and universal approaches have positive overall effects, about an additional four months' progress. Schools should consider the appropriate combination of behaviour approaches to reduce overall disruption and provide tailored support where required.
Global	There is evidence for a range of different interventions with the highest impacts for approaches that focus on self-management or role-play and rehearsal.
Global	Even within programme types there is a range of impacts. In selecting a behaviour intervention, schools should look for programmes that have been evaluated and shown to have a positive impact.
Global	When adopting behaviour interventions – whether targeted or universal – it is important to consider providing professional development to staff to ensure high-quality delivery and consistency across the school.

How effective is the approach?

Global	The average impact of behaviour interventions is four months' additional progress over the course of a year. Evidence suggests that, on average, behaviour interventions can produce moderate improvements in academic performance along with a decrease in problematic behaviours. However, estimated benefits vary widely across programmes. Approaches such as improving teachers' behaviour management and pupils' cognitive and social skills are both effective, on average. School-level behaviour approaches are often related to improvements in attainment, but there is a lack of evidence to show that the improvements are actually caused by the behaviour interventions, rather than other school interventions happening at the same time. Parental and community involvement programmes are often associated with reported improvements in school ethos or discipline and so are worth considering as alternatives to direct behaviour interventions.

Behind the Average

Global	Effects are slightly lower for secondary school pupils, about an additional three months’ progress.
Global	Impact seems to apply across the curriculum with slightly greater impact, about an additional five months’ progress, for mathematics than for literacy or science.
Global	Frequent sessions several times a week over an extended period of up to a term appear to be the most successful.
Global	Approaches that focus on self-management and those involving role play or rehearsal are associated with greater impact.

Closing the disadvantage gap

Local	According to figures from the Department for Education in the United Kingdom (UK), pupils who receive free school meals are more likely to receive a permanent or fixed period exclusion compared to those who do not. The most common reason for exclusion is persistent disruptive behaviour. Pupil behaviour will have multiple influences, some of which teachers can directly control though universal or classroom management approaches. Some pupils will require more specialist support to help manage their self-regulation or social-emotional skills.

How could you implement it in your setting?

Global	Behaviour interventions have an impact by increasing the time that pupils have for learning. This might be achieved by reducing low-level disruption that impacts learning time in the classroom or by preventing exclusions that remove pupils from school for periods of time. If interventions take up more classroom time than the disruption they displace, engaged learning time is unlikely to increase. In most schools, a combination of universal and targeted approaches will be most appropriate. - Universal approaches to classroom management can help prevent disruption but often require professional development to administer effectively. - Targeted approaches that are tailored to pupils’ needs such as regular report cards or functional behaviour assessments may be appropriate where pupils are struggling with behaviour. In all approaches it is crucial to maintain high expectations for pupils and to embed a consistent approach across the school. Successful approaches may also include SEL interventions and parental engagement.
Global	Evidence suggests that programmes delivered over two to six months produce more long-lasting results. Whole school strategies usually take longer to embed than individually tailored or single classroom strategies.
Local	The costs of behaviour interventions vary widely and overall are estimated to range between very low to moderate. The costs to schools to deliver whole school strategies are largely based on staff time and training. More intensive, targeted interventions are likely to incur higher staffing and training costs. Behavioural interventions can require a large amount of staff time, compared with other approaches. Targeted or one-on-one approaches, delivered by trained school staff or specialists, will require additional staff time compared to universal approaches. Overall, effective approaches can promote better engagement with teaching and learning by reducing challenging behaviour and improving pupil engagement. In addition to time and cost, school leaders should reflect on the impact of whole school behaviour policies and support staff in maintaining a consistent approach. When adopting new approaches, school leaders should consider programmes with a track record of effectiveness. Improving classroom management may involve intensive training where teachers reflect on their practice, implement new strategies and review progress over time. When introducing new approaches, schools should consider implementation. For more information see Putting Evidence to Work – A School’s Guide to Implementation.

How secure is the evidence?

Global

The security of the evidence around behaviour interventions is rated as low. Eighty-nine studies that meet the inclusion criteria for the Toolkit were identified. Overall, the topic lost two additional padlocks because:

- only a small percentage of studies were conducted recently, which might mean that the research is not representative of current practice;

- a large percentage of the studies were not independently evaluated. Evaluations conducted by organizations connected with the approach, for example, commercial providers, typically have larger impacts, which may influence the overall impact of the strand.

As with any evidence review, the Toolkit summarizes the average impact of approaches as researched in academic studies. It is important to consider context and apply professional judgement when implementing an approach.

Strand title	Collaborative learning
Update date	23 June 2021
Number of studies	212
Summary	Moderate impact for very low cost, based on moderate evidence
Cost	Very low
Padlocks	3
Impact	+5 months

What is it?

Global

A collaborative (or cooperative) learning approach involves pupils working together on activities or learning tasks in a group small enough to ensure that everyone participates. Pupils in the group may work on separate tasks contributing to a common overall outcome, or work together on a shared task. This is distinct from unstructured group work.

Some collaborative learning approaches put pairs, groups or teams of mixed attainment to work in competition with each other in order to drive more effective collaboration. There is a very wide range of approaches to collaborative and cooperative learning involving many different kinds of organization and tasks. Peer tutoring can also be considered as a type of collaborative learning, but is reviewed as a separate topic in the Toolkit.

The collaborative learning approaches adopted by schools were typically implemented for 8 weeks over the course of a school year. The average impact, however, includes longer and shorter collaborative learning approaches.

Key findings

Global	Collaborative learning approaches have a positive impact, on average, and may be a cost-effective approach for raising attainment.
Global	Pupils need support and practise in working together; it does not happen automatically. Professional development can support the effective management of collaborative learning activities.
Global	Tasks and activities need to be designed carefully so that working together is effective and efficient, otherwise some pupils may struggle to participate or try to work on their own. It is important to ensure that all pupils talk and articulate their thinking in collaborative tasks to ensure they benefit fully.
Global	Competition between groups can be used to support pupils in working together more effectively. However, an over emphasis on competition can cause learners to focus on winning rather than succeeding in their learning.
Global	The most promising collaborative learning approaches tend to have group sizes between three and five pupils who have a shared outcome or goal.

How effective is the approach?

Global	The impact of collaborative approaches on learning is consistently positive, with pupils making about an additional five months' progress, on average, over the course of an academic year. However, the size of impact varies, so it is important to get the detail right. Collaborative learning can describe a large variety of approaches, but effective collaborative learning requires much more than just sitting pupils together and asking them to work in pairs or a group; structured approaches with well-designed tasks lead to the greatest learning gains. There is some evidence that collaboration can be supported with competition between groups, but this is not always necessary, and can lead to learners focusing on the competition rather than the learning it aims to support. Most of the positive approaches include the promotion of conversation and interaction between learners. The evidence indicates that groups of three to five are most effective for collaborative learning approaches – there are smaller positive impacts for both paired work and collaborative learning activities with more than five pupils in a group. There is also some evidence that collaborative learning approaches are particularly promising when they are used to teach science.

Behind the average

Global	The effects of collaborative learning are slightly higher in secondary schools (about an additional six months' progress) than in primary schools (about an additional five months' progress).
Global	The impact of collaborative learning is slightly lower in literacy (about an additional three months' progress) than mathematics (about an additional five months' progress) and science (about an additional ten months' progress).
Global	Small groups of three to five pupils with responsibility for a joint outcome appear to be the most successful structure.
Global	Studies that deliver collaborative learning through digital technology tend to have a lower impact, about an additional three months' progress overall.

Closing the disadvantage gap

Local	There is limited evidence on differential impact for pupils from disadvantaged backgrounds. There is some evidence that collaborative learning approaches may benefit those with low prior attainment by providing opportunities for pupils to work with peers to articulate their thinking, share knowledge and skills, and address misconceptions through peer support and discussion. It is crucial that support is provided through well-structured and carefully designed learning activities to ensure that lower-attaining pupils are involved, challenged and learn successfully. If collaborative learning approaches just involve high-attaining pupils solving problems with no input from their peers, this is likely to widen existing gaps in attainment.

Applications and Approaches	
Name	Collaborative learning with joint outcomes
Impact	+7 months
Number of studies	111
Summary	When groups conducting collaborative learning activities are given a joint group outcome to work towards, the impact of the approach is typically higher than average. One hundred and eleven studies in which pupils worked towards a joint outcome were identified.

Applications and Approaches	
Name	Collaborative learning with individual outcomes
Impact	+4 months
Number of studies	101
Summary	Some collaborative learning activities give different children within the group different objectives to accomplish. Overall, these approaches have positive outcomes, but the impact is typically slightly lower than those with shared group outcomes. One hundred and one studies in which individual outcomes were given to pupils within collaborative learning activities were identified.

How could you implement it in your setting?

Global

There are many theories about how collaborative learning might benefit pupil outcomes. Through collaboration, pupils may develop explanation, demonstration, problem-solving and metacognitive skills, or pupils may benefit from sharing the load of challenging tasks. It is important that schools ensure that within collaborative learning:

- all pupils, particularly pupils with low prior attainment, are supported to fully participate;
- the make-up of pairings and groups is carefully considered;
- teachers promote good practice in collaboration, for example, modelling high-quality discussions so that collaborative activities are productive;
- teachers carefully monitor collaborative activities and support pupils who are struggling or not contributing.

Global

There is a broad range of approaches to collaborative or cooperative learning involving different kinds of organization and tasks across the curriculum. Not all of the specific approaches to collaborative learning adopted by schools have been evaluated, so it is important to evaluate any new initiative in this area. Professional development is likely to be required to maximize the effectiveness of approaches and monitor the impact of different approaches in the classroom.

Local

The average cost of collaborative learning is expected to be very low with the cost to schools largely in teacher training and resources. As a classroom-based approach, implementing collaborative learning will also require a small amount of staff time for planning and monitoring, compared with other approaches.

In addition to time and cost, school leaders should consider how to maximize the effectiveness of collaborative learning through teacher professional development to support the use of well-designed tasks. They should carefully monitor the impact of approaches on lower-attaining pupils.

When introducing new approaches, schools should consider implementation. For more information see Putting Evidence to Work – A School's Guide to Implementation.

How secure is the evidence?

Global

The security of the evidence around collaborative learning interventions is rated as low. Two hundred and twelve studies that meet the inclusion criteria of the Toolkit were identified. The topic lost three padlocks because:

- only a small percentage of studies have taken place recently, which might mean that the research is not representative of current practice;
- a large percentage of the studies were not independently evaluated. Evaluations conducted by organizations connected with the approach, for example, commercial providers, typically have larger impacts, which may influence the overall impact of the strand;
- there is a large amount of unexplained variation between the results included in the topic. All reviews contain some variation in results, which is why it is important to look behind the average. Unexplained variation (or heterogeneity) reduces certainty in the results in ways that we have been unable to test by looking at how context, methodology or approach is influencing impact.

