

CELEBRATING

100
YEARS

of RCTs in education

100 years of education trials: no significant difference?



@TheNFER @RoyalStatSoc



www.nfer.ac.uk www.rss.org.uk

#EducationRCTs100



NFER

National Foundation for
Educational Research



ROYAL
STATISTICAL
SOCIETY

DATA | EVIDENCE | DECISIONS

The history and future of randomised controlled trials in education



Professor Carole Torgerson

School of Education

Durham University

carole.torgerson@durham.ac.uk

NFER & RSS

100 years of RCTs in education: no significant difference?

23rd September 2019

Improvement and the Distribution of Practice

By

ROBERT ALEXANDER CUMMINS

B.A. (1881), M.A. (1883), Ph.D. (1885)

Head of Department of Education, Simpson College,
Indianapolis, Iowa

Teachers College, Columbia University
Contributions to Education, No. 97

125479
11 | 8 | 20

Published by
Teachers College, Columbia University
NEW YORK CITY
1919



Durham
University

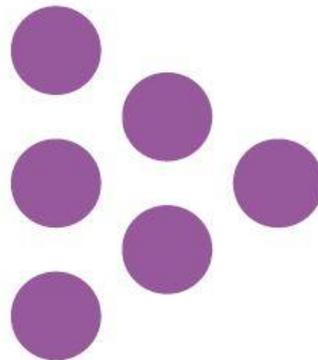


Educational
Research

Volume 60 Issue 3, 2018
ISSN: 0013-1881

**Special Issue: Randomised controlled trials (RCTs) in
education research – methodological debates, questions,
challenges**

Guest Editors: Ben Styles and Carole Torgerson



Contents

EDITORIAL

Randomised controlled trials (RCTs) in education research –
methodological debates, questions, challenges
Ben Styles and Carole Torgerson

ARTICLES

Randomised trials in education in the USA
Larry V. Hedges and Jacob Schauer

The trials of evidence-based practice in education: a systematic review
of randomised controlled trials in education research 1980–2016
Paul Connolly, Ciara Keenan and Karolina Urbanska

Methodological challenges in education RCTs: reflections from
England's Education Endowment Foundation
Anneka Dawson, Emily Yeomans and Elena Rosa Brown

Randomised controlled trials in Scandinavian educational research
*Maiken Pontoppidan, Maria Keilow, Jens Dietrichson, Oddny Judith Solheim,
Vibeke Opheim, Stefan Gustafson and Simon Calmar Andersen*

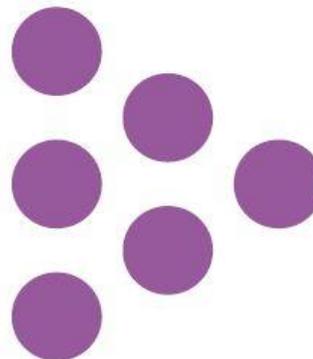
Innovation, evaluation design and typologies of professional learning
Mark Boylan and Sean Demack

The importance of process evaluation for randomised control trials
in education
Nadia Siddiqui, Stephen Gorard and Beng Huat See

BOOK REVIEW

**Special Issue: Randomised controlled trials (RCTs) in
education research – methodological debates, questions,
challenges**

Guest Editors: Ben Styles and Carole Torgerson



Is an intervention using computer software effective in literacy learning? A randomised controlled trial

G Brooks^{a*}, JNV Miles^b, CJ Torgerson^{b*} and DJ Torgerson^b

^a*University of Sheffield, UK;* ^b*University of York, UK*

Background: computer software is widely used to support literacy learning. There are few randomised trials to support its effectiveness. Therefore, there is an urgent need to rigorously evaluate computer software that supports literacy learning.

Methods: we undertook a pragmatic randomised controlled trial among pupils aged 11–12 within a single state comprehensive school in the North of England. The pupils were randomised to receive 10 hours of literacy learning delivered via laptop computers or to act as controls. Both groups received normal literacy learning. A pre-test and two post-tests were given in spelling and literacy. The main pre-defined outcome was improvements in spelling scores.

Results: 155 pupils were randomly allocated, 77 to the ICT group and 78 to control. Four pupils left the school before post-testing and 25 pupils did not have both pre- and post-test data. Therefore, 63 and 67 pupils were included in the main analysis for the ICT and control groups respectively. After adjusting for pre-test scores there was a slight increase in spelling scores, associated with the ICT intervention, but this was not statistically significant (0.954, 95% confidence interval (CI) – 1.83 to 3.74, $p = 0.50$). For reading scores there was a statistically significant decrease associated with the ICT intervention (–2.33, 95% CI –0.96 to –3.71, $p = 0.001$).

Conclusions: we found no evidence of a statistically significant benefit on spelling outcomes using a computer program for literacy learning. For reading there seemed to be a reduction in reading scores associated with the use of the program. All new literacy software needs to be tested in a rigorous trial before it is used routinely in schools.

BRITISH JOURNAL OF EDUCATIONAL STUDIES, ISSN 0007-1005
VOL. 49, No. 3, SEPTEMBER 2001, PP 316-328

THE NEED FOR RANDOMISED CONTROLLED TRIALS IN EDUCATIONAL RESEARCH

by CAROLE J. TORGERSON, *The University of York*, and DAVID J. TORGERSON, *The University of York*

ABSTRACT: This paper argues for more randomised controlled trials in educational research. Educational researchers have largely abandoned the methodology they helped to pioneer. This gold-standard methodology should be more widely used as it is an appropriate and robust research technique. Without subjecting curriculum innovations to a RCT then potentially harmful educational initiatives could be visited upon the nation's children.

Some early milestones in a brief history of RCTs in (education) research

1911: First ‘pragmatic’ trial (quasi-experiment)? – Pearson’s spelling experiment (Torgerson and Torgerson, 2007)

1919: First randomised trials in American education? – Cummings’ practice experiments (Hedges and Schauer, 2018)

[1919: Theory of experimentation and randomisation – Fisher (Stephen Senn, September 16th, 2019)]

1923 McCall’s textbook on the design of educational experiments (Hedges and Schauer, 2018) – focus on ‘matching’

Some early milestones in a brief history of RCTs in (education) research

1931: First known trial in modern period? - Walters' counseling experiment (Forsetland *et al*, 2007); 1932: Walters, replication trial, followed by 6 further trials in 1930s

1940: Lindquist, *Statistical Analysis in Educational Research* : framework for cluster randomisation followed by appropriate analysis of cluster means

[1944: patulin trial - first modern placebo controlled health care trial]

[1948: streptomycin trial]

Pragmatic experimentation in education

“The type of experimentation employed by the trained psychologist in his laboratory is exceedingly useful, but it has its limitations. Its chief defect is that *it isolates from its natural setting the issue to be tested.*” [italics added]

“Educational investigation...should test the efficiency and the economy of a single factor in the teaching process *when surrounded by the normal accompaniments of its classroom situation.*” [italics added]

[H. C. Pearson, 1911]

[Torgerson and Torgerson, 2007]

First known RCT?

The Ohio experiments

“The factor of the teacher was equalized by a *random* selection of the classes which made up the two groups.” [italics added]
(p.51)

“Yet, the *random* method of selecting the pupils would tend to favour one group as much as the other, in so far as this factor [previous training] was concerned.” (p.51)

[R.A. Cummings, 1919]

[Hedges and Schauer, 2018]

First known RCT?

The Ohio experiments

“The *Equal* and *Reducing* groups were made up from the pupils of the seven villages as follows: The *Equal* group included all the classes at Rocky Ridge, Lakeside, and Greenwich, and grades 3, 5 and 7 from Oak Harbor. The *Reducing* group included grades 6 and 8 from Oak Harbor and all the classes at Elmore, Waterville, and Weston.” (p.50)

[R.A. Cummings, 1919]

First known RCT?

The Lyndhurst experiments: pragmatic RCT?

“ In comparing pupils of the same grade, taken at random as ours were, it would seem fair to suppose, for example, that thirty pupils (class c) in the fourth grade in one building would have had as much practice...as had thirty-one pupils (class f) in the fourth grade in another building...” (pp.30-31)

“We have, then, for the *Equal* group, classes ‘e, f, g, h, i’ and ‘j’,...The *Reducing* group includes classes ‘a, b, c, d’ and ‘k’...” (p.35)

[R.A. Cummings, 1919]

First known RCT?

The Lyndhurst experiments

“The records of those who happened to be absent on either of the test days, together with those who were transferred in or out while the experiment was going on, were eliminated because of incompleteness.” (p.12)

“Differences in initial ability, however, may be eliminated by the ‘pairing off’ method, i.e., leaving out the initially better from one group and the initially poorer from the other group, until the *Equal* and *Reducing* groups consist of pupils of average initial ability.” (p.32)

[R.A. Cummings, 1919]

First known RCT?

“Five seniors, each of whom had a good scholarship record, pleasing personality, excellent health and fine social environment, were chosen to act as personnel counselors for the members of the freshman class, who at the end of the first eight weeks of school happened to be delinquent in scholarship in the School of Mechanical Engineering in 1929-30. The 220 delinquent freshmen were divided into two groups by *random sampling*.” [italics added]

[J.E. Walters, 1931]

[Forsetlund *et al*, 2007]

Education trials in 20th Century

1900s to 1940s

- many ‘explanatory’ experiments in educational psychology, probably sometimes using randomisation, more often using ‘matching’
- a few ‘pragmatic’ experiments
- Lindquist’s book describing random sampling and random allocation and the correct analysis for cluster trials (1940)

1960s to 1980

- ‘first flowering’ (Hedges and Schauer, 2018)
- many RCTs in education (US), e.g., HighScope Perry Pre-school Study
- 1963: Campbell and Stanley
- [1967: Schwartz and Lellouch: Explanatory and pragmatic ‘attitudes’ in experimentation]
- 1979: Cook and Campbell
- dearth of high quality RCTs in education (UK), with some exceptions, e.g., i.t.a. trials

Education trials in 20th Century

1980s to 1999

- ‘low point for educational trial in US’ (Hedges and Schauer, 2018)
- dearth of many large scale RCTs in education (US) with notable exceptions: e.g., Tennessee class-size experiment (1985)
- Many ‘explanatory’ experiments in educational psychology
- and dearth of high quality RCTs in education (UK), with some exceptions

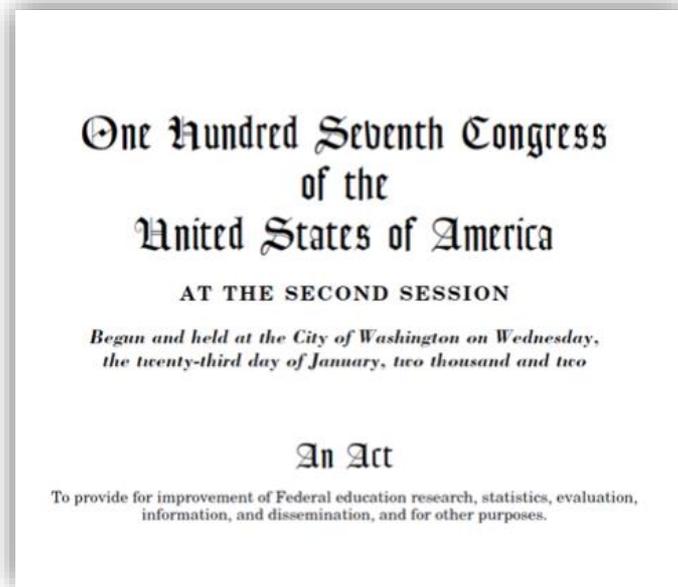
Education trials in 21st Century

2000 - present

- 2002: Shadish Cook and Campbell
- 2002: US legislation to create Institute of Education Sciences (IES)
- 2002 to present: numerous RCTs in education in US including large scale RCTs; pre- post-doctoral training in RCT design and research training for established researchers
- 2009: UK government funding of first RCT evaluation of curriculum intervention (Torgerson et al, 2011)
- 2011: UK government setting up of Education Endowment Foundation (EEF); 152+ RCTs funded by EEF; increasing funding of RCTs by other grant awarding bodies, e.g., ESRC, Nuffield Foundation



US legislation: Institute of Education Sciences



“Scientifically valid educational evaluation employs experimental designs using random assignment, when feasible, and other research methodologies that allow for the strongest possible causal inferences when random assignment is not feasible.” [page 5]
[italics added]



Education Endowment Foundation (EEF)