As with any evidence review, the Toolkit summarizes the average impact of approaches as researched in academic studies. It is important to consider context and apply professional judgement when implementing an approach.

3 CHAPTER

Strand title	Extending school time
Update date	9 June 2021
Number of studies	74
Summary	Moderate impact for moderate cost, based on limited evidence
Cost	Moderate
Padlocks	2
Impact	+3 months

What is it?

Global	Extending school time involves increasing learning time in schools during the school day or changing the school calendar. This can include extending core teaching and learning time in schools as well as the use of targeted before and after school programmes (including additional small group or one-on-one tuition). It also includes revisions to the school calendar to extend the total number of days in the school year. The median intervention length for approaches that extended the school year was forty weeks (an extended school day for one academic year). Other approaches to increasing learning time, such as homework and summer schools, are included in other sections of the Toolkit.

Key Findings

Global	Programmes that extend school time have a positive impact on average but are expensive and may not be cost-effective for schools to implement. Schools will also need to consider the workload and well-being of staff.
Global	Planning to get the most from any extra time is important. It should meet pupils' needs and build on their capabilities. Where additional time is voluntary, it is important to monitor attendance to ensure pupils who need additional support can benefit.
Global	Before and after school programmes with a clear structure, a strong link to the curriculum, and well-qualified and well-trained staff are more clearly linked to academic benefits than other types of extended hours provision.
Global	Additional school time may be more effective if it is used for one-on-one support, in contrast to small or large group teaching.
Global	Enrichment activities without a specific focus on learning can have an impact on attainment, but the effects tend to be lower and the impact of different interventions can vary a great deal (see entries for physical activity or arts participation). These interventions may, however, be beneficial for their own sake outside of any attainment impacts.

How effective is the approach?

Global	The average impact of approaches involving extending school time is about an additional three months' progress over the course of a year. The average impact is influenced by the targeted use of before and after school programmes, which have higher impacts, on average. The impact is also slightly lower when school time is extended in secondary school. In addition to providing academic support, some school programmes aim to provide stimulating environments and activities or develop additional personal and social skills. These programmes are more likely to have an impact on attainment than those that are solely academic in focus. However, it is not clear whether this is due to the additional activities or to improved attendance and greater engagement. The research also indicates that attracting and retaining pupils in before and after school programmes is harder at secondary level than at primary level. To be successful, any extension of school time should be supported by both parents and staff. It should also be noted that more extreme increases may have diminishing effects if engagement of pupils is reduced. While the impact on academic attainment is, on average, positive, the cost of extending school times might mean that it is not a cost-effective approach to implement at the school level without additional funding.

Behind the average

Global	More studies have been undertaken in primary schools. Effects are higher for primary (about an additional three months' progress) than secondary (about an additional two months' progress) schools.
Global	Most of the evidence relates to literacy and mathematics with similar effects in both subjects.
Global	More intensive approaches in extended time, such as one-on-one, appear to be more effective than either small group or large group teaching.
Local	Most studies have been conducted in the United States of America (USA) – this could pose a risk to the transferability of findings as impacts may be influenced by the average length of regular education in any given context.

Closing the disadvantage gap

Local	There is some evidence to suggest that disadvantaged pupils might benefit more from additional time at school. To increase the likelihood of additional school time benefiting disadvantaged pupils, school leaders should consider how to secure engagement and attendance among those from disadvantaged backgrounds. It is possible that if targeted tuition or enrichment activities are offered universally, those who could benefit the most would be the least likely to participate or engage. However, adopting a more targeted approach also has its challenges, as selected pupils may feel singled out and stigmatized. Additional non-academic activities may also provide free or low-cost alternatives to sport, music and other enrichment activities that more advantaged families are more likely to be able to pay for outside of school.

How could you implement it in your setting?

Global

The theory behind extending school time is that extra hours of allocated learning mean that pupils have more exposure to teaching, more time to engage with content and a greater amount of learning overall. When implementing approaches that extend school time it is important to acknowledge that allocated learning time and actual learning time are not the same thing. Schools should:

- carefully monitor attendance to ensure that extensions to the school day or term do not lead to reductions in overall learning time for some pupils;

- carefully consider and monitor pupil engagement – if more time is spent managing pupil behaviour in a longer school day then engaged learning time may not increase;

- monitor staff well-being and workload to ensure that additional teaching time does not reduce quality (e.g. through less time for professional development or planning lessons).

Extending school time is likely to require a significant reconfiguration of working patterns for staff, especially if this involves an altered school calendar. It is important that school leaders are clear regarding the purpose of introducing additional learning time and secure parental support prior to making changes.

Global

Approaches to extending school time are likely to be spread over an academic year. Some schools may also decide to target additional support at specific classes or pupils during particular school terms or times of the year.

Local

If additional teachers are not hired to cover the increase in teaching time that comes from extending school time, any increases to school calendars or timetables may also require a large amount of staff time, compared with other approaches.

In addition to time and cost, school leaders should consider how to ensure the quality of teaching during additional school time and avoid approaches that could increase teacher workload without making significant impacts on pupil learning.

When introducing new approaches, schools should consider implementation. For more information see Putting Evidence to Work – A School's Guide to Implementation.

What does it cost?

Local

Overall, costs are estimated as moderate. The basic cost of teaching a pupil is about £3,120 a year (£16 per day) in primary school and about £4,680 a year (£25 per day) in secondary school. Extending the school year by two weeks would therefore require about £160 per pupil per year for primary schools and about £250 per pupil per year for secondary schools. Estimates suggest that after school clubs cost, on average, £7 per session per pupil. A weekly session would therefore cost £273 per pupil over the course of a thirty-nine-week school year.

How secure is the evidence?

Global

The security of the evidence around extending school time is rated as moderate. Seventy-four studies that meet the inclusion criteria of the Toolkit were identified. Overall, the topic lost an additional padlock because a large percentage of the studies are not randomized controlled trials. While other study designs still give important information about the effectiveness of the approaches, there is a risk that results are influenced by unknown factors that are not part of the intervention.

As with any evidence review, the Toolkit summarizes the average impact of approaches when researched in academic studies. It is important to consider context and apply professional judgement when implementing an approach.

Next steps	
Local	Putting Evidence to Work – A School's Guide to Implementation.

Strand title	Feedback
Update date	4 June 2021
Number of studies	155
Summary	High impact for very low cost, based on extensive evidence
Cost	Very low
Padlocks	4
Impact	+6 months

What is it?	
Global	Feedback is information given to the learner about their performance relative to learning goals or outcomes. It should aim to produce (and be capable of producing) improvement in students' learning. Feedback redirects or refocuses the learner's actions to achieve a goal by aligning effort and activity with an outcome. It can be about the output or outcome of the task, the process of the task, the student's management of their learning or self-regulation, or about them as individuals (which tends to be the least effective). Feedback can be verbal or written or can be given via tests or digital technology. It can come from a teacher or someone taking a teaching role, or from peers (see Peer tutoring).

Key findings	
Global	Providing feedback is well-evidenced and has a high impact on learning outcomes. Effective feedback tends to focus on the task, subject and self-regulation strategies: it provides specific information to the student on how to improve.
Global	Feedback can be effective during, immediately after and some time after learning. Feedback policies should not over specify the frequency of feedback.
Global	Feedback can come from a variety of sources – studies have shown positive effects of feedback from teachers and peers. Feedback delivered by digital technology also has positive effects (albeit slightly lower than the overall average).
Global	Different methods of feedback delivery can be effective and feedback should not be limited exclusively to written marking. Studies of verbal feedback show slightly higher impacts overall an additional seven months' progress). Written marking may form one part of an effective feedback strategy but it is crucial to monitor impacts on staff workload.
Global	It is important to give feedback when work is correct – not just when it is incorrect. High-quality feedback may focus on a task, subject or self-regulation strategies.

How effective is the approach?	
Global	Feedback studies tend to show high effects on learning. However, there are a wide range of effects and some studies show that feedback can have negative effects and even make things worse. There are positive impacts from a wide range of feedback approaches – including when feedback is delivered by technology or peers. Impacts are greatest when feedback is delivered by teachers. It is particularly important to provide feedback when work is correct, rather than just using it to identify errors. Many studies of feedback also include other practices. For example, mastery learning approaches combine feedback with additional support for pupils who are falling behind, while approaches like formative assessment also include work to understand specific gaps in learning that need to be addressed and how the teacher wants the pupil to progress. Feedback has effects across all age groups. Research in schools has focused particularly on its impact on English, mathematics and, to a lesser extent, science.
Local	Embedding formative assessment explicitly can be a key component in laying the foundations for effective feedback. The EEF has trialled 'Embedding Formative Assessment' in English schools and found a positive impact, on average.

Behind the average	
Global	Feedback appears to have slightly greater effects for primary school pupils (about an additional seven months' progress) than for secondary school pupils (about an additional five months' progress).
Global	Effects are high across all curriculum subjects, with slightly higher effects in mathematics and science.
Global	Low-attaining pupils tend to benefit more from explicit feedback than high attainers.
Global	Although some studies have successfully demonstrated the benefits of digital feedback, effects are typically slightly smaller (about an additional four months' progress).

Closing the disadvantage gap	
Local	There is evidence to suggest that feedback involving metacognitive and self-regulatory approaches may have a greater impact on disadvantaged pupils and lower prior attainers than other pupils. Pupils require clear and actionable feedback to employ metacognitive strategies as they learn, as this information informs their understanding of their specific strengths and areas for improvement, thereby indicating which learning strategies have been effective for them in previously completed work.

Applications and Approaches	
Name	Written feedback
Impact	5 months
Number of studies	104

Summary	Written feedback typically involves both marks or grades and comments. It is generally given to pupils after they have completed a task and is usually intended for them to read on their own. The impact of written feedback is typically a little lower than the overall impact. Average progress is five months. This impact includes all forms of written feedback. The evidence for specific approaches such as 'triple marking' is much more limited. It is especially important that schools monitor teachers' workload in the use of written feedback. Given it is not clear when feedback can be most effective, feedback policies should not over specify the timing of feedback.

Applications and Approaches	
Name	Oral feedback
Impact	7 months
Number of studies	67
Summary	Oral feedback typically involves spoken comments from the teacher, either to an individual, group or class. It tends to be more immediate than written feedback and is usually given either during or at the conclusion of a task or activity. The impact of oral feedback is higher, on average, than the impact of feedback overall. Average progress is seven months. Whilst recognizing the potential benefits of oral feedback, this finding should not supplant the necessity to consider the principles that underpin the teacher feedback to improve pupil learning guidance report. While oral feedback has a slightly higher positive effect on average, most schools will want to use a range of methods for providing feedback and it is important to focus on quality within each medium.