“...all EEF-funded projects are independently and rigorously evaluated...The impact of projects on attainment will be evaluated where possible, *using randomised controlled trials.*”
[italics added]



CONSORT 2010 checklist of information to include when reporting a randomised trial*

Section/Topic	Item No	Checklist item	Reported on page No
Title and abstract			
	1a	Identification as a randomised trial in the title	_____
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	_____
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale	_____
	2b	Specific objectives or hypotheses	_____
Methods			
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio	_____
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	_____
Participants	4a	Eligibility criteria for participants	_____
	4b	Settings and locations where the data were collected	_____
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	_____
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed	_____
	6b	Any changes to trial outcomes after the trial commenced, with reasons	_____
Sample size	7a	How sample size was determined	_____
	7b	When applicable, explanation of any interim analyses and stopping guidelines	_____
Randomisation:			
Sequence generation	8a	Method used to generate the random allocation sequence	_____
	8b	Type of randomisation; details of any restriction (such as blocking and block size)	_____
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned	_____
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions	_____
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those	_____

		assessing outcomes) and how	_____
	11b	If relevant, description of the similarity of interventions	_____
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes	_____
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses	_____
Results			
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome	_____
	13b	For each group, losses and exclusions after randomisation, together with reasons	_____
Recruitment	14a	Dates defining the periods of recruitment and follow-up	_____
	14b	Why the trial ended or was stopped	_____
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group	_____
Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups	_____
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)	_____
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended	_____
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory	_____
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	_____
Discussion			
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses	_____
Generalisability	21	Generalisability (external validity, applicability) of the trial findings	_____
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	_____
Other information			
Registration	23	Registration number and name of trial registry	_____
Protocol	24	Where the full trial protocol can be accessed, if available	_____
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	_____

*We strongly recommend reading this statement in conjunction with the CONSORT 2010 Explanation and Elaboration for important clarifications on all the items. If relevant, we also recommend reading CONSORT extensions for cluster randomised trials, non-inferiority and equivalence trials, non-pharmacological treatments, herbal interventions, and pragmatic trials. Additional extensions are forthcoming: for those and for up to date references relevant to this checklist, see www.consort-statement.org.

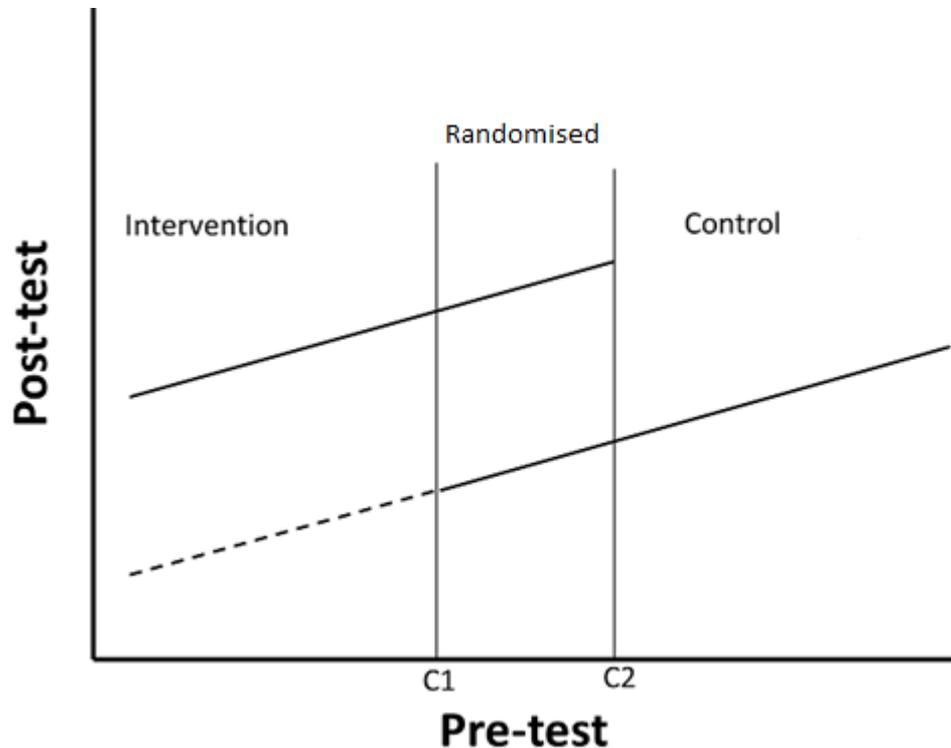
Randomisation: Challenges

Encouraging randomisation conducted and reported to
CONSORT standards

Explaining randomisation to facilitate acceptability

Ensuring strict adherence to randomisation

EEF-funded *SHINE in Secondaries* evaluation: Regression discontinuity design (RDD)



RDD with two cut points & tie-breaker randomisation

[Shadish, W. R., Cook, T. D., & Campbell, D. T., 2002]

Below C1 (1st cut point) all students are invited to the intervention

Above C2 (2nd cut point) no students are invited to intervention

Between C1 & C2 students are randomised to receive intervention or control

[Menzies *et al*, 2015]

Acknowledgements and thanks

***Educational Research Journal* Volume 60 Issue 3, 2018**

Editor Special Issue: Frances Brill NFER

Authors: Larry Hedges and Jacob Schauer for uncovering the first RCT (Cummings, 1919)

Authors, peer reviewers, members of editorial board and panel

Published RCTs and methodological papers

All my co-authors and collaborators

Thanks for listening!

References

- Campbell, D.T. and Stanley, J.C. (1966) *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
- Cook, T.D. and Campbell, D.T. (1979) *Quasi-experimentation: Design and analysis for field settings*. MA: Houghton-Mifflin
- Cummings, R.A. *Improvement and the distribution of practice*. New York: Teachers College, Columbia University.
- Forsetlund, L., Chalmers, I. and Bjorndal, A. (2007) When was random allocation first used to generate comparison groups in experiments to assess the effects of social interventions? *Economics of Innovation and New Technology*, 16(5 & 6).
- Hedges, L. and Schauer, J., Randomised trials in education in the USA, in Styles, B. and Torgerson, C. (2018) Randomised controlled trials (RCTs) in education research – methodological debates, questions, challenges, *Educational Research*, 60(3).
- Lindquist, E.F., (1940) *Statistical Analysis in Educational Research*. Boston: Houghton-Mifflin.
- McCall, W.A. (1923) *How to Experiment in Education*. New York: Macmillan.
- Pearson, H.C. (1911) The Scientific Study of the Teaching of Spelling. *Journal of Educational Psychology*, 2, 241–252.

References (cont.)

- Schwartz, D. and Lellouch, J. (1967) Explanatory and pragmatic attitudes in therapeutic trials, *Journal of Chronic Diseases* 20:636-48
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Styles, B. and Torgerson, C. (2018) Special Issue: Randomised controlled trials (RCTs) in education research – methodological debates, questions, challenges, *Educational Research*, 60(3).
- Torgerson, C.J. & Torgerson, D.J. (2001). The Need for Randomised Controlled Trials in Educational Research. *British Journal of Educational Studies* 49(3): 316-328.
- Torgerson, C.J., Wiggins, A., Torgerson, D.J., Ainsworth, H., Hewitt, C. (2013) *Every Child Counts: Testing policy effectiveness using a RCT, designed, conducted and reported to CONSORT standards*, *Journal of Research in Mathematics Education*
- Torgerson, C.J. and Torgerson, D.J. (2007) The need for pragmatic experimentation in educational research, *Economics of Innovation and New Technology*, 16(5 & 6)
- Walters, J.E. (1931) Seniors as Counselors, *The Journal of Higher Education*: 2(8)
- Walters, J.E. (1932) Measuring Effectiveness of Personnel Counseling, *Personnel Journal*: 11(4)



QUEEN'S
UNIVERSITY
BELFAST

CENTRE FOR
EVIDENCE AND
SOCIAL INNOVATION

The Use of RCTs in Education: Opportunities and Challenges

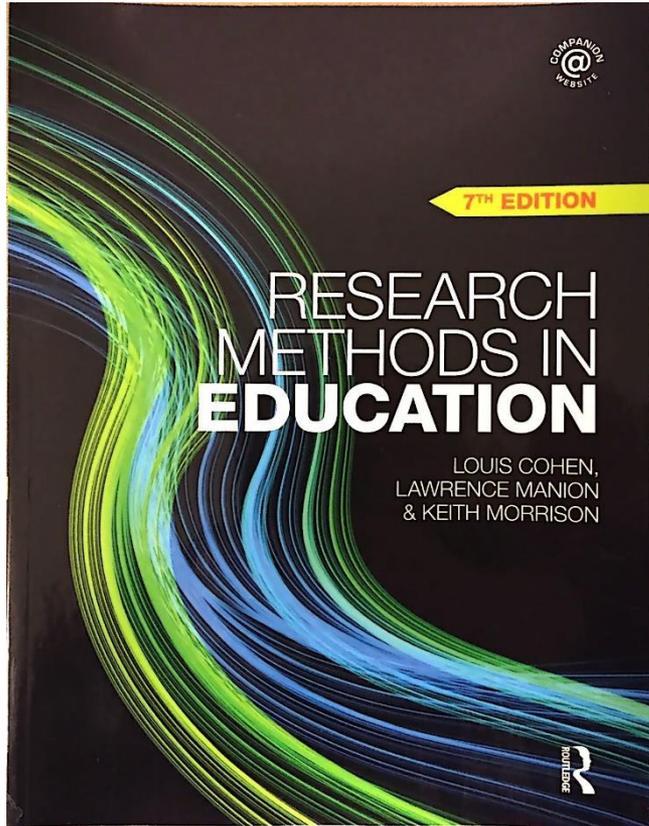
Professor Paul Connolly

Centre for Evidence and Social Innovation
Queen's University Belfast

“100 years of RCTs in education: No significant difference?”, Symposium hosted by the Royal Statistical Society and the National Foundation for Educational Research

Monday 23 September 2019, Royal Statistical Society, London

RCTs and the education research community



“This model [the RCT], premised on notions of isolation and control of variables in order to establish causality, may be appropriate for a laboratory, though whether, in fact, a social situation either ever could become the antiseptic, artificial world of the laboratory or should become such a world is both an empirical and a moral question respectively. Further, the ethical dilemmas of treating humans as manipulable, controllable and inanimate are considerable”

“Randomised controlled trials belong to a discredited view of science as positivism”

(Cohen, Manion & Morrison, 2011: p. 314)



QUEEN'S
UNIVERSITY
BELFAST

CENTRE FOR
EVIDENCE AND
SOCIAL INNOVATION

The paradigm wars

Practitioner research	“What works” research
For practitioners	Against practitioners
Encourages professional autonomy	Undermines professional autonomy
Small-scale, action research	Large-scale, surveys and RCTs
Qualitative	Quantitative
Emancipatory	Oppressive
Democratic	Dictatorial
Theoretically-informed	Descriptive and theoretically naïve
Encourages reflective practice	Stifles reflective practice



Four key criticisms

1. It is just not possible to do RCTs in education
2. RCTs ignore context and experience
3. RCTs seek to generate universal laws of 'cause and effect'
4. RCTs are inherently descriptive and contribute little to theory



**QUEEN'S
UNIVERSITY
BELFAST**

**CENTRE FOR
EVIDENCE AND
SOCIAL INNOVATION**

And what is the actual evidence?

- Systematic search and review of all RCTs undertaken in education since 1980
- Studies included only if:
 1. the study design involved the random allocation of subjects (either individually or as groups) to a control group and at least one intervention group
 2. the intervention was undertaken in and with the involvement of an educational institution:
 - preschool/kindergarten
 - primary/elementary
 - secondary/middle/high
 - college/university
 3. the intervention focused on improving at least one educational outcome (i.e. relating to the acquisition of knowledge and/or skills)

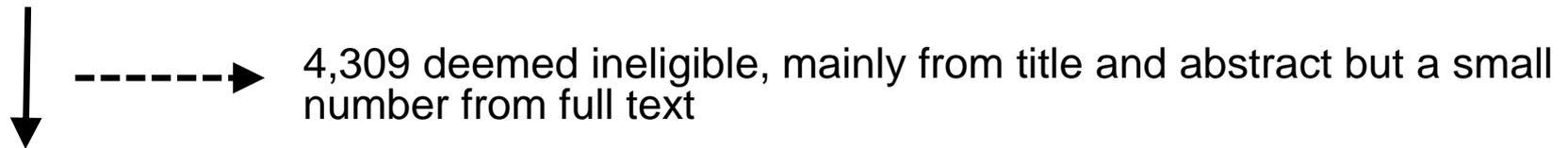


Studies retrieved

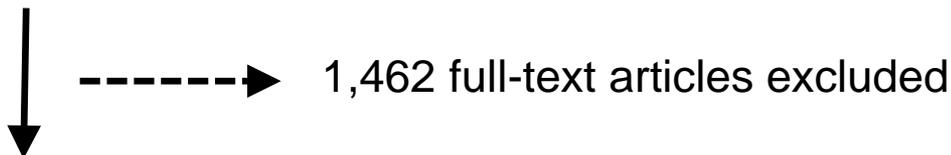
10,286 studies found in total (8,172 from databased; 2,114 from grey literature)



6,788 abstracts screened



2,479 full-text articles assessed for eligibility



1,017 unique studies included in the analysis

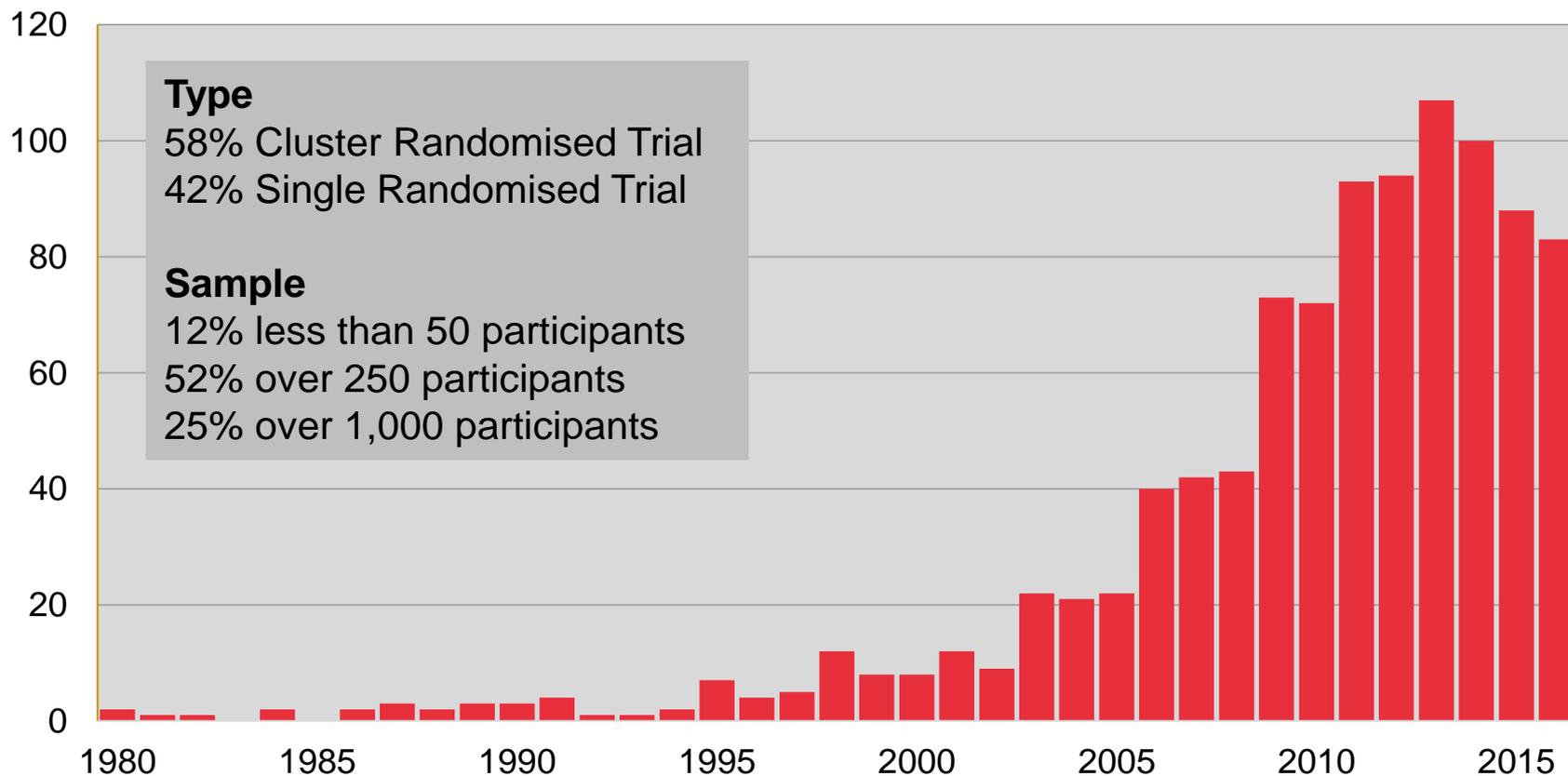


QUEEN'S
UNIVERSITY
BELFAST

CENTRE FOR
EVIDENCE AND
SOCIAL INNOVATION

Key findings

RCTs in Education Published Between 1980 - 2015

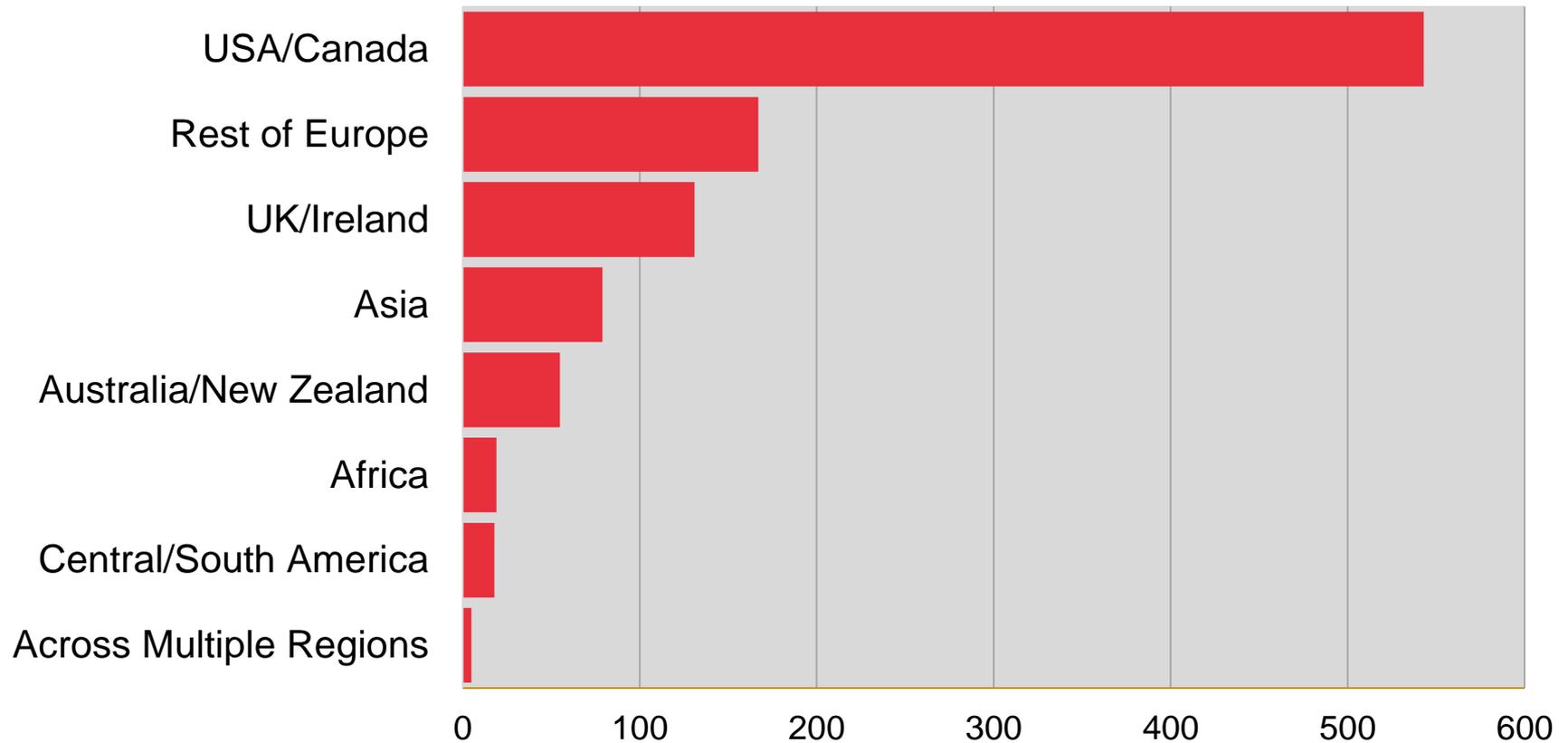


QUEEN'S
UNIVERSITY
BELFAST

CENTRE FOR
EVIDENCE AND
SOCIAL INNOVATION

Key findings

Location of RCTs Published Between 1980 - 2015

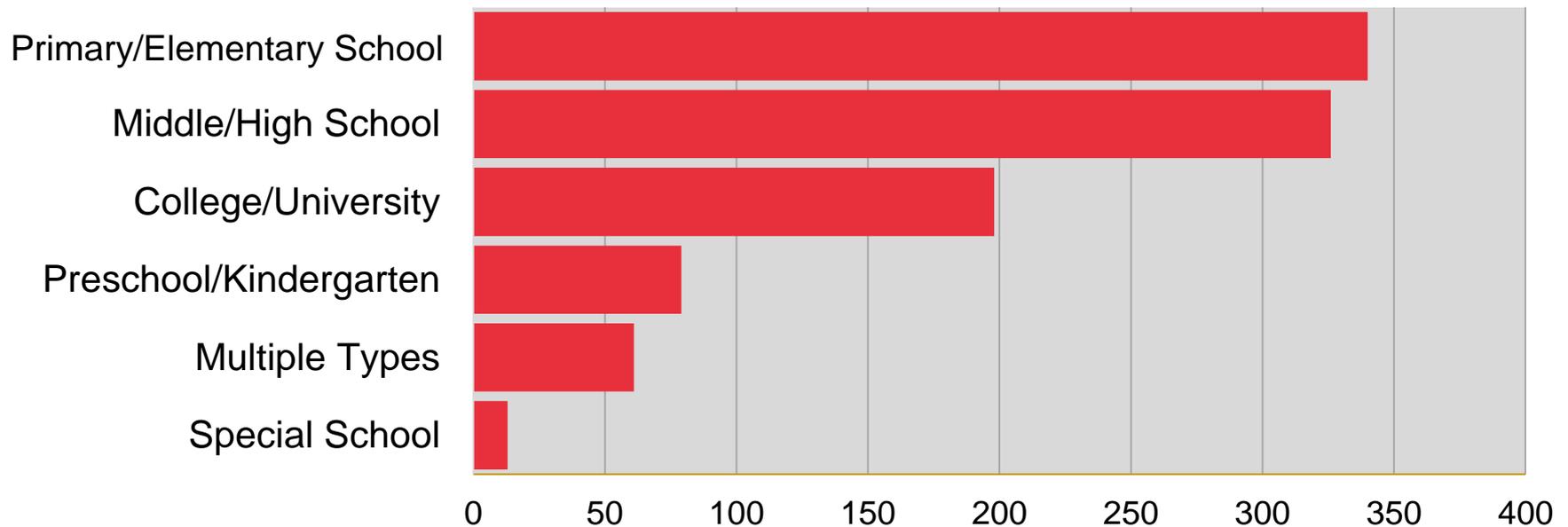


**QUEEN'S
UNIVERSITY
BELFAST**

**CENTRE FOR
EVIDENCE AND
SOCIAL INNOVATION**

Key findings

Type of Educational Institutions Providing the Focus for RCTs Published Between 1980 - 2015