How could you implement it in your setting?

Global

Feedback may have a positive impact by: supporting pupils to focus future learning on areas of weakness; identifying and explaining misconceptions; supporting them in taking greater responsibility for their own improvement; or increasing pupils' motivation to improve.

Implementing feedback successfully will require:
- communication with pupils, teachers and parents/caregivers about practices and expectations that relate to feedback policies;
- assessment of pupil understanding to ascertain what needs to be improved;
- consideration of the 'opportunity cost' associated with different feedback practices;
- ensuring that feedback can be acted upon, for example, by including specific information regarding what a pupil has done successfully or not, and an explanation as to why;
- careful consideration of how feedback will be received, including impacts on self-confidence and motivation;
- providing opportunities for pupils to act upon the feedback after it has been given;
- evaluation of how effective the feedback has been.

Global

Feedback interventions vary in length. Some are short, targeted approaches that address pupil misconceptions within weeks or even days. Others are used as more extended methods of tracking and supporting pupil progress over many months.

Local

The average cost of feedback and feedback interventions is very low. The cost to schools is largely in training.

Implementing feedback and feedback interventions will also require a moderate and sustained amount of staff time, compared with other approaches.

In addition to time and cost, school leaders should consider how to maximize teacher professional development in supporting them to deliver effective feedback and avoid approaches that increase teacher workload without providing pupils with the necessary information to improve performance.

When introducing new approaches, schools should consider implementation. For more information see Putting Evidence to Work – A School's Guide to Implementation.

Relevant EEF studies

Embedding Formative Assessment

Anglican Schools Partnership

How secure is the evidence?

Global

The security of the evidence around feedback is rated as high. One hundred and fifty-five studies that meet the inclusion criteria of the Toolkit were identified. The topic lost a padlock because a large percentage of the studies are not randomized controlled trials. While other study designs still give important information about the effectiveness of approaches, there is a risk that the results are influenced by unknown factors that are not part of the intervention.

As with any evidence review, the Toolkit summarizes the average impact of approaches as researched in academic studies. It is important to consider context and apply professional judgement when implementing an approach.

FEEDBACK & THEORY OF CHANGE

Evidence suggests that disadvantaged pupils and low prior attainers can benefit more from meta-cognitive interventions than the average pupil.

In order for pupils to employ metacognitive processes and thinking to their learning, they need clear and accurate feedback on their strengths and areas of improvement to know what actions and practices to repeat, and what to do differently.

Therefore, feedback that is clear and accurately reflects pupils' strengths and weaknesses may therefore bring about larger improvements for disadvantaged pupils and/or low prior attainers.

Strand title	Mastery learning
Update date	24 June 2021
Number of studies	80
Summary	Moderate impact for very low cost, based on very limited evidence
Cost	Very low cost
Padlocks	2
Impact	+5 months

What is it?

Global	Mastery learning was originally developed in the 1960s. According to an early definition of mastery learning, learning outcomes are kept constant but the time needed for pupils to become proficient or competent in these objectives varies. Subject matter is broken into blocks or units with predetermined objectives and specified outcomes. Learners must demonstrate mastery on unit tests, typically 80 per cent, before moving on to new material. Pupils who do not achieve mastery are provided with extra support through a range of teaching strategies such as more intensive teaching, tutoring, peer-assisted learning, small group discussions or additional homework. Learners continue the cycle of studying and testing until the mastery criteria are met. More recent mastery approaches do not always have all the characteristics of mastery learning. Some approaches without a threshold typically involve the class moving on to new material when the teacher decides that the majority of pupils have mastered the unit. Curriculum time varies according to the progress of the class. In other approaches, pupils are required to demonstrate mastery on a test to progress to new material, but there is not a specified threshold of at least 80 per cent.
Local	Mastery learning should be distinguished from a related approach which is sometimes known as 'teaching for mastery'. This term is often used to describe the approach to mathematics teaching found in high-performing places in East Asia, such as Shanghai and Singapore. Like mastery learning, teaching for mastery aims to support all pupils to achieve deep understanding and competence in the relevant topic. However, teaching for mastery is characterized by teacher-led, whole class teaching; common lesson content for all pupils; and use of manipulatives and representations. Although some aspects of teaching for mastery are informed by research, relatively few interventions of this nature have been evaluated for impact. Most of the studies in this strand should be distinguished from this related approach.

WORKING GROUP 4

3

CHAPTER

Key findings

Global	Mastery learning is a cost-effective approach, on average, but is challenging to implement effectively. Schools should plan for changes and assess whether the approach is successful within their context.
Global	A high level of success should be required before pupils move on to new content – it is crucial to monitor and communicate pupil progress and to provide additional support for pupils who take longer to reach the required level of knowledge.
Global	Mastery learning approaches are often associated with direct instruction, but many of the high-impact studies identified included elements of collaborative learning.
Global	There is a large variation in the average impact – mastery learning approaches have consistently positive impacts, but effects are higher for primary school pupils and in mathematics.

How effective is the approach?

Global	The impact of mastery learning approaches is an additional five months' progress, on average, over the course of a year. There is a lot of variation in this average. It seems to be important that a high bar is set for achievement of 'mastery' (usually 80 per cent to 90 per cent on the relevant test). By contrast, the approach appears to be much less effective when pupils work at their own pace (see also Individualized instruction). Mastery learning also appears to be particularly effective when pupils are given the opportunity to work in groups or teams and take responsibility for supporting each other's progress (see also Collaborative learning and Peer tutoring).
Local	The EEF evaluation of 'Maths Mastery' – an example of the 'Teaching for Mastery' approach, found positive impacts overall – but with a slightly lower effect than the average impact for more traditional mastery approaches.

Behind the average

Global	Studies involving primary school pupils have tended to be more effective (about an additional eight months' progress) than for secondary school pupils (about an additional three months' progress).
Global	Mastery learning has been used successfully across the curriculum but particularly for reading, mathematics and science. Effects are higher in mathematics and science (about an additional six months' progress) than reading (about an additional three months' progress).
Global	A high level of mastery of about 80 per cent is associated with more successful approaches.
Global	Mastery learning approaches that include collaborative learning can be particularly effective.

Closing the disadvantage gap

Local	Mastery learning approaches aim to ensure that all pupils have mastered key concepts before moving on to the next topic – in contrast with traditional teaching methods whereby pupils may be left behind, with gaps of misunderstanding widening. Mastery learning approaches could address these challenges by giving additional time and support to pupils who may have missed learning, or who take longer to master new knowledge and skills. In order for mastery approaches to be effective for pupils with gaps in understanding, it is crucial that additional support is provided. Approaches that simply build upon foundational knowledge without targeting support for pupils who fall behind are unlikely to narrow disadvantage gaps.

How could you implement it in your setting?

Global

Mastery learning works through designing units of work so that each task has a clear learning outcome, which pupils must master prior to moving on to the next task. Core components of the mastery approach that schools should be careful to implement include:

- effective diagnostic assessment to identify areas of strength and weakness;

- careful sequencing of topics so that they gradually build on foundational knowledge;

- flexibility for teachers on how long they need to spend on any particular topic;

- monitoring of pupil learning and regular feedback so that pupils can master topics prior to moving to the next;

- additional support for pupils who struggle to master topic areas.

Global

Mastery learning interventions are typically delivered over the course of an academic year, as choosing to take longer on a topic or scheme of work requires flexibility in the planning and teaching of curriculum content.

Some schools may decide that certain topics are more suited to a mastery approach than others, and therefore the delivery time could be as short as half a term.

Local

Overall, the median costs of implementing mastery learning approaches are estimated as very low. The costs associated with mastery learning approaches mostly arise from professional development training for teaching staff, which is most commonly a start-up cost for introducing the new approach.

Whilst the average cost estimate for mastery learning is very low, the range in costs of professional development training, and the option to pay for ongoing training and additional staff to provide greater timetable flexibility, mean that costs can range from very low to moderate.

Implementing mastery learning also requires a moderate amount of staff time, compared with other approaches. School leaders should be aware of the extra staff time required and think carefully about other activities they might need to cut back on to provide this additional support.

In addition to time and cost, school leaders should consider how to maximize support for struggling learners and avoid some pupils getting bored or frustrated whilst they wait for others to master content.

When introducing new approaches, schools should consider implementation. For more information see Putting Evidence to Work – A School's Guide to Implementation.

How secure is the evidence?

Global

The security of the evidence around mastery learning is rated as low. Eighty studies that meet the inclusion criteria of the Toolkit were identified. Overall, the topic lost two additional padlocks because:

- only a small percentage of studies have taken place recently, which might mean that the research is not representative of current practice;

- a large percentage of the studies are not randomized controlled trials. While other study designs still give important information about the effectiveness of approaches, there is a risk that results are influenced by unknown factors that are not part of the intervention.

As with any evidence review, the Toolkit summarizes the average impact of approaches when researched in academic studies. It is important to consider context and apply professional judgement when implementing an approach.

MASTERY LEARNING – THEORY OF CHANGE

On average, disadvantaged pupils are more likely have higher school absence rates

Disadvantaged pupils are more likely to be previously low-attaining pupils – meaning they gain less knowledge or proficiency in skills than their more privileged counterparts.

Mastery learning approaches moderate against lost learning by giving pupils with less proficiency in a topic further support and opportunity to address gaps in their understanding.

Mastery learning mitigates against gaps in understanding widening as pupils progress through school, and as a result can help support pupils who may be left behind through other approaches.

Strand title	Mentoring
Update date	4 June 2021
Number of studies	44
Summary	Low impact for moderate cost, based on moderate evidence
Cost	Moderate
Padlocks	3
Impact	+2 months

What is it?

Global

Mentoring in education involves pairing young people with an older peer or adult volunteer who acts as a positive role model. In general, mentoring aims to build confidence and relationships, develop resilience and character, or raise aspirations, rather than develop specific academic skills or knowledge.

Mentors typically build relationships with young people by meeting with them one-on-one for about an hour a week over a sustained period, either during school, at the end of the school day or at weekends. In some approaches mentors may meet with their mentees in small groups.

Activities vary between different mentoring programmes. While some include academic support with homework or other school tasks, approaches focused primarily on direct academic support (sometimes referred to as 'academic mentoring') are not covered in this strand. See One-on-one tuition and Peer tutoring.

Mentoring has increasingly been offered to young people who are deemed to be hard to reach or who are at risk of educational failure or exclusion.