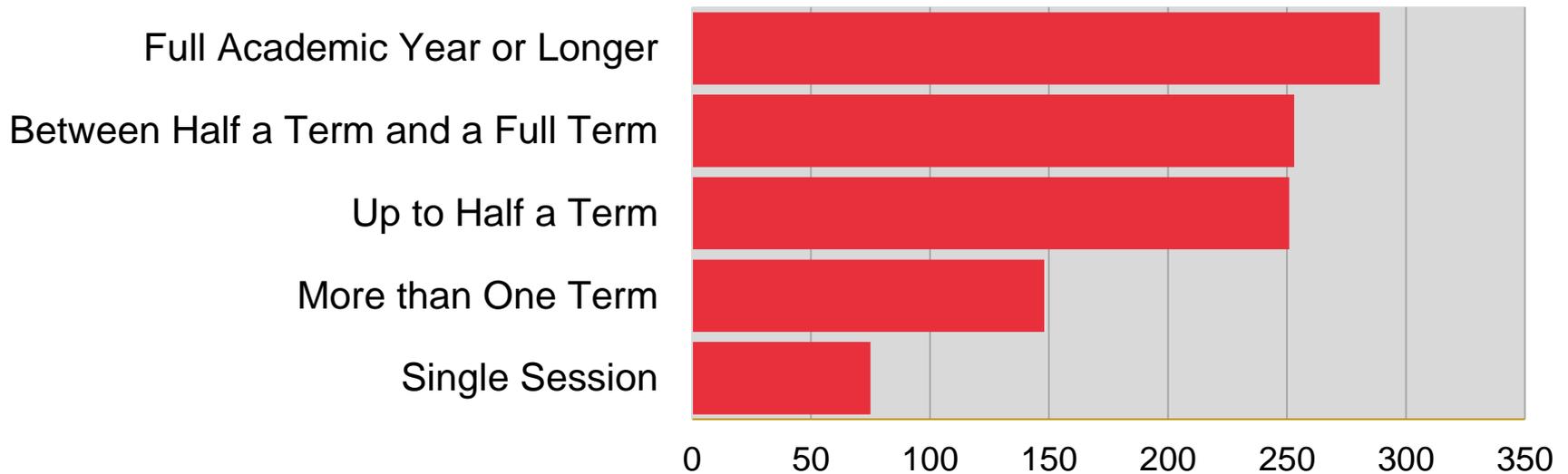


**QUEEN'S
UNIVERSITY
BELFAST**

**CENTRE FOR
EVIDENCE AND
SOCIAL INNOVATION**

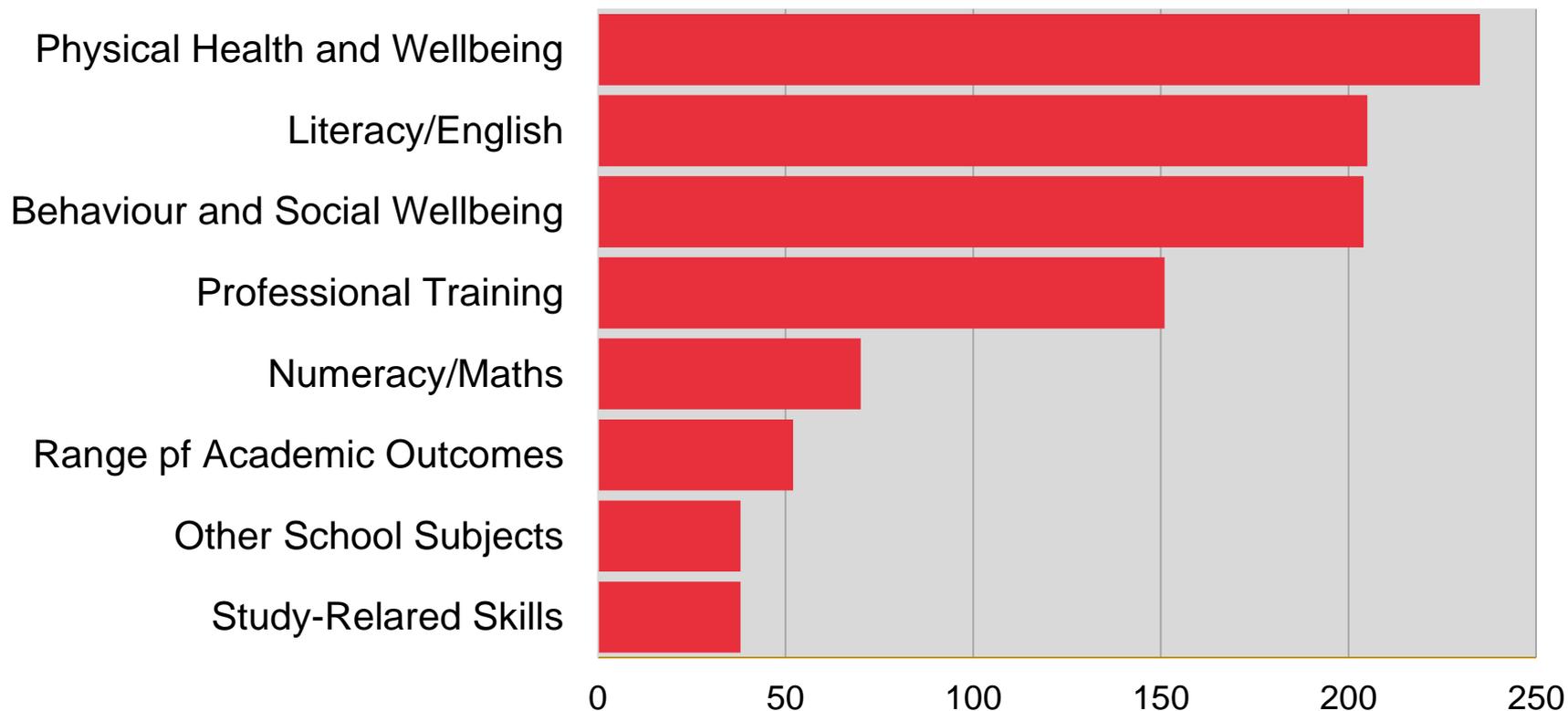
Key findings

Duration of Interventions Providing the Focus for RCTs Published Between 1980 - 2015



Key findings

Primary Outcomes of RCTs Published Between 1980 - 2015



QUEEN'S
UNIVERSITY
BELFAST

CENTRE FOR
EVIDENCE AND
SOCIAL INNOVATION

Back to the four key criticisms

1. Is it possible to do RCTs in education?

- Yes, see above!

2. Do RCTs ignore context and experience?

- 49% reported some sub-group analysis
- 31% reported some qualitative findings (further 7% gathered qualitative data but did not report it in the publication)

3. Do RCTs seek simply to generate universal laws of ‘cause and effect’?

- See above re: sub-group analyses
- 78% included some discussion regarding the limits to generalisability

4. Are RCTs inherently descriptive and atheoretical?

- 35% included explicit discussion of specific theorists/theories
- A further 43% included discussion of a descriptive theory of change
- But also, note the potential of systematic reviews and meta-analysis of trials to advance theory



Conclusions

- The use of RCTs in education is growing rapidly
- Whilst it is a developing field of research, there is already clear evidence that significant progress is being made to address the criticisms levelled at RCTs
- RCTs are quite capable of moving beyond the question of ‘what works?’ to ‘what works, for whom and in what contexts and under what circumstances?’
- RCTs, and especially the synthesis and meta-analysis of findings from RCTs, have significant potential to contribute to theory testing and development



QUEEN'S
UNIVERSITY
BELFAST

CENTRE FOR
EVIDENCE AND
SOCIAL INNOVATION

Future challenges

1. Urgent need to increase awareness and understanding of RCTs amongst the wider education research community
2. More RCTs should be encouraged to include qualitative components and to engage explicitly with theory
3. More education researchers should be encouraged to bring their subject and methodological expertise to RCTs through multi-method research designs
4. Need to further develop collaborative approaches to RCTs with teachers and schools and also other key stakeholders (children and young people, parents, policy makers)
5. Much more investment is needed in systematic reviews and the development of meta analytic techniques



QUEEN'S
UNIVERSITY
BELFAST

CENTRE FOR
EVIDENCE AND
SOCIAL INNOVATION

More information

Article

Connolly, P., Keenan, C. and Urbanska. K. (2018) The trials of evidence-based practice in education: a systematic review of randomized controlled trials in education research 1980-2016, Educational Research, 60:3, pp. 276-291.

Full text available, open access:

<https://doi.org/10.1080/00131881.2018.1493353>

Contact

paul.connolly@qub.ac.uk



QUEEN'S
UNIVERSITY
BELFAST

CENTRE FOR
EVIDENCE AND
SOCIAL INNOVATION

Are rigorous educational trials producing useful evidence?

Hugo Lortie-Forgues



UNIVERSITY
of York

In collaboration with Matthew Inglis



**Loughborough
University**

RCTs in Education

- Growing number of Randomized Control Trials (RCTs) in the last 10 years.

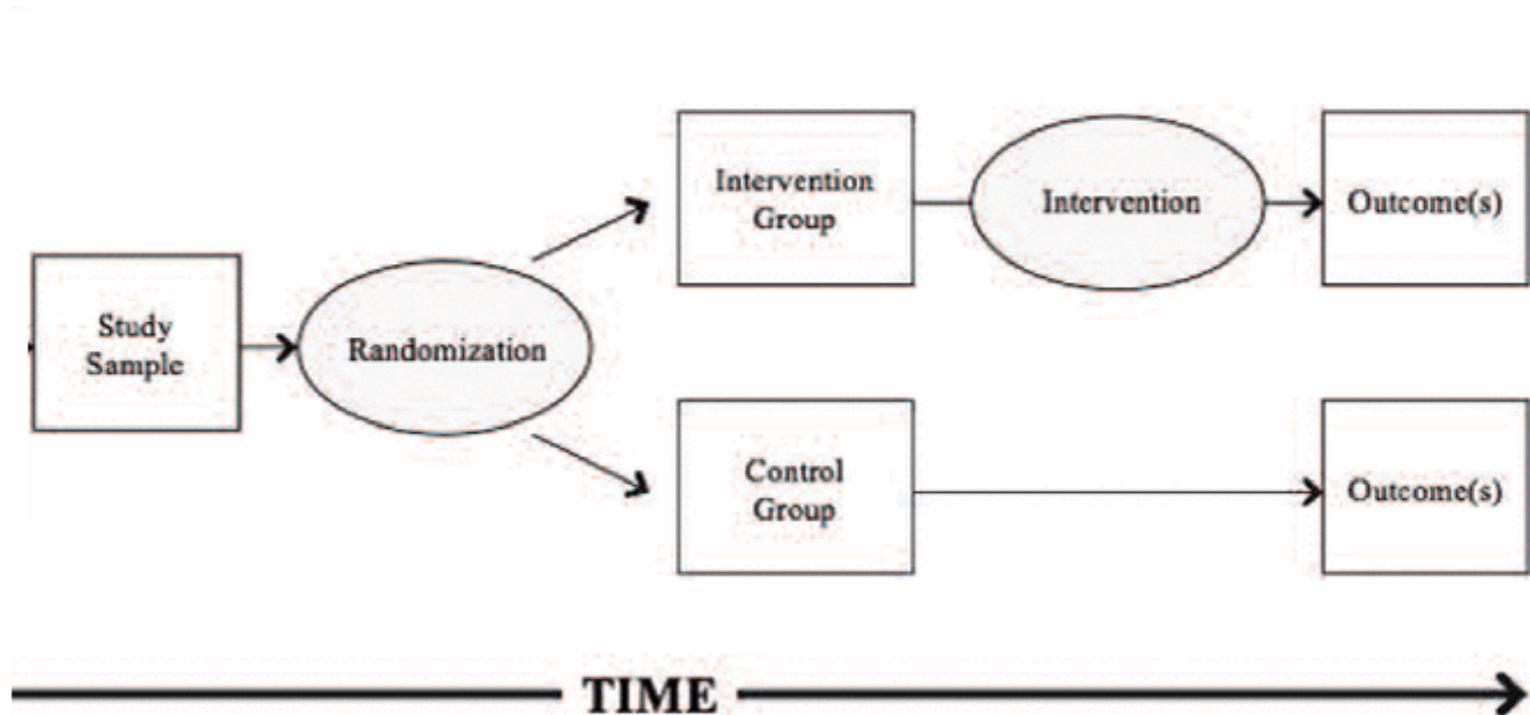


RCTs in Education

- Growing number of Randomized Control Trials (RCTs) in the last 10 years.
- Usually very expensive (\approx £500,000 each RCT)
- Important to reflect on how successful this dramatic change of approach has been.

Effect Size

- Effect size: standardized measure of the magnitude of a phenomenon.



Effect Size

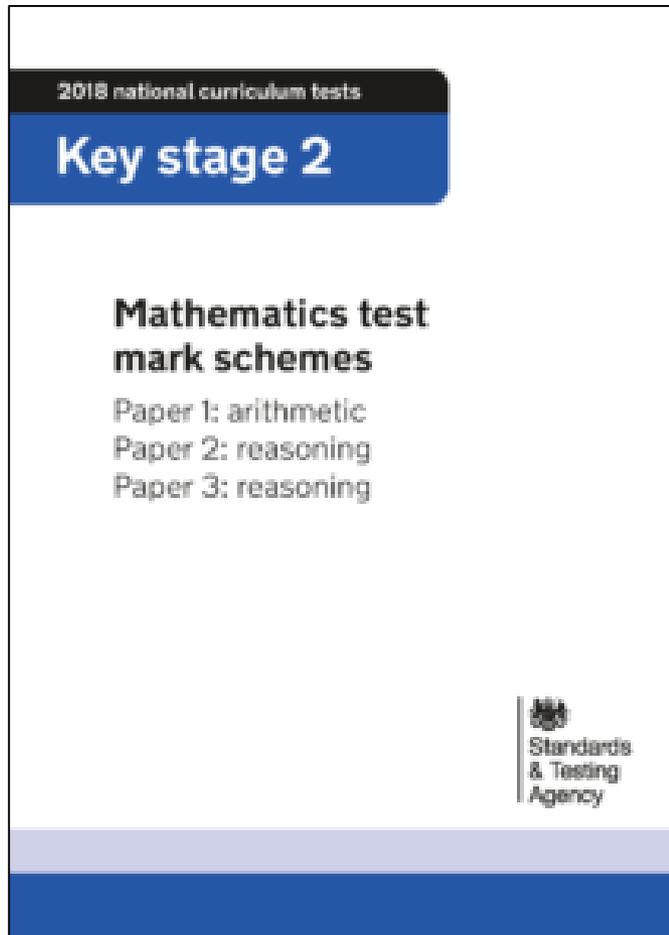
Benchmarks in social sciences

0.2 small effect size

0.5 medium effect size

0.8 large effect size

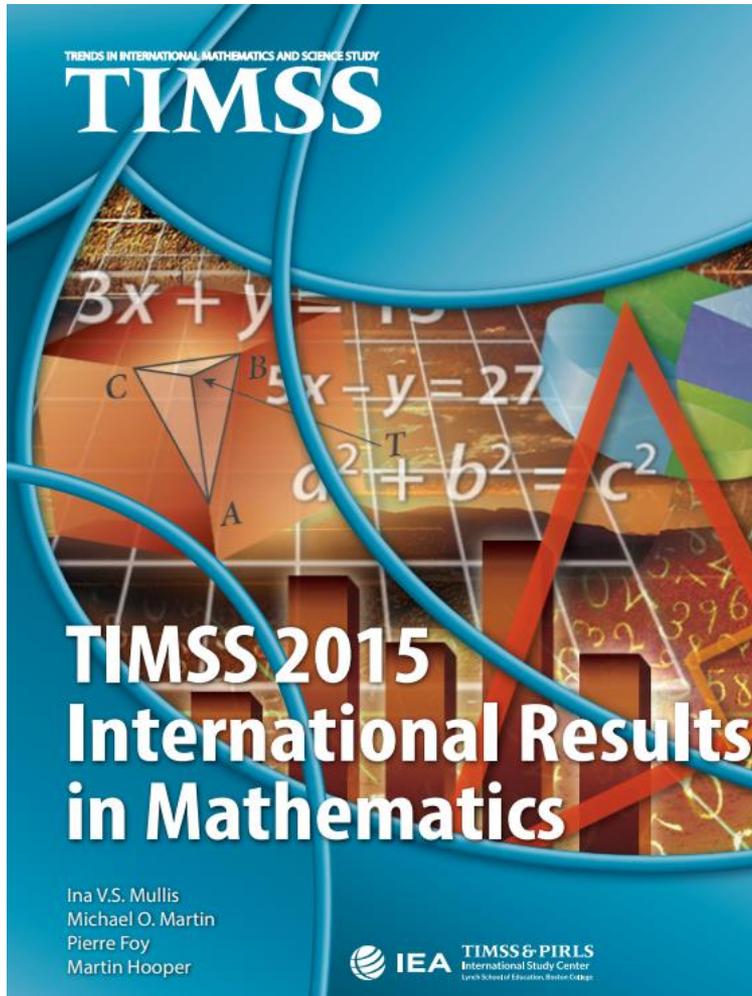
Effect Size



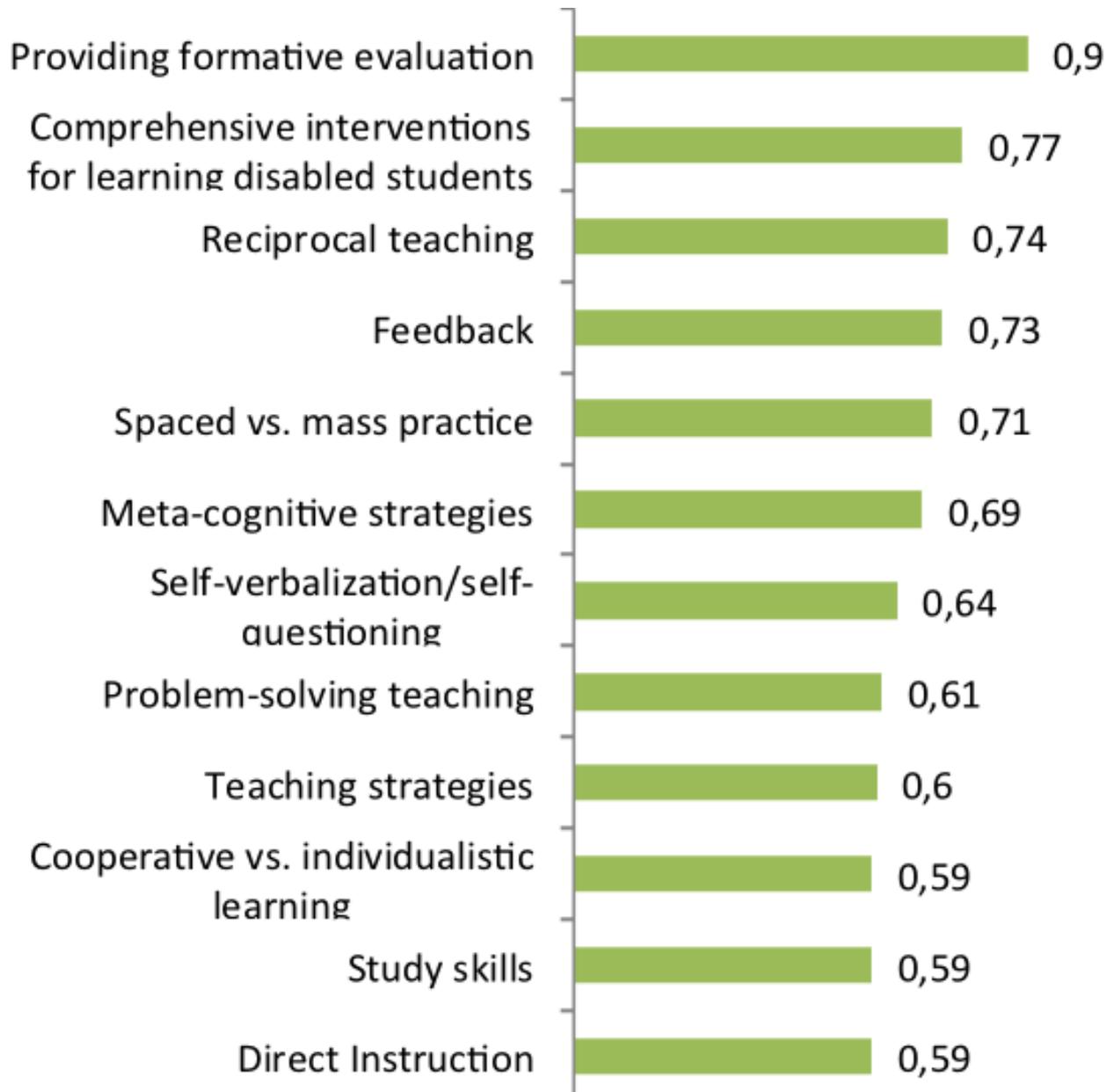
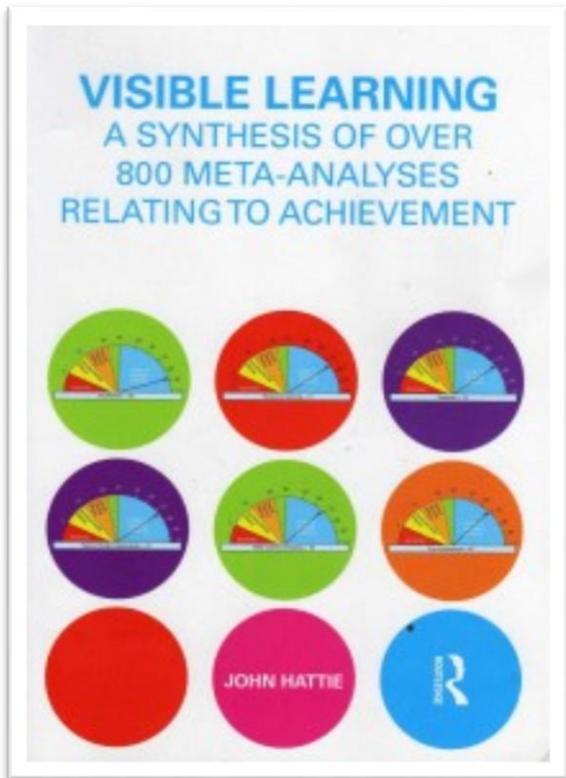
Scores range from 80 to 120.

Increasing performance by **1 point** corresponds to an effect size of **0.14**.

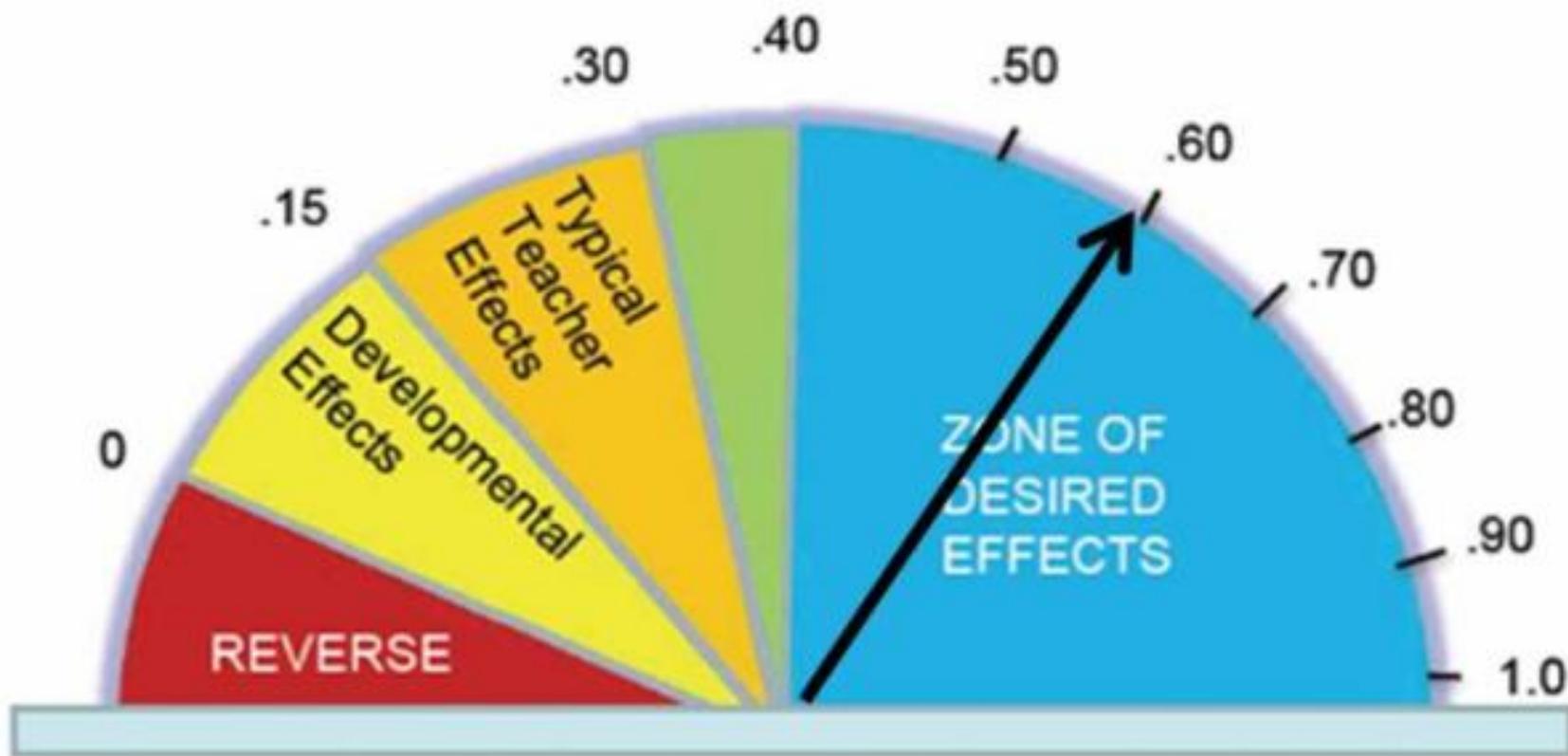
Effect Size



The difference in mathematical ability between the pupils in Singapore and in England corresponds to an effect size of **0.84**.