Key findings

Global	The impact of mentoring varies but, on average, it is likely to have a small positive impact on attainment.
Global	Positive effects on attainment tend not to be sustained once the mentoring stops, so care must be taken to ensure that benefits are not lost. It is important to consider how pupils who have benefited from mentoring can be supported to retain positive changes in their
Global	Both community-based and school-based approaches can be successful.
Global	Mentor drop-out can have detrimental effects on mentees. It is important to consider how to support mentors.

How effective is the approach?

Global	On average, mentoring appears to have a small but positive impact on academic outcomes. The impacts of individual programmes vary. Some studies have found more positive impacts for pupils from disadvantaged backgrounds, and for non-academic outcomes such as attitudes to school, attendance and behaviour. There are risks associated with unsuccessful mentor pairings, which may have a detrimental effect on the mentee, and some studies report negative overall impacts. Programmes that have a clear structure and expectations, provide training and support for mentors, and recruit mentors who are volunteers, are associated with more successful outcomes. There is no evidence that approaches with a single focus on improving academic attainment or performance are more effective; programmes with multiple objectives can be equally or more effective.

Behind the average

Global	Studies have been undertaken in both primary and secondary school settings with similar impacts.
Global	Overall impact on mathematics and general school subjects tends to be higher than on reading or science outcomes.
Global	Regular meetings of once a week or more frequently appear to be most effective.

Closing the disadvantage gap

Local	While mentoring is not generally as effective in raising attainment outcomes as small group or one-on-one tuition, it is possible to target the approach to pupils from disadvantaged backgrounds and those with particular needs. Some evidence suggests that some pupils from disadvantaged backgrounds show low engagement with or have low expectations of schooling. Mentoring interventions may be more beneficial for these pupils, as the development of trusting relationships with an adult or older peer can provide a different source of support. Mentors dropping out of programmes can result in detrimental effects for pupils, particularly for those who may have already experienced disillusionment at their perceived lack of support from teachers and school. Therefore, additional care should be taken in the recruitment of reliable mentors when interventions are being used to support disadvantaged pupils.

How could you implement it in your setting?	
Global	Mentoring requires close interaction between an adult or older peer and one pupil or a small group of pupils. Conversations between mentors and mentees may address but would not be limited to: attitudes to school; specific academic skills or knowledge; self-perception and belief, particularly in relation to schoolwork; and aspirations for future studies and career options. It is important to consider what support mentors might require to effectively deliver mentoring. Mentoring interactions normally occur one-on-one between mentor and one mentee – although mentors can mentor multiple pupils. Some mentoring approaches also include small group interactions.
Global	Mentoring interventions are typically delivered over an extended period of time (often at least the length of a school year) in order to allow mentors and mentees to develop more lasting and trusting relationships. Frequent regular meetings of once a week or more tend to be more beneficial.
Local	The average cost of a mentoring intervention is moderate. The cost to schools largely involves mentor training, salary costs (for non-volunteer mentors) and resources. Some programmes also include continuous training and support for mentors which may increase costs. Compared with other approaches, implementing mentoring interventions requires a moderate and sustained amount of staff time. In addition to time and cost, school leaders should consider how to maximize the recruitment of effective and reliable mentors who are well matched to mentees. Consideration should also be given to how any gains made in pupil confidence, resilience or aspiration are to be maintained after the intended period of mentoring, as studies show these changes can be difficult to sustain. When introducing new approaches, schools should consider implementation. For more information see Putting Evidence to Work – A School's Guide to Implementation.

NOTES – MENTORING THEORY OF CHANGE

Pupils may experience negative effects of teacher labelling, poor prior attainment, or low-ability grouping. As a result, pupils may become disaffected with school and teacher authority.

Pupils may lack the necessary confidence, resilience or aspiration to apply themselves at school, which prevents them from achieving their potential.

Mentors are paired with pupils who model the skills and knowledge the pupils are lacking, as well as offering some support and advice about school and future aspirations.

Pupils experience an increase in self-belief due to the trusted relationship with their mentors, investing more time and effort into their studies and future educational and career goals.

Pupils improve attainment outcomes by investing more time and effort, motivated by revised aspirations and self-belief.

Strand title	Metacognition and self-regulation
Update date	29 June 2021
Number of studies	246
Summary	+7 months' impact for very low cost, based on extensive evidence
Cost	Very low
Padlocks	4
Impact	+7 months

What is it?	
Global	Metacognition and self-regulation approaches to teaching support pupils to think about their own learning more explicitly, often by teaching them specific strategies for planning, monitoring and evaluating their learning. Interventions are usually designed to give pupils a repertoire of strategies to choose from and the skills to select the most suitable strategy for a given learning task. Self-regulated learning can be broken into three essential components: - cognition – the mental process involved in knowing, understanding and learning; - metacognition – often defined as 'learning to learn'; and - motivation – willingness to engage our metacognitive and cognitive skills.

Key findings	
Global	The potential impact of metacognition and self-regulation approaches is high (more than seven months of additional progress), although it can be difficult to realize this impact in practice as such methods require pupils to take greater responsibility for their learning and develop their understanding of what is required to succeed.
Global	The evidence indicates that explicitly teaching strategies to help plan, monitor and evaluate specific aspects of students' learning can be effective.
Global	These approaches are more effective when they are applied to challenging tasks rooted in standard curriculum content.
Global	Teachers can demonstrate effective use of metacognitive and self-regulatory strategies by modelling their own thought processes. For example, teachers might explain their thinking when interpreting a text or solving a mathematical task, alongside promoting and developing metacognitive talk related to lesson objectives.
Local	Professional development can be used to develop a mental model of metacognition and self-regulation, alongside an understanding of teaching metacognitive strategies.

How effective is the approach?	
Global	The average impact of metacognition and self-regulation strategies is an additional seven months' progress over the course of a year. Metacognition and self-regulation strategies can be effective when taught in collaborative groups so that learners can support each other and make their thinking explicit through discussion.

Behind the average

Global	Studies involving primary school pupils have typically been more effective (about an additional eight months' progress) than those involving secondary school pupils (about an additional seven months' progress).
Global	Metacognitive and self-regulation strategies have been used across curricula, with approaches in mathematics and science particularly successful.
Global	Studies that use digital technology, for example, intelligent tutoring systems that scaffold learning, have particularly high impacts on pupil outcomes.

Closing the disadvantage gap

Local	There is some evidence to suggest that disadvantaged pupils are less likely to use metacognitive and self-regulatory strategies without being explicitly taught them. Explicit teaching of metacognitive and self-regulatory strategies could therefore encourage such pupils to practise and use these skills more frequently in the future. With explicit teaching and feedback, pupils are more likely to use these strategies independently and habitually, enabling them to manage their own learning and overcome challenges on their own in the future.

How could you implement it in your setting?

Global	Self-regulation and metacognition strategies work through learners monitoring and evaluating their own learning strategies. Some necessary components for successful metacognitive strategies might include: - explicit teaching of metacognitive strategies; - teachers modelling their own thinking to demonstrate metacognitive strategies; - opportunities for pupils to reflect on and monitor their strengths and areas for improvement, and plan how to overcome current difficulties; - providing enough challenges for learners to develop effective strategies, but not so difficult that they struggle to apply them.
Global	Metacognition and self-regulation strategies are most effective when they are embedded in a school's curriculum and a specific subject lesson. For example, teaching metacognitive strategies to self-evaluate an essay in history will be different for a pupil evaluating their methods for mathematical problem-solving. When introducing new approaches, schools should consider implementation. For more information see Putting Evidence to Work – A School's Guide to Implementation.

Local

Overall, the median costs of implementing metacognition and self-regulation strategies are very low. The costs associated with metacognition and self-regulation are mostly in professional development training for staff, which is most commonly a start-up cost to embed the approach into the school's curriculum.

Whilst the median cost estimate for metacognition and self-regulation strategies is very low, the variation in the cost of professional development training, and the option to purchase additional materials and provide ongoing training and support, means that costs can range from very low to low. Evidence suggests that the effectiveness of metacognition and self-regulation strategies is influenced by teachers' understanding of how to develop pupils' metacognitive knowledge.

These cost estimates assume that schools are already paying for staff salaries, materials and equipment for teaching, and facilities to host lessons. These are all prerequisite costs of implementing metacognition and self-regulation strategies, without which the cost is likely to be higher.

Implementing metacognition and self-regulation strategies also requires a small amount of staff time, compared with other approaches, as staff need to develop their own understanding of metacognitive and self-regulatory processes to model effective use of these strategies and skills to pupils.

In addition to time and cost, school leaders should consider how to maximize explicit teaching of metacognitive strategies by supporting teachers to use these approaches in their practice. At the same time, school leaders should be careful to avoid alienating teachers who do not feel confident in their knowledge or implementation of these strategies.

How secure is the evidence?

Global

The security of the evidence around metacognition and self-regulation strategies is rated as high. Two hundred and forty-six studies were identified. The topic lost a padlock because a large percentage of the studies were not independently evaluated. Evaluations conducted by organizations connected with the approach, for example, commercial providers, typically have larger impacts, which may influence the overall impact of the strand.

As with any evidence review, the Toolkit summarizes the average impact of approaches when researched in academic studies. It is important to consider context and apply professional judgement when implementing an approach.

METACOGNITION AND SELF-REGULATION – THEORY OF CHANGE

There is some evidence to suggest that disadvantaged pupils are less likely to use metacognitive and self-regulatory strategies.

Explicit teaching of metacognitive strategies should help learners who are less likely to practise/use these skills to use them more frequently in the future.

Disadvantaged pupils may develop these skills and start to use them independently and out of habit. This will help them overcome challenges in the future.

Strand title	One-on-one tuition
Update date	27 May 2021
Number of studies	123
Summary	Moderate impact for moderate cost, based on moderate evidence
Cost	Moderate
Padlocks	3
Impact	+5 months

What is it?

Global	One-on-one tuition involves a teacher, teaching assistant (TA) or other adult providing intensive individual support to a pupil. It may happen outside of normal lessons as additional teaching, for example, as part of extending school time or a summer school, or as a replacement for other lessons.

Key findings

Global	On average, one-on-one tuition is very effective at improving pupil outcomes. It might be an effective strategy for providing targeted support for pupils who have low prior attainment or are struggling in particular areas.
Global	Tuition is more likely to make an impact if it is additional to and explicitly linked with normal lessons.
Global	One-on-one tuition can be expensive to deliver, particularly when it is delivered by teachers. Approaches that deliver either instruction through TAs or in small groups rather than one-on-one have smaller positive effects, on average, but may be a cost-effective solution to providing targeted support.
Global	For one-on-one tuition led by TAs, interventions are likely to be particularly beneficial when the TAs are experienced, well-trained and supported, for example, delivering a structured intervention.