Visible Learning (John Hattie)



Effect sizes likely to be small in RCTs

- Large, heterogenous population
- Active control group
- Outcome measures are standardized tests
- Pre-registered measures and analyses
 - No p-hacking
- Findings published regardless of outcome
 - No Publication bias

How large are the effect sizes observed in rigorous educational RCTs?

The Education Endowment Foundation (EEF)

- Started in 2011
- Conducts “rigorous” RCTs
 - Majority of trials have > 500 participants
 - In multiple schools (on average, 44 different schools)
 - Relatively long interventions
 - Mean grant: £500,000

EEF RCTs

- 82 RCTs
- 140 Distinct Effect Sizes
- 790,279 students
- 37 million pounds

School Year	Nbr
Kindergarten	5
Elementary	86
Secondary	36
Elem and Secondary	13

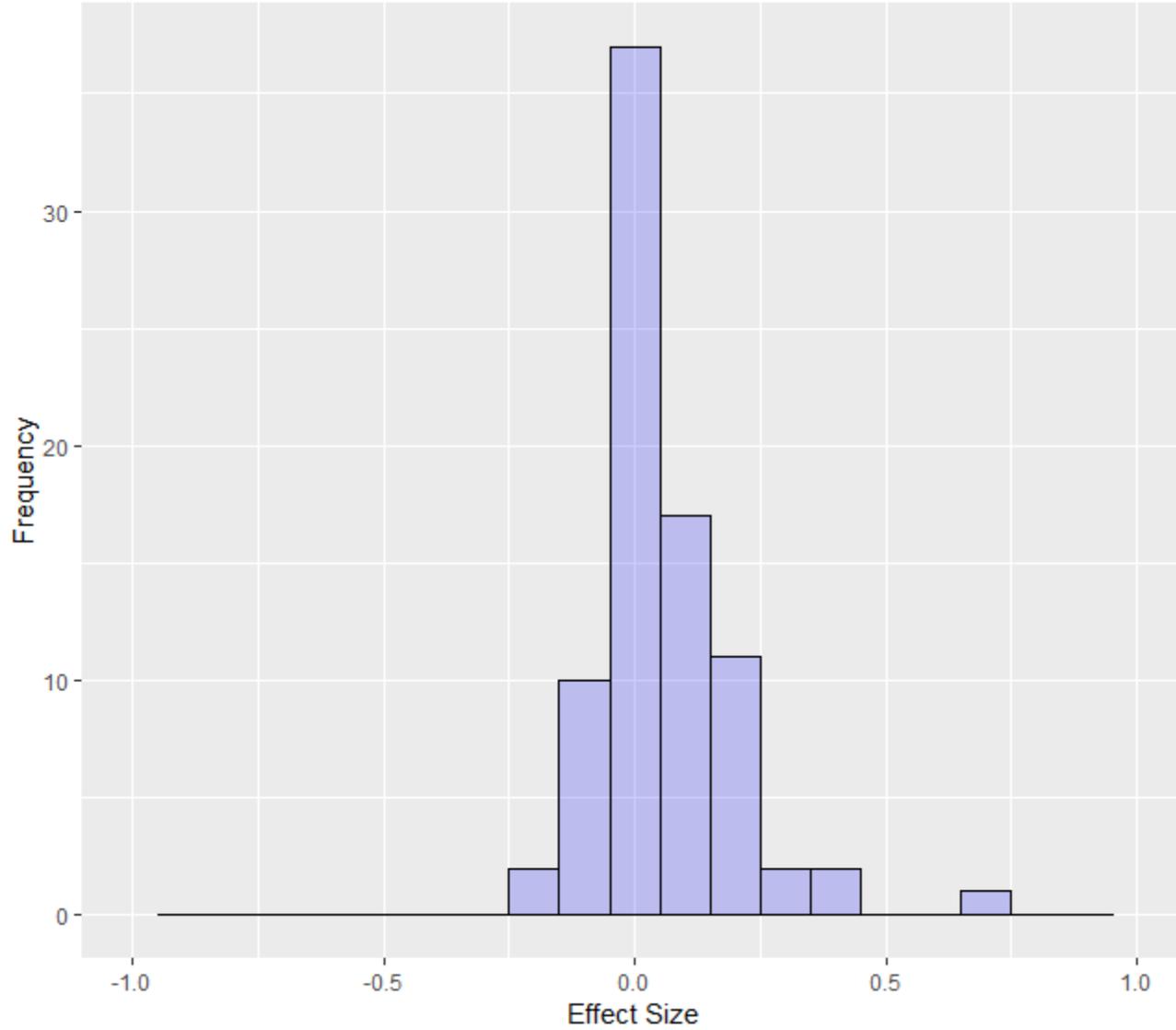
Topic	Nbr
Language: Reading	63
Mathematics	35
Language: General	20
Language: Writing	8
Sciences	4
Combination of topics	10

EEF RCTs

- Average effect size?

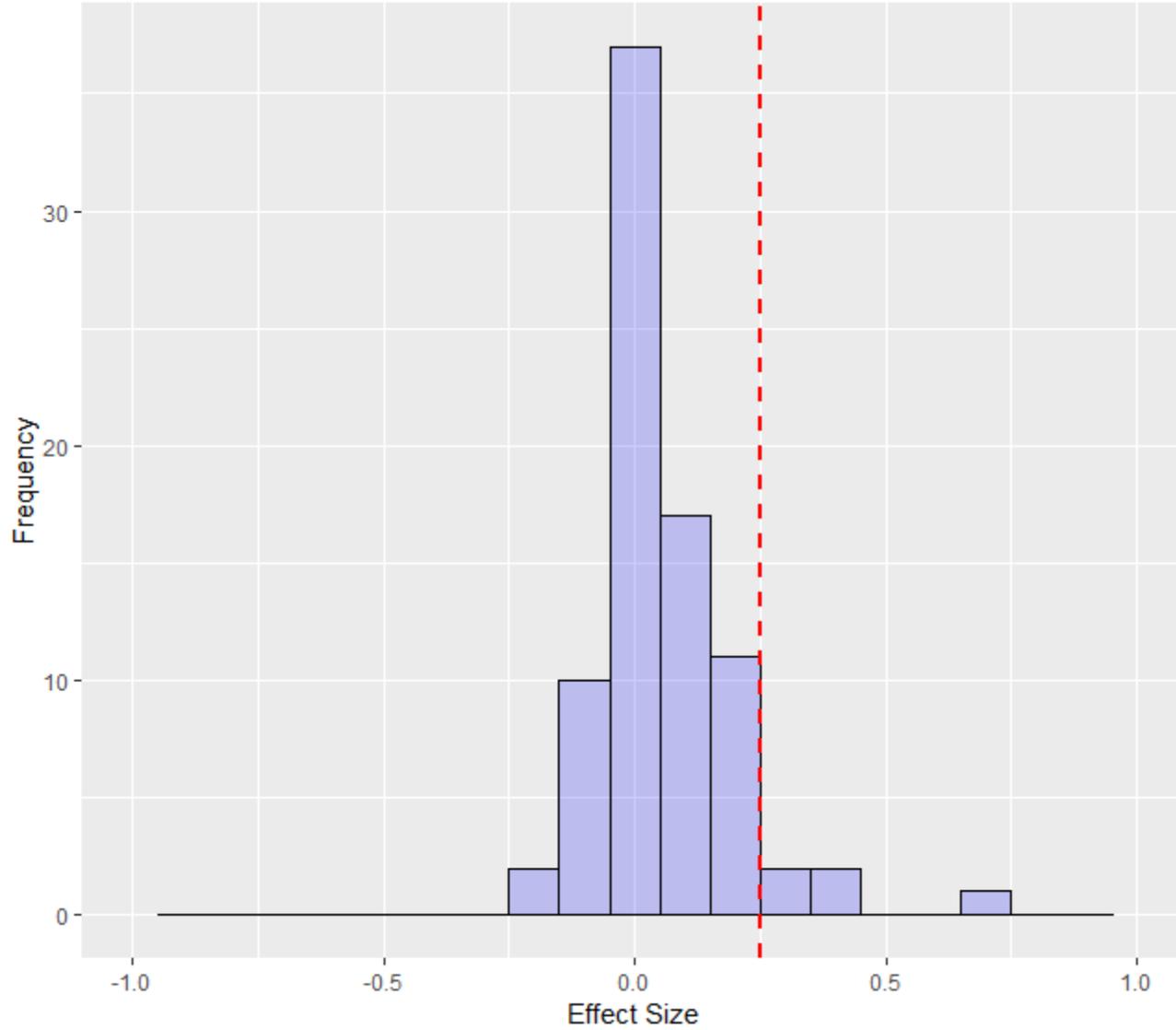
0.06

Effect Sizes from the EEF (UK)



Effect Sizes from the EEF (UK)

0.4



Factors influencing effect sizes?

Age of the participants?

Topic of the intervention?

Cost of the intervention?

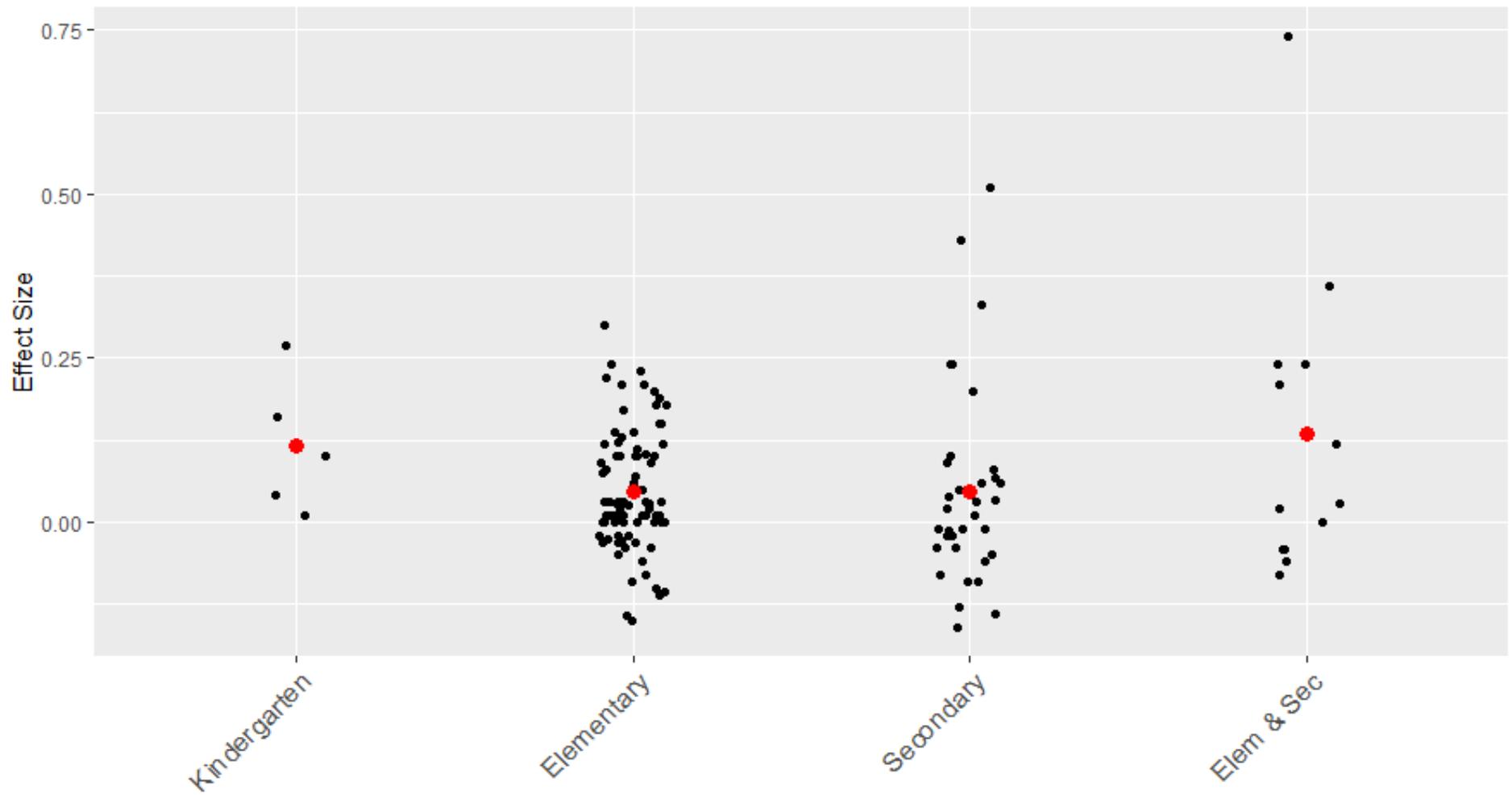
Size of the trial?