How effective is the approach?

Global	Evidence indicates that one-on-one tuition can be effective, providing approximately five additional months' progress on average. Short, regular sessions (about thirty minutes, three to five times a week) over a set period of time (up to ten weeks) appear to have optimum impact. Evidence also suggests tuition should be additional to, but explicitly linked with, normal teaching, and that teachers should monitor progress to ensure the tutoring is beneficial. Studies comparing one-on-one with small group tuition show mixed results. In some cases one-on-one tuition has led to greater improvement, while in others tuition in groups of two or three has been equally or even more effective. The variability in findings may suggest the importance of the particular type or quality of teaching enabled by very small groups, rather than the precise size of the group. Programmes involving TAs or volunteers can have a valuable impact, but may be less effective than those using experienced and specifically trained teachers. Where tuition is delivered by volunteers or TAs, training and the use of a structured programme is advisable.

Behind the average

Global	Studies undertaken in primary schools tend to show greater impact (about an additional six months' progress) than those undertaken in secondary schools (about an additional four months' progress).
Global	Effects in mathematics appear to be substantially lower (about an additional two months' progress) than in literacy (about an additional six months' progress).
Global	Short, regular sessions (about thirty minutes, three to five times a week) over a set period of time (up to ten weeks) appear to result in optimum impact.
Global	Studies involving digital technology show broadly similar effects.
Local	Studies have been undertaken in seven countries around the world with broadly similar effects.

Closing the disadvantage gap

Local	Studies in England have shown that pupils eligible for free school meals typically receive additional benefits from one-on-one tuition. Low-attaining pupils are particularly likely to benefit. One-on-one tuition approaches can enable pupils to make effective progress by providing intensive, targeted academic support to those identified as having low prior attainment or who are at risk of falling behind. The approach allows the teacher or tutor to focus exclusively on the needs of the learner and provide teaching that is closely matched to each pupil's understanding. One-on-one tuition offers greater levels of interaction and feedback compared to whole class teaching and can support pupils in spending more time on new or unfamiliar material, overcome barriers to learning, and increase their progress through the curriculum.

How could you implement in your setting?

Global	One-on-one tuition provides additional support that is targeted at a pupil's specific needs. Reducing the ratio of pupils to teacher allows for closer interaction between educators and pupils. When adopting one-on-one tuition, schools should consider how to ensure that these active ingredients have a positive impact by: - accurately identifying the pupils who require additional support; - understanding the learning gaps of the pupils who receive tuition and using this knowledge to select curriculum content appropriately; - ensuring that teachers are well prepared for high-quality interactions with pupils, such as providing well-planned feedback; - ensuring that tuition is well linked to classroom content and allowing time for the teacher and tutor to discuss the tuition; - monitoring the impact of tuition on pupil progress and adjusting provision accordingly.
Global	One-on-one tuition may be delivered by teachers, trained TAs, academic mentors or tutors. Interventions are typically delivered over an extended period, often over the course of several weeks or a term.

Local

The average cost of one-on-one tuition is moderate. The costs to schools are largely in additional salary costs and learning resources, the majority of which are recurring costs. Through the UK's National Tutoring Programme Year 1 (2020–21), schools could purchase subsidized in-person or online one-on-one sessions in fifteen-hour blocks for an average cost of £167 to £180 per pupil. Costs are lower for online delivery compared to in-person tuition and are higher when provided by qualified or specialist teachers.

When delivering teacher or TA-led small group tuition, implementation is likely to require a large amount of staff time compared with whole class approaches. Given the lower costs, small group tuition may be a sensible approach to trial before considering one-on-one tuition. See Small group tuition.

In addition to time and cost, school leaders should consider using providers with a track record of effectiveness. To increase the impact of school-led one-on-one tuition, school leaders might consider professional development for teachers, TAs and tutors to support high-quality teaching in areas such as formative assessment, curriculum knowledge, instruction and feedback, which will build capacity in schools.

When introducing new approaches, schools should consider implementation. For more information see Putting Evidence to Work – A School's Guide to Implementation.

How secure is the evidence?

Global

The security of the evidence around one-on-one tuition is rated as moderate. One hundred and twenty-three studies that meet the inclusion criteria for the Toolkit were identified. The topic lost padlocks because:

- a large percentage of the studies were not independently evaluated. Evaluations conducted by organizations connected with the approach, for example, commercial providers, typically have larger impacts, which may influence the overall impact of the strand;

- there is a large amount of unexplained variation between the results included in the topic. All reviews contain some variation in results, which is why it is important to look behind the average. Unexplained variation (or heterogeneity) reduces certainty in the results in ways that we have been unable to test by looking at how context, methodology or approach is influencing impact.

As with any evidence review, the Toolkit summarizes the average impact of approaches as researched in academic studies. It is important to consider context and apply professional judgement when implementing an approach.

Strand title	Peer tutoring
Update date	9 June 2021
Number of studies	127
Summary	Moderate impact for very low cost, based on extensive evidence
Cost	Very low
Padlocks	4
Impact	+5 months

What is it?

Global	Peer tutoring includes a range of approaches in which learners work in pairs or small groups to provide each other with explicit teaching support, such as: - fixed-role, cross-ability tutoring in which one learner, who is often older, takes the tutoring role and is paired with a tutee or tutees, who are often younger; - reciprocal-role tutoring, in which learners alternate between the roles of tutor and tutee. The common characteristic is that learners take on responsibility for aspects of teaching and for evaluating their success.

Key findings

Global	Peer tutoring, on average, has a positive impact on both tutors and tutees and may be a cost-effective approach for delivering one-on-one or small group tuition in a school.
Global	Peer tutoring seems to be most effective when it is used to review or consolidate learning, rather than introduce new material.
Global	Training for staff and tutors is essential for success. It is crucial to allocate sufficient time to train both staff and tutors, to ensure training provides structure to the tutoring, and to identify and implement improvements as the programme progresses.
Global	Four- to ten-week intensive blocks with regular sessions (four to five times a week) appear to provide maximum impact for both tutors and tutees.

How effective is the approach?

Global	Peer tutoring approaches have been shown to have a positive impact on learning, with an average positive effect equivalent to approximately five additional months' progress within one academic year. Studies have identified benefits for both tutors and tutees, and for a wide range of age groups. Though all types of pupils appear to benefit from peer tutoring, there is some evidence that pupils who are low-attaining and those with special educational needs make the biggest gains. Peer tutoring appears to be particularly effective when pupils are provided with support to ensure that the quality of peer interaction is high, for example, questioning frames to use in tutoring sessions, and training and feedback for tutors. In cross-age peer tutoring some studies have found that a gap of less than three years is optimal, although ensuring that the gap is wide enough so that the work is challenging to the tutee whilst easy enough for the tutor to support them is key. Regular tutoring sessions (four to five times a week) of up to ten weeks appear to be more effective than less intensive or longer programmes. Successful approaches may also have other benefits, such as supporting the social and personal development of pupils and boosting their self-confidence and motivation for learning.

Behind the average

Global	Effects are similar (about an additional five months' progress) for both primary and secondary pupils.
Global	Impact is similar (about an additional five months' progress) for both literacy and mathematics.
Global	Lower-attaining pupils tend to benefit more (about an additional six months' progress) than higher-attaining pupils.
Global	A number of studies involving digital technology have been undertaken, with similar overall impact.

Closing the disadvantage gap

Local	While there is limited evidence that specifically examines pupils from disadvantaged backgrounds, studies have shown that pupils who are low-attaining typically receive additional benefits from peer tutoring. Peer-led tutoring approaches may help pupils to close gaps in their learning by offering targeted, peer-led support to consolidate within-class learning, practise skills, and identify and overcome misconceptions. There is also some evidence to suggest that peer-led tutoring can offer tutors the chance to revisit and revise skills and prior knowledge, and develop metacognitive understanding of topics.
Name	Peer tutoring: tutors
Impact	+6 months
Number of studies	12
Summary	In peer tutoring pupils are taught by other pupils, of the same age or sometimes older. This section focuses on the academic impact of delivering tuition on the tutors themselves. Some schools are concerned that the tutors may not benefit and may be losing learning time. However, the impact of peer tutoring on tutors is typically slightly higher than the overall impact of six months' additional progress, on average. The evidence base for this is weaker than the overall evidence for peer tutoring, as only twelve of the 127 studies examined the impact of the approach on tutors.

Applications and Approaches

Name	Reciprocal tutoring
Impact	+5 months
Number of studies	43
Summary	In reciprocal tutoring pupils take turns to be the tutor and the tutee, usually in the same session. Each pupil experiences being taught by a peer and being the tutor. The impact of reciprocal peer tutoring is typically about the same as the overall effect. The average months' progress is five.

How could you implement it in your setting?

Global	Peer tutoring relies on close interaction between two or more students with learners taking responsibility for aspects of teaching and for evaluating their success. When implementing peer tutoring approaches, schools should consider how to ensure high-quality interactions between pupils. This might include: - carefully structuring tasks so sessions focus on existing knowledge; - training peer tutors on teaching approaches, such as modelling knowledge, overcoming common misconceptions, feedback and evaluating progress; - carefully considering appropriate pairing of tutors and tutees; - providing teaching aids and learning frames to guide tutors on how to structure learning, or the types of questions to ask tutees.

Global

Peer tutoring interventions are typically delivered over four- to ten-week intensive blocks. Approaches may involve cross-age or same-age tutoring, usually in pairs. Approaches may be based on a fixed tutee–tutor relationship, while others may be reciprocal.

Local

The average cost of peer tutoring is expected to be very low. The cost to schools is largely in teacher training and learning resources. Implementing peer tutoring also requires a moderate amount of staff time, compared with other approaches.

In addition to time and cost, school leaders should consider how to maximize the quality of peer tutoring interactions and ensure sufficient time is allocated to identify and implement improvements to approaches. When utilizing programmes, school leaders should assess the quality and strength of evidence behind them.

When introducing new approaches, schools should consider implementation. For more information see Putting Evidence to Work – A School's Guide to Implementation.

How secure is the evidence?

Global

The security of the evidence around peer tutoring is rated as high. One hundred and twenty-seven studies that meet the inclusion criteria of the Toolkit were identified. The topic lost a padlock because a large percentage of the studies were not independently evaluated. Evaluations conducted by organizations connected with the approach, for example, commercial providers, typically have larger impacts, which may influence the overall impact of the strand.

As with any evidence review, the Toolkit summarizes the average impact of approaches as researched in academic studies. It is important to consider context and apply professional judgement when implementing an approach.

Strand title	Phonics
Update date	10 June 2021
Number of studies	121
Summary	Moderate impact for very low cost, based on extensive evidence
Cost	Very low
Padlocks	5
Impact	+5 months

What is it?

Global

Phonics is an approach to teaching some aspects of literacy by developing pupils' knowledge and understanding of the relationship between written symbols and sounds. It involves hearing, and identifying and using sound patterns or phonemes to read written language. The aim is to systematically teach pupils the relationship between these sounds and the written spelling patterns, or graphemes, which represent them. Phonics emphasizes the skills of decoding new words by sounding them out and combining or 'blending' the sound-spelling patterns.

3 CHAPTER

Key findings	
Global	Phonics is an important component in the development of early reading skills, particularly for children from disadvantaged backgrounds.
Global	The teaching of phonics should be explicit and systematic to support children in making connections between the sound patterns they hear in words and the way these words are written.
Global	The teaching of phonics should be matched to children’s current level of skill in terms of their phonemic awareness and their knowledge of letter sounds and patterns (graphemes).
Global	Phonics improves the accuracy of a child's reading but not necessarily their comprehension. It is important that children are successful in making progress in all aspects of reading, including comprehension, and the development of vocabulary and spelling, which should also be taught explicitly.

How effective is the approach?	
Global	The average impact of the adoption of phonics approaches is about an additional five months’ progress over the course of a year. Phonics approaches have been consistently found to be effective in supporting younger pupils to master the basics of reading, with an average impact of an additional five months’ progress. Research suggests that phonics is particularly beneficial for younger learners (four- to seven-year-olds) as they begin to read. Teaching phonics is more effective on average than other approaches to early reading (such as whole language or alphabetic approaches), though it should be emphasized that effective phonics techniques are usually embedded in a rich literacy environment for early readers and are only one part of a successful literacy strategy. While there have been fewer studies examining phonics with older readers, there is evidence that it can be a positive approach. With any reading intervention, careful diagnosis of the difficulties the reader is experiencing, regardless of age, is required. If an older reader is struggling with decoding, phonics approaches will still be appropriate. Where readers are struggling with vocabulary or comprehension, other interventions may be more appropriate. There is some variation in impact between different phonological approaches. Synthetic phonics approaches have higher impacts, on average, than analytic approaches. Analytic phonics approaches have also been studied less overall (only nine studies). The small number of analogic phonics approaches identified in this review (six studies) have a negative impact on average.

Behind the average	
Global	The majority of studies have been conducted in primary schools, though there are a number of successful studies with secondary school pupils with a similar overall impact (about an additional five months' progress).
Global	Most studies of phonics are of intensive support in small groups and one-on-one with the aim of supporting pupils to catch up with their peers. The effects of one-on-one tend to be a little higher (about an additional five months' progress) compared with small group interventions (about an additional four months' progress), but this needs to be offset by the number of pupils who can receive support.
Global	Approaches using digital technology tend to be less successful than those led by a teacher or TA. Studies of intensive support involving TAs show slightly lower overall impact (about an additional four months' progress) compared to those involving teachers. This indicates the importance of training and support in phonics for interventions led by TAs.
Global	Synthetic phonics approaches have higher impacts, on average, than analytic phonics approaches.
Local	Studies have been conducted internationally (seven countries), mainly in English-speaking countries. Those conducted outside of the USA have typically shown greater impact.

Closing the disadvantage gap	
Local	Studies in England have shown that pupils eligible for free school meals typically receive similar or slightly greater benefit from phonics interventions and approaches. This is likely due to the explicit nature of the instruction and the intensive support provided. It is possible that some disadvantaged pupils may not develop phonological awareness at the same rate as other pupils, having been exposed to fewer words spoken and books read in the home. Targeted phonics interventions may therefore improve decoding skills more quickly for pupils who have experienced these barriers to learning.

How could you implement it in your setting?	
Global	Phonics approaches aim to quickly develop pupils' word recognition and spelling by developing pupils' ability to hear, identify and manipulate phonemes (the smallest unit of spoken language), and to teach them the relationship between phonemes and the graphemes (written letters or combinations of letters) that represent them. Successfully implementing a phonics might involve: - using a systematic approach that explicitly teaches pupils a comprehensive set of letter–sound relationships through an organized sequence; - training staff to ensure they have the necessary linguistic knowledge and understanding; - carefully monitoring progress to ensure that phonics programmes are responsive and provide extra support where necessary; - carefully considering any adaptions to systematic programmes that might reduce impact. Good implementation of phonics programmes will also consider pupils' wider reading skills and will identify where pupils are struggling with aspects of reading other than decoding that might be targeted through other approaches such as the explicit teaching of reading comprehension strategies.
Global	Where phonics is delivered as an intervention targeted at specific pupils, regular sessions (up to four times a week) of thirty minutes or so over a period of up to twelve weeks appear to be the most successful structure.
Local	Overall, the median costs of implementing a phonics intervention are estimated as very low. The costs associated with teaching phonics arise from the need for specific resources and professional training, the majority of which are initial start-up costs paid during the first year of delivery. Whilst the median cost estimate for phonics programmes is very low, the range of prices between available programmes and the option to purchase additional ongoing training and support for teaching staff means that costs can range from very low to low. Evidence suggests that the effectiveness of phonics is related to the pupil's stage of reading development, so it is important that teachers have professional development in effective assessment as well as in the use of particular phonics techniques and materials. These cost estimates assume that schools are already paying for staff salaries to deliver interventions, facilities to host lessons, and basic stationery materials for staff and pupils. These are all pre-requisite costs of implementing a phonics intervention, without which the cost is likely to be higher. When introducing new approaches, schools should consider implementation. For more information see https://educationendowmentfoundation.org.uk/tools/guidance-reports/a-schools-guide-to-implementation/" Putting Evidence to Work – A School's Guide to Implementation.

How secure is the evidence?

Global

The security of the evidence around phonics is rated as very high. One hundred and twenty-one studies that meet the inclusion criteria of the Toolkit were identified.

As with any evidence review, the Toolkit summarizes the average impact of approaches as researched in academic studies. It is important to consider context and apply professional judgement when implementing an approach.

NOTES – PHONICS THEORY OF CHANGE

Some disadvantaged pupils may have less advanced decoding skills due to lack of exposure to books and print as young children.

Phonics interventions aim to improve decoding skills, so that pupils read words more fluently and therefore comprehend sentences and whole texts.

Strand title	Reading comprehension strategies
Update date	8 June 2021
Number of studies	141
Summary	High impact for very low cost, based on extensive evidence.
Cost	Very low
Padlocks	4
Impact	+6 months

What is it?

Global

Reading comprehension strategies focus on learners’ understanding of written text. Pupils learn a range of techniques which enable them to comprehend the meaning of what they read. These can include: inferring meaning from context; summarizing or identifying key points; using graphic or semantic organizers; developing questioning strategies; and monitoring their own comprehension and then identifying and resolving difficulties for themselves (see also Metacognition and self-regulation).

Strategies are often taught to a class and then practised in pairs or small groups (see also Collaborative learning).

WORKING GROUP 4

3

CHAPTER

Key Findings

Global	Reading comprehension strategies have a high impact on average (about an additional six months' progress). In addition to phonics it is a crucial component of early reading instruction.
Global	It is important to identify the appropriate level of text difficulty, to provide appropriate context to practise the skills, desire to engage with the text and enough of a challenge to improve reading comprehension.
Global	Effective diagnosis of reading difficulties is important in identifying possible solutions, particularly for older struggling readers. Pupils can struggle with decoding words, understanding the structure of the language used, or understanding particular vocabulary, which may be subject-specific.
Global	A wide range of strategies and approaches can be successful, but many pupils need to be taught explicitly and consistently.
Global	It is crucial to support pupils to apply the comprehension strategies independently to other reading tasks, contexts and subjects.

How effective is the approach?

Global	The average impact of reading comprehension strategies is an additional six months' progress over the course of a year. Successful reading comprehension approaches allow activities to be carefully tailored to pupils' reading capabilities, and involve activities and texts that provide an effective, but not overwhelming, challenge. Many of the approaches can be usefully combined with collaborative learning techniques and phonics activities to develop reading skills. Techniques such as graphic organizers and drawing pupils' attention to text features are likely to be particularly useful when reading expository or information texts. There are some indications that approaches involving digital technology can be successful in improving reading comprehension (although there are relatively few studies in this area), particularly when they focus on the application and practice of specific strategies and the use of self-questioning skills. Supporting struggling readers is likely to require a coordinated effort across the curriculum and a combination of approaches that include phonics, reading comprehension and oral language. No particular strategy should be seen as a panacea, and careful diagnosis of the reasons why an individual pupil is struggling should guide the choice of intervention strategies.

Behind the average

Global	More studies have been conducted with primary school pupils, but the teaching of reading comprehension strategies appears effective across both primary (about an additional six months' progress) and secondary (about an additional seven months' progress) schools.
Global	Although the main focus is on reading, comprehension strategies have been successfully used in a number of curriculum subjects where it is important to be able to read and understand text.
Global	Lower-attaining pupils in particular appear to benefit from the explicit teaching of strategies to comprehend text.
Global	There are some indications that approaches involving digital technology can be successful in improving reading comprehension, particularly when they focus on the application and practice of specific strategies and the use of self-questioning skills.
Global	Shorter interventions of up to ten weeks tend to be more successful. However, there are some examples of successful longer interventions.

Closing the disadvantage gap

Local

Studies in England have shown that pupils eligible for free school meals may receive additional benefits from being taught how to use reading comprehension strategies. However, the UK evidence base is less extensive than the global average, and UK studies show lower impact for all pupils.

Reading comprehension strategies involve the teaching of explicit approaches and techniques a pupil can use to improve their comprehension of written text. Many learners will develop these approaches without teacher guidance, adopting the strategies through trial and error as they look to better understand texts that challenge them. However, we know that, on average, disadvantaged children are less likely to own a book of their own and read at home with family members, and for these reasons may not acquire the necessary skills to read and understand challenging texts.

How could you implement it in your setting?

Global

Reading comprehension strategies work through a number of different mechanisms – all focused on improving the understanding of meaning of text effectively. Common elements include:

- explicit teaching of strategies;
- teachers questioning pupils to apply key steps;
- summarizing or identifying key points;
- metacognitive talk to model strategies;
- using graphic or semantic organizers;
- using peer and self-questioning strategies to practise the strategies (such as reciprocal questioning);
- pupils monitoring their own comprehension and identifying difficulties themselves.

Global

Reading comprehension strategy interventions are typically delivered between one to three terms of a school year, either by teachers within class settings, or by TAs with smaller groups.

Evidence suggests that reading comprehension approaches need to be tailored to pupils' current reading capabilities, so it is important that teachers receive professional development in effective diagnosis as well as training in the use of particular techniques and materials.