Year the trial was conducted?

Effect Sizes by Age

School Year	Nbr
Kindergarten	5
Elementary	86
Secondary	36
Elem and Secondary	13

Age



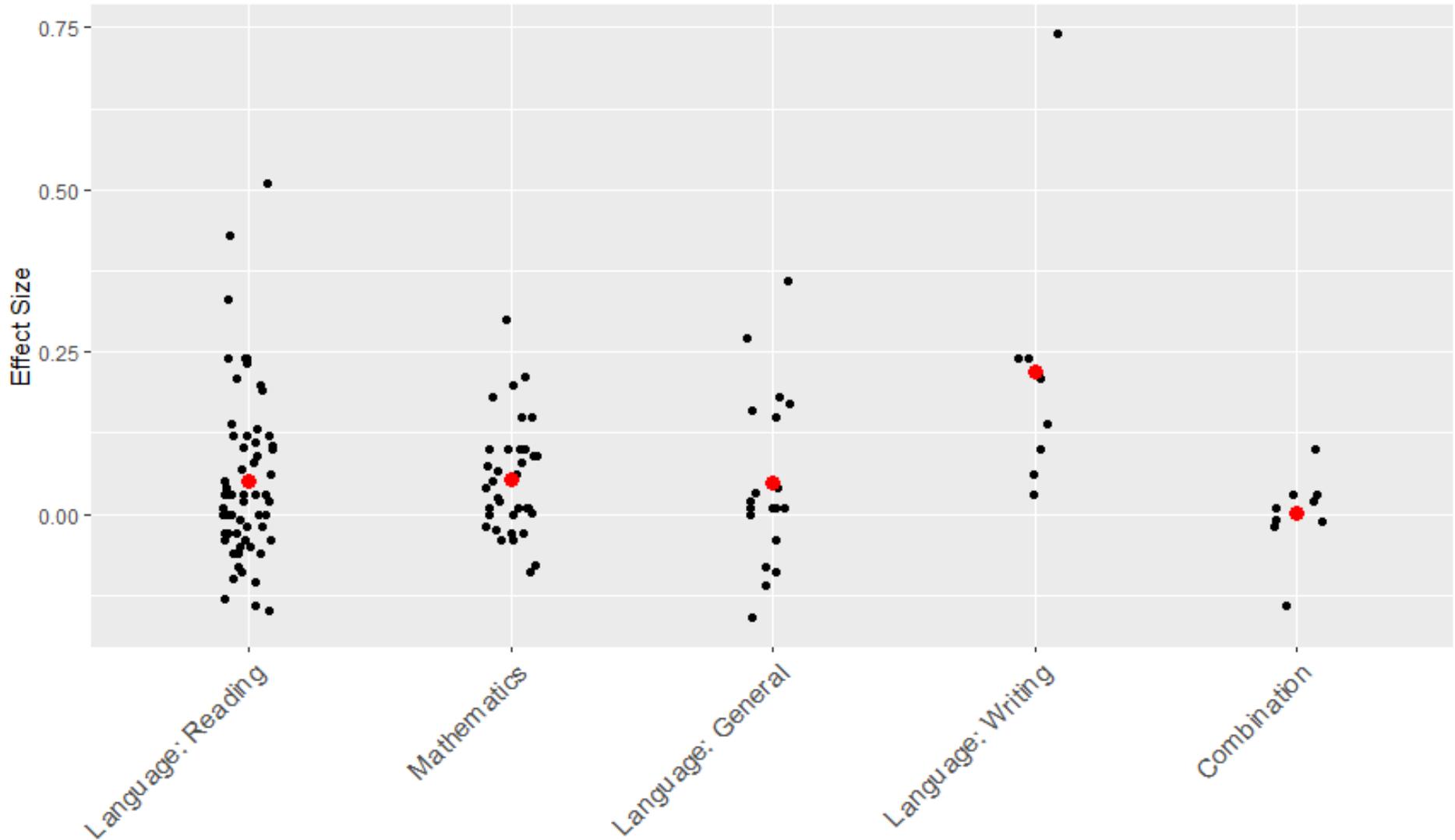
$Q(3) = 4.45, p = 0.216$ (not sig)

Topic

Topic	Nbr
Language: Reading	63
Mathematics	35
Language: General	20
Language: Writing	8
Sciences	4
Combination of topics	10

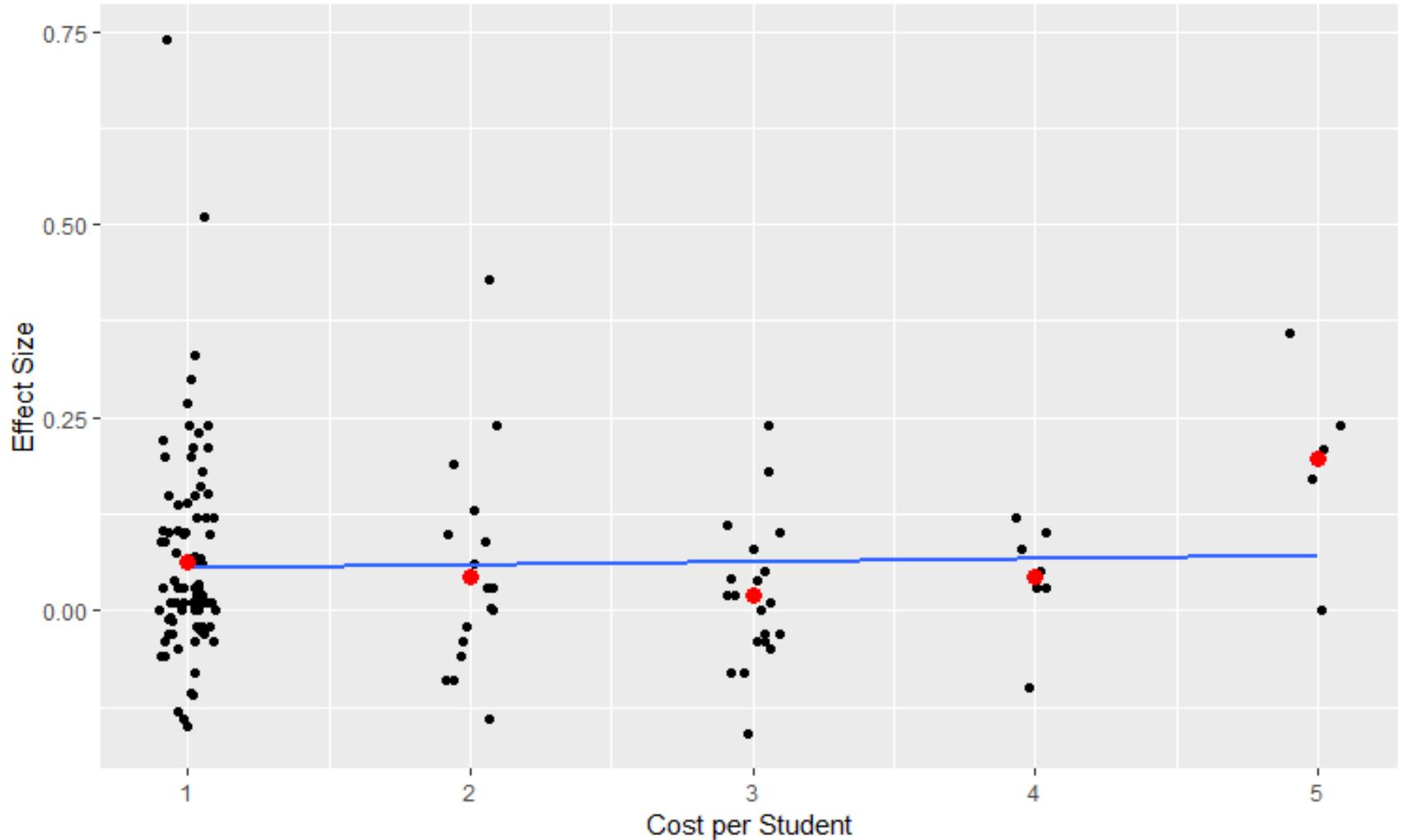
Topic

$Q(4) = 8.89, p = 0.064$ (not sig)