Local

The average cost of reading comprehension strategies is estimated as very low. The cost to schools is largely in training and professional development, books and learning resources, the majority of which are initial start-up costs paid during the first year of delivery. Whilst the median cost estimate for reading comprehension programmes is very low, the range of prices between available programmes and the option to purchase additional ongoing training and support for teaching staff means that costs can range from very low to low.

Effective teaching of reading comprehension strategies also requires a moderate amount of staff time, compared with other approaches. In addition to time and cost, school leaders should consider how to develop teachers' ability to use specific techniques for particular pupils' needs and ensure they use texts that provide an effective challenge to readers.

When introducing new approaches, schools should consider implementation. For more information see Putting Evidence to Work – A School's Guide to Implementation.

How secure is the evidence?

Global

The security of the evidence around reading comprehension strategies is rated as high. One hundred and forty-one studies that met the inclusion criteria for the Toolkit were identified. The topic lost a padlock because a large percentage of the studies were not independently evaluated. Evaluations conducted by organizations connected with the approach, for example, commercial providers, typically have larger impacts, which may influence the overall impact of the strand.

As with any evidence review, the Toolkit summarizes the average impact of approaches as researched in academic studies. It is important to consider context and apply professional judgement when implementing an approach.

REFERENCES

Borenstein M., Hedges L.V., Higgins, J.P.T. and Rothstein, H.R. (2009) 'Subgroup analyses', in Introduction to meta-analysis. London: John Wiley & Sons, pp. 59-86.

Button, K.S., Ioannidis, J. P., Mokrysz, C., Nosek, B.A., Flint, J., ... and Munafò, M. R. (2013) 'Power failure: why small sample size undermines the reliability of neuroscience', Nature Reviews Neuroscience, 14(5), pp. 365-376.

EEF (2018) Sutton Trust-EEF Teaching and Learning Toolkit & EEF Early Years Toolkit: technical appendix and process manual (working document v.01). London: Education Endowment Foundation.

Higgins, S. (2018) Improving learning: meta-analysis of intervention research in education. Cambridge: Cambridge University Press.

Higgins, S., Katsipataki, M., Villanueva Aguilera, A.B., Dobson, E., ... and Uwimpuhwe, G. (2022) 'The Teaching and Learning Toolkit: communicating research evidence to inform decision-making for policy and practice in education', Review of Education. https://doi.org/10.1002/rev3.3327.

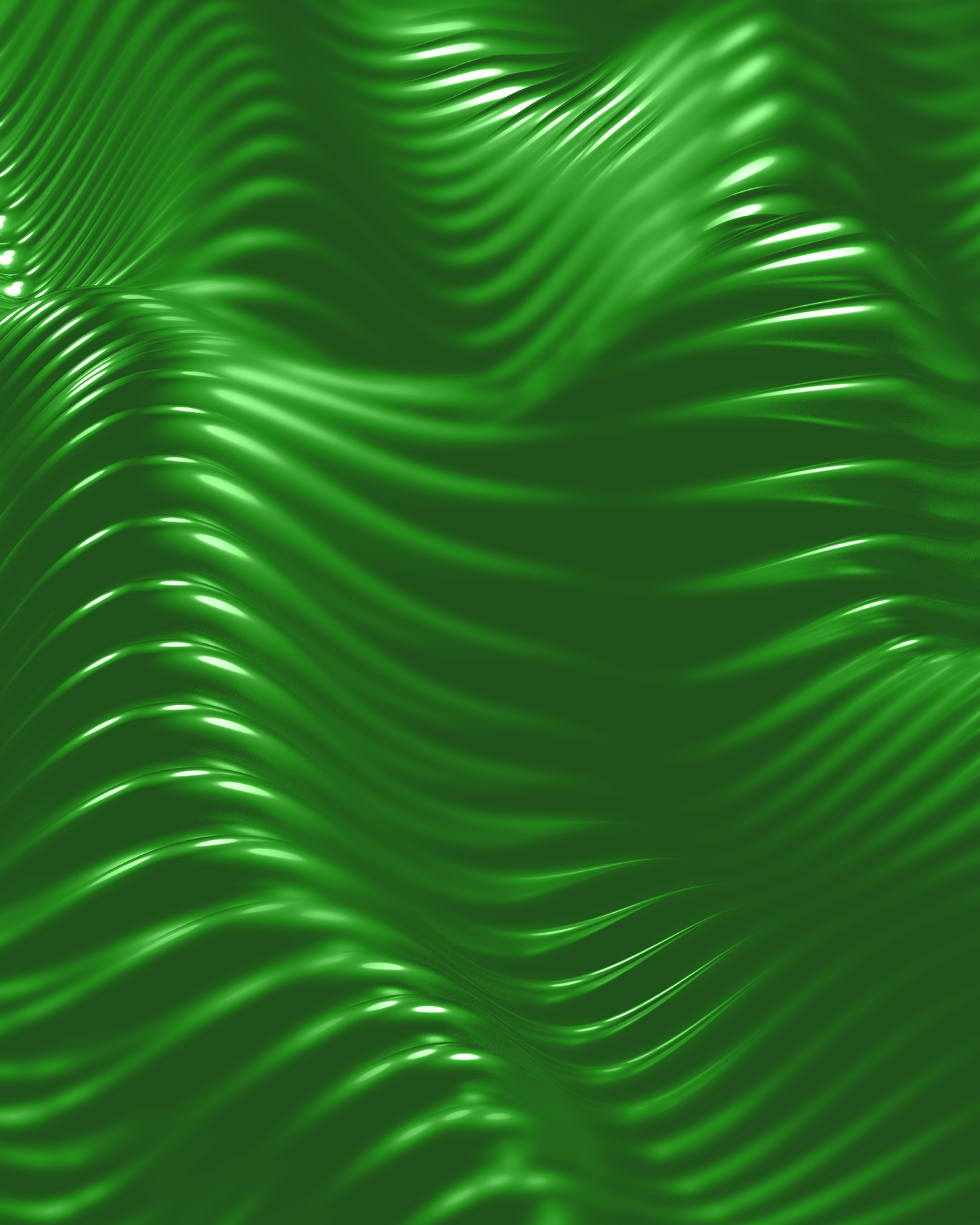
Kühberger, A., Fritz, A. and Scherndl, T. (2014) 'Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size', PLOS ONE, 9(9). https://doi.org/10.1371/journal.pone.0105825.

Methley, A.M., Campbell, S., Chew-Graham, C., McNally, R. and Cheraghi-Sohi, S. (2014) 'PICO, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews', BMC Health Services Research, 14(1). https://doi.org/10.1186/s12913-014-0579-0.

Papaioannou, D., Sutton, A., Carroll, C., Booth, A. and Wong, R. (2010) 'Literature searching for social science systematic reviews: consideration of a range of search techniques', Health Information & Libraries Journal, 27(2), pp. 114–122.

Schlosser, R.W., Wendt, O., Bhavnani, S. and Nail-Chiwetalu, B. (2006) 'Use of information-seeking strategies for developing systematic reviews and engaging in evidence-based practice: the application of traditional and comprehensive pearl growing: a review', International Journal of Language & Communication Disorders, 41(5), pp. 567–582.

Slavin, R. and Smith, D. (2009) 'The relationship between sample sizes and effect sizes in systematic reviews in education', Educational Evaluation and Policy Analysis, 31(4), pp. 500–506.



KEY ACRONYMS

123

3D: Three-Dimensional

ABC

AAC: Augmentative and Alternative Communication

ABI: Acquired Brain Injury

ACC: Anterior Cingulate Cortex

ADHD: Attention Deficit Hyperactivity Disorder

AI: Artificial Intelligence

AIED: Artificial Intelligence in Educational Development

ALE: Activation Likelihood Estimation

ASC: Autism Spectrum Condition

ASD: Autism Spectrum Disorder.

AT: Assistive Technology

BDNF: Brain Derived Neurotrophic Factor

BMI: Body Mass Index

BPEB: Building Performance Evaluation

CA: Canada

CARE: Cultivating Awareness and Resilience in Education

CASEL: Collaborative for Academic, Social, and Emotional Learning

CBTS: Computer Based Tutoring Systems

CCA: Canadian Council for the Arts

CCE: Climate Change Education

CCL: Canadian Council on Learning

CD: Conduct Disorder

CDA: Cognitive Diagnosis Assessment

CNAT: Clasby Neurodiversity Assessment Tool

CPS: Collaborative Problem-Solving

CRPD: Convention on the Rights of Persons with Disabilities.

CSCL: Computer Supported Collaborative Learning

CVT: Control-Value Theory

DEF

DA: Dynamic Assessment

DBCFSN: Detroit Black Community Food Security Network

DESD: Decade of Education for Sustainable Development

DfE: Department for Education

DFID: Department for International Development

DH: Department of Health.

DI: Differentiated Instruction

DNA: Deoxyribonucleic Acid

DSD: Department of Social Development

DSM: Diagnostic and Statistical Manual of Mental Disorders

DSMMD: Diagnostic and Statistical Manual of Mental Disorders

DT: Design Thinking

DTI: Diffusion Tensor Imaging

DWCPD: Department for Women, Children and Persons with Disabilities

EBE: Evidence Based Education

ECCE: Early Childhood Care and Education

ECE: Early Childhood Education

EdTech: Education Technology

EE: Environmental Education

EEF: Education Endowment Foundation

EEG: Electroencephalography

EF: Executive Functions

EFA: Education for All

EFL: English as a Foreign Language

EfS : Education for Sustainability

EI: Education International

EN: Educational Neuroscience

ePEN: Electronic Performance Evaluation Network

ESD: Education for Sustainable Development

ESE: Environmental and Sustainability Education

FEC: Futures of Education Commission

fMRI: functional Magnetic Resonance Imaging

fNIRS: functional Near-Infrared Spectroscopy

GHI

GDP: Gross Domestic Product

GEB: General Ecological Behaviour

GHG: Greenhouse Gas

GIFT: Generalized Intelligent Framework for Tutoring

GIRFEC: Getting It Right For Every Child

GNP: Gross National Product

GPE: Global Partnership for Education

GWAS: Genome-Wide Association Study

HCT: Human Capital Theory

IPCC: Intergovernmental Panel on Climate Change

IPS: Intraparietal Sulcus

IQ: Intelligence Quotient

IRT: Item Response Theory

ISEE Assessment: International Science and Evidence based Education Assessment

ISTE: International Society for Technology in Education

JKL

J-PAL: Abdul Latif Jameel Poverty Action Lab

KBS: Keep Back Straight

LA: Learning Analytics

LAC: Latin American Country

LATAM: Latin America

LGBTQ+: Lesbian, Gay, Bisexual, Transgender, Queer or Questioning

LMICs: Low- and Middle-Income Countries

LTD: Long-Term Depression

LTP: Long-Term Potentiation

LUOTS: Lightning Up the Old Train Station

MNO

MA: Millennium Ecosystem Assessment

MBE: Mind, Brain and Education

MDES: Minimum Detectable Effect Size

MDG: Millennium Development Goal

MEG: Magnetoencephalography

ACRONYMS

MOOC: Massive Open Online Course

MRI: Magnetic Resonance Imaging

MTSS: Multi-Tier Systems of Support

NAPLAN: National Assessment Program – Literacy and Numeracy

NCEE: National College Entrance Exam

NCLB-Act: No Child Left Behind-Act

NCP: Nature's Contribution to People

NEA: National Education Association

NEP: New Ecological Paradigm

NGO: Non-Governmental Organization

NRC: National Research Council

OECD: Organisation for Economic Co-operation and Development

PQRS

PBL: Project Based Learning

PE: Physical Education

PERMA: Positive Emotions, Engagement, (positive) Relationships, Meaning, and Accomplishment

PET: Positron Emission Tomography

PFC: Prefrontal Cortex

PGS: Polygenic Score

PISA: Programme for International Student Assessment

PISA-D: PISA for Development

POC: People of Colour

POE: Post Occupancy Evaluation

PTE: Pearson Test of English

PTSD: Post-Traumatic Stress Disorder

R&D: Research and Development

RAN: Rapid Automatized Naming

RCP: Representative Concentration Pathways

RCT: Randomized Controlled Trial

RD: Reading Disorder

REM: Rapid Eye Movement

ROI: Return on Investment

RtI: Response to Intervention

SCS: Sustainable Community Schools

SDG: Sustainable Development Goal

SDM: Summary for Decision-Makers

SEAL: Social and Emotional Aspects of Learning

SEF: Stage-Environment Fit

SEL: Social and Emotional Learning

SEND: Special Educational Needs and Disabilities

SES: Socio-economic Status
SLD: Specific Learning Disability

SMART: Stress Management and Resiliency Training

SNP: Single Nucleotide Polymorphisms

SOGIE: Sexual Orientation and Gender Identity Expression

STEAM: Science, Technology, Engineering, Arts and Mathematics

STEM: Science, Technology, Engineering, and Mathematics

TUV

TALIS: Teaching and Learning International Survey

TBI: Traumatic Brain Injury

TFI: Teach for India

ToM: Theory of Mind

TPB: Theory of Planned Behaviour

TPJ: Temporoparietal Junction

UDL: Universal Design for Learning

UK (or U.K.): United Kingdom

UKABIF: United Kingdom Acquired Brain Injury Forum

UN: United Nations

UNCRC: United Nations Convention on the Rights of the Child

UNDESA: United Nations Department of Economic and Social Affairs

UNDESD: United Nations Decade of Education for Sustainable Development

UNEP: United Nations Environment Programme

UNESCO: United Nations Educational, Scientific and Cultural Organization

UNESCO MGIEP: UNESCO Mahatma Gandhi Institute of Education for Peace and Sustainable Development

UNFCCC: United Nations Framework Convention on Climate Change

UNICEF: United Nations International Children's Emergency Fund
UNPF: United Nations Population Fund

UNPFA: United Nations Fund for Population Activities

USA: United States of America

USSR: Union of Soviet Socialist Republics

VRU: Violence Reduction Unit

VS: Ventral Striatum

VUCA: Volatile, Uncertain, Complex and Ambiguous.

WXYZ

WEIRD: Western, Educated, Industrialised, Rich and Democratic

WG1: Working Group 1 (of the ISEE Assessment)

WG2: Working Group 2 (of the ISEE Assessment)

WG3: Working Group 3 (of the ISEE Assessment)

WG4: Working Group 4 (of the ISEE Assessment)

WHO: World Health Organization

WSSD: World Summit on Sustainable Development

WWF: World Wide Fund for Nature

ZPD: Zone of Proximal Development

GLOSSARY

WORKING GROUP- 4

ABC

Allocation uncertainty

Allocation uncertainty means that even if a randomly selected sample (e.g. of schools) is randomly assigned to an intervention condition (e.g., experimental or control group), there might be differences between experimental versus control groups that we might not be able to identify.

Causal ascriptions

Causal ascriptions of 'what works' in education are causal relationships between an intervention and outcomes, which arise from a comparison of an experimental group with a control group.

Cognition

Cognition is the mental process involved in knowing, understanding and learning.

Confidence interval

A confidence interval is a range that is often used to measure uncertainty around an estimated value, such as an effect size or the mean of a distribution. This range of values is bounded above and below the statistic's mean. A 95 per cent ‘confidence interval’ includes a range of values for which 95 per cent of the confidence intervals computed from many hypothetical studies would contain the unknown population parameter if all the conditions under which the intervals are built hold.

DEF

Effect size

The effect size is a number that conveys the strength of the relationship between two variable factorst. An example of an effect size often used in intervention research is the difference between the means of two groups, scaled by a measure of how variable or dispersed outcomes are: it divides the mean difference between groups by the standard deviation. In intervention studies, effect size refers to the estimate of impact of an intervention measured (how well did this intervention work?) as the standardized difference in outcomes between the treatment and the comparison group. The larger the effect size, the larger the difference between the two groups and the stronger the association between the intervention and the outcomes being measured.

Evidence (Scientific)

Scientific evidence in the context of applied educational research, is meant to provide limited empirical indications about the efficacy of a given intervention. Scientific evidence comes from rigorous research answering valid research questions. This category of evidence includes, among others, experimental studies, quasi-experimental studies, correlational studies, etc.

GLOSSARY OF KEY TERMS WORKING GROUP- 4

GHI

Grand scientific theory

Grand scientific theories aim at generalizing theoretical understanding.

MNO

Measurement uncertainty

Measurement uncertainty refers to the margin of doubt that exists for the result of any measurement which could be due both to the instrument being used (e.g. a test, a timer) and how this translates the relevant behaviour into a quantitative value (e.g. a score). This means that every measurement differs from the ‘true’ value that it is trying to capture.

Metacognition

Metacognition is 'thinking about thinking' or 'learning to learn' and refers to processes such as monitoring of attention, emotion and behaviour. Students can use metacognitive processes and strategies to monitor and reflect on their own learning.

Mid-range scientific theory

Mid-range (or middle range) scientific theories consist of representations or abstractions of aspects of reality that can be approximated by conceptual models, which can be subjected to empirical tests. This type of theory is detailed enough and ‘close enough to the data’ that testable hypotheses can be derived from it, but abstracted enough to apply to other situations as well.

PQRS

p-value

In intervention research the *p*-values are the probability, under a specified statistical model (the hypothetical scenario), that the sample mean between two groups (i.e., experimental and control group) would be equal to or more extreme than the observed value in the study. Or, when an intervention has no true impact on the population, the *p*-value expresses how likely it would be (in this hypothetical situation) to observe a difference at least as big as the difference they observe due to statistical uncertainty. It answers the question 'how rare would this result be, in a world where the actual result (e.g., actual mean difference between groups) is zero?'.

Permuted p-value

Permuted *p*-values do not rely on assumptions underlying ‘normal’ *p*-values (e.g. randomization, normal data distribution) and therefore do not attempt to make generalizations beyond the sample. Permutation tests work by resampling the observed data many times in order to determine a p-value for the test.

Probative evidence

The probative nature of research results (evidence) refers to the best level of confidence that can be placed in the results of scientific studies aimed at establishing the effectiveness of interventions. This category of evidence includes, for example, mega-analysis, meta-analyses, evidence-based reviews, etc.

Pseudoscientific evidence

Pseudoscientific evidence in the context of applied educational research is a level of evidence and refers to non-scientific 'findings' that 1) pertain to an issue within the domains of science, 2) are not epistemically warranted, but 3) are part of a doctrine creating the impression that it is epistemically warranted. This category of 'evidence' comes from belief, biased observation and absence of research.

Publication bias

Publication bias is a phenomenon in the publication of scientific articles meaning that articles with greater effect sizes or statistical significance are more likely to be published, thereby leaving in the shade articles with mixed or not statistically significant results.

Qualitative research

Qualitative research is research using methods such as participant observation or case studies which result in a narrative, descriptive account of a setting or practice.

Quantitative research

Quantitative research refers to studies that use numerical data to falsify hypotheses in order to develop theories and assumptions. Quantitative research is used to establish generalizable findings.

Randomised controlled trial (RCT)

A Randomised controlled trial is a research design in intervention research which offers (insight into) causal inference.

Relative evidence

Relative evidence for educational intervention effects (i.e., what works?) refers to a level of evidence and comes from thorough comparisons of extant interventions, under the assumption that combined results coming from meta-analyses or systematic reviews are much more informative than single - albeit excellent – studies when necessary precautions are taken. Relative evidence arises from the combined results of multiple studies, using meta-analysis and made possible by thorough comparisons of multiple extant interventions.

Sampling uncertainty

Sampling uncertainty means that even a randomly selected sample (e.g. of schools) might be different from the population at large for reasons we might not be able to identify.

Science

Science is the pursuit and application of knowledge and understanding of the natural and social world following a systematic methodology based on evidence.

Scientific theory

Scientific theories are conceptual models used to explain phenomena.

Statistical significance

Statistical significance is a term used in quantitative research. A result is deemed as 'statistically significant' if the confidence interval does not include zero or if a *p*-value is below a given threshold, often 0.05. In intervention research, when a result is 'statistically significant' it is often interpreted as meaning that the intervention 'had an effect'.

TUV

(Statistical) Uncertainty

Uncertainty in quantitative scientific studies measures how likely it is that the same results that are found in one study are also found in other studies, if replicated under the same conditions. For example, how likely is it that the same experiment, repeated under the same conditions, would find a similar effect? Uncertainty means that we are never sure of the true test-estimate (such as an effect size) observed in a study, for example, in an RCT. It is always possible that the estimate observed in a study will differ from the true estimate in the sample, or in the population. Uncertainty can be statistically expressed as e.g., *p*-value or the amount of variability (i.e. interval) around an estimate. Sources of uncertainty are, among others, sampling uncertainty, allocation uncertainty, and measurement uncertainty.

Validity (external)

External validity refers to the possibility of applying the conclusions of an empirical study outside the context of the study.

Validity (internal)

Internal validity is the extent to which an empirical study establishes and univocally explains a relationship between an intervention and its outcome.

UNESCO MGIEP

unesco

Mahatma Gandhi Institute of
Education for Peace and
Sustainable Development

isee

ASSESSMENT

4 QUALITY EDUCATION

Sustainable Development Goals