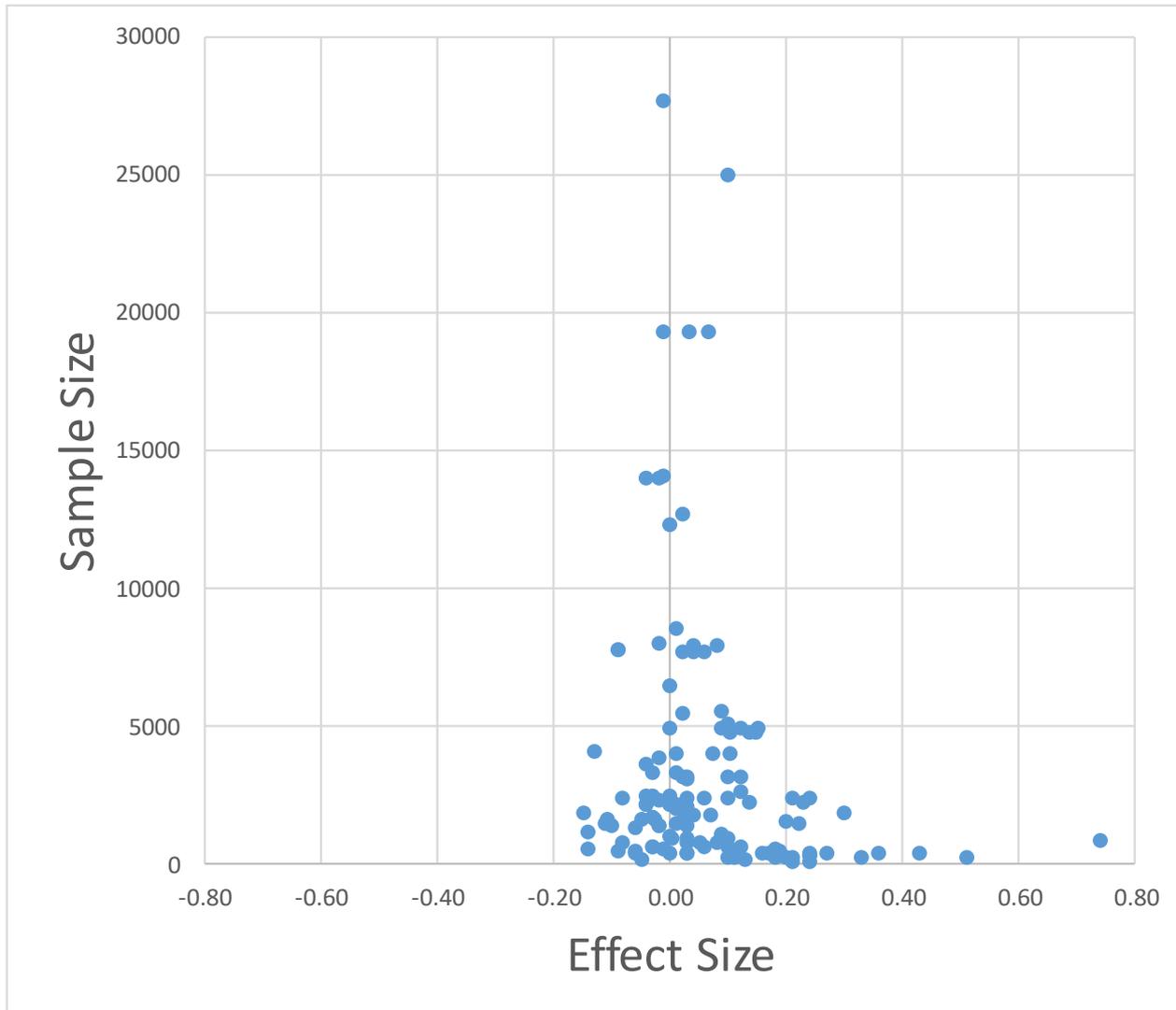


Cost per student £££££

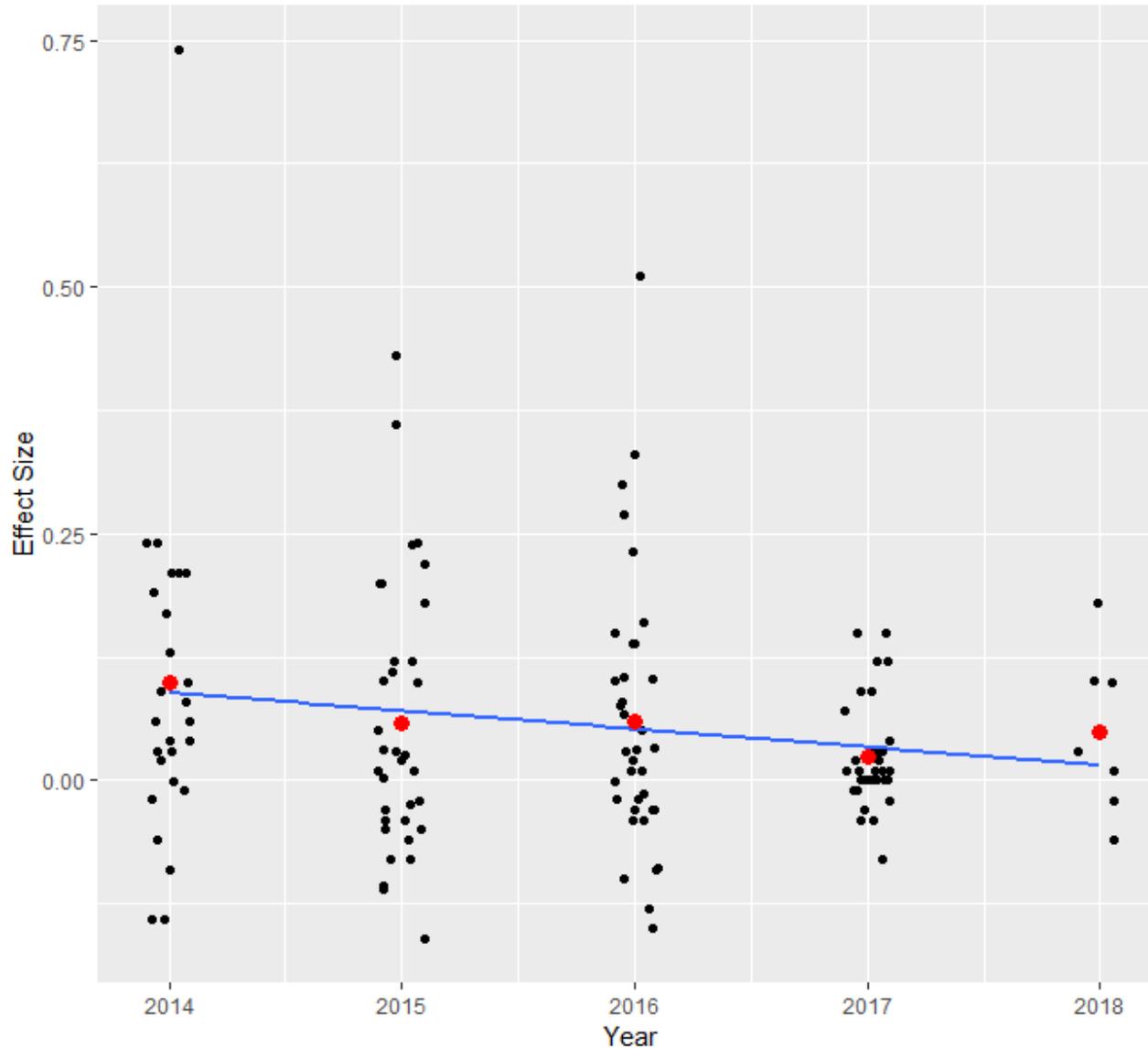
$r = 0.03$



Effect Sizes by Sample Size



Are ES becoming larger over time?



$r = -0.17$

How about in the US?



- The Institute of Education Science (IES)
 - The National Center for Education Evaluation (NCEE)
 - Since 2007

School Year	Nbr
Kindergarten	10
Kinder & Elementary	2
Elementary	73
Secondary	24
Elem & Secondary	22
Total	131

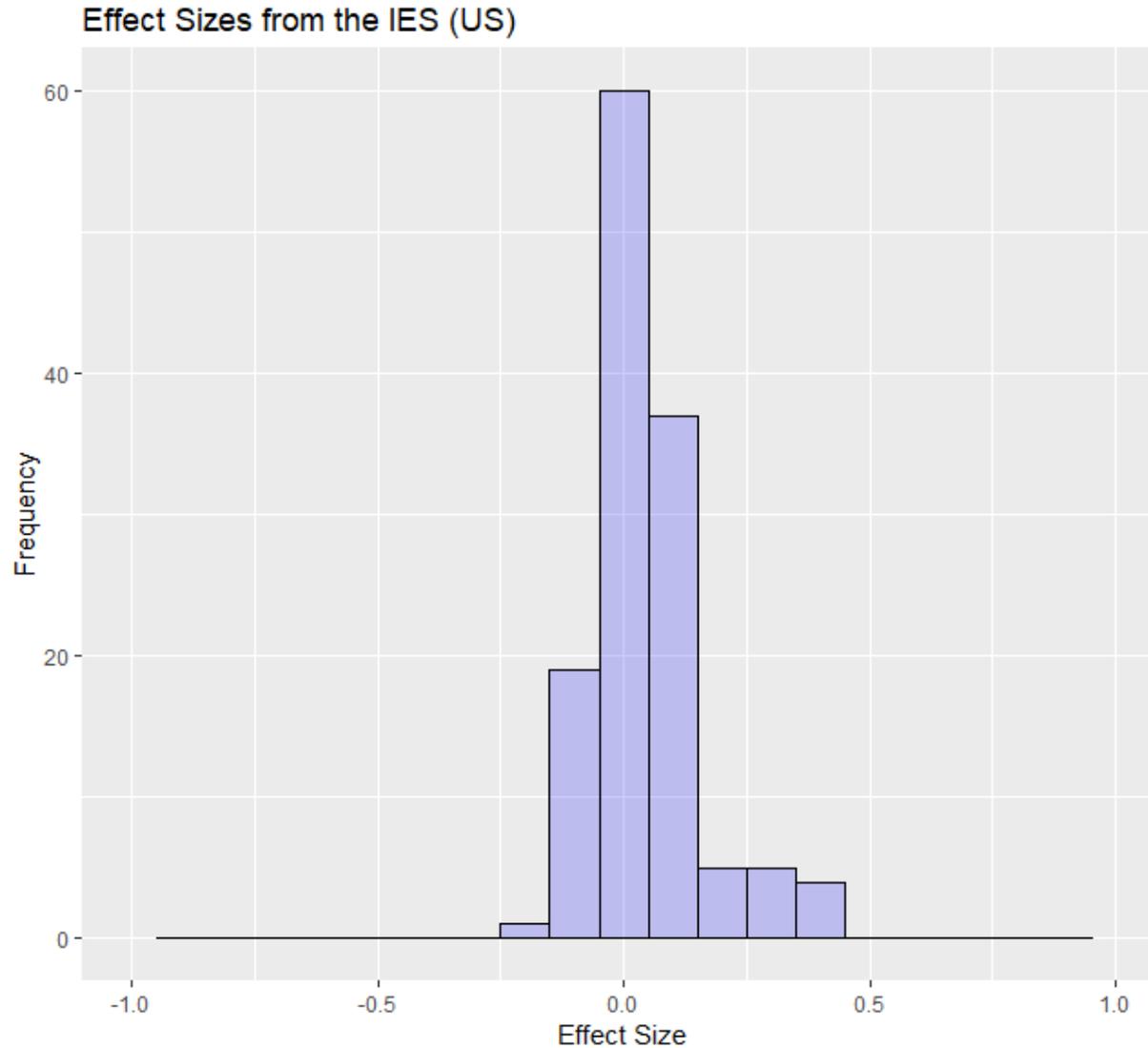
Topic	Nbr
Language: Reading	61
Mathematics	39
Language: General	17
Sciences	4
Economics	2
Social Studies	1
Language: Writing	1
Combination	6
Total	131

How about in the US?

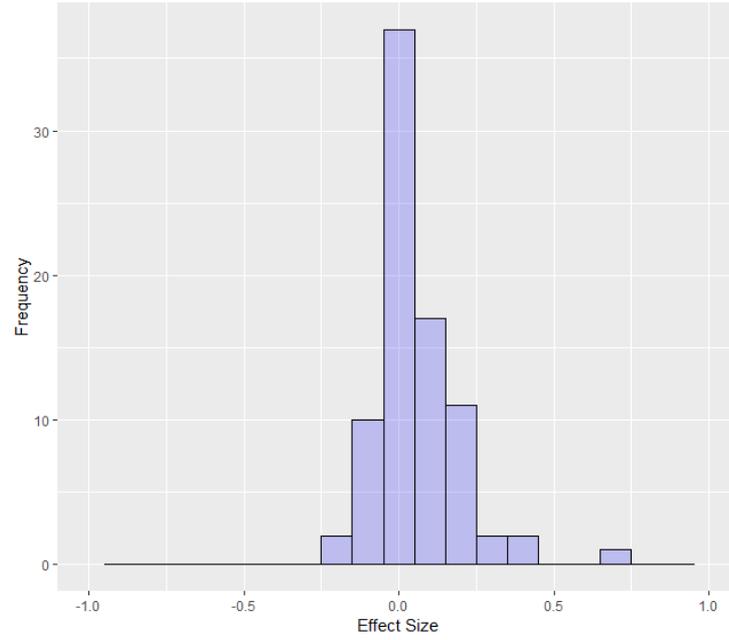
- EEF (UK)
 - Mean effect size: **0.06**

- IES (US)
 - Mean effect size: **0.06**

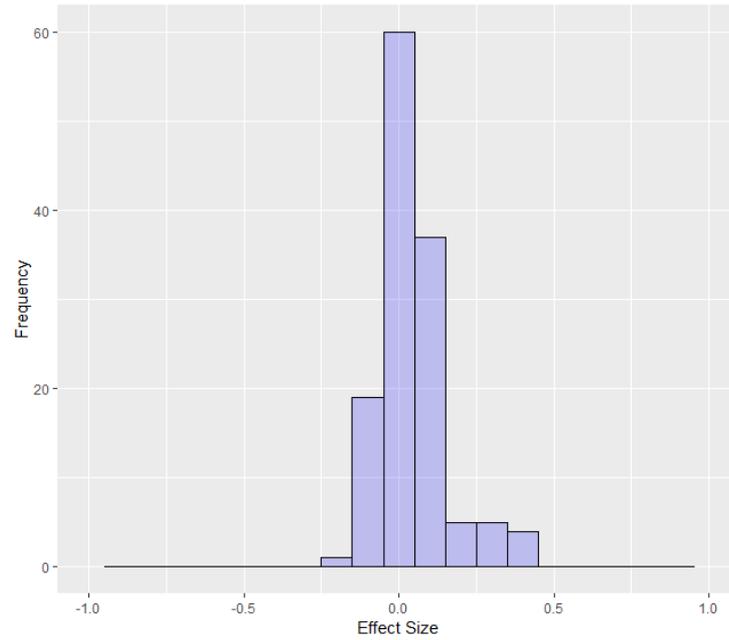
How about in the US?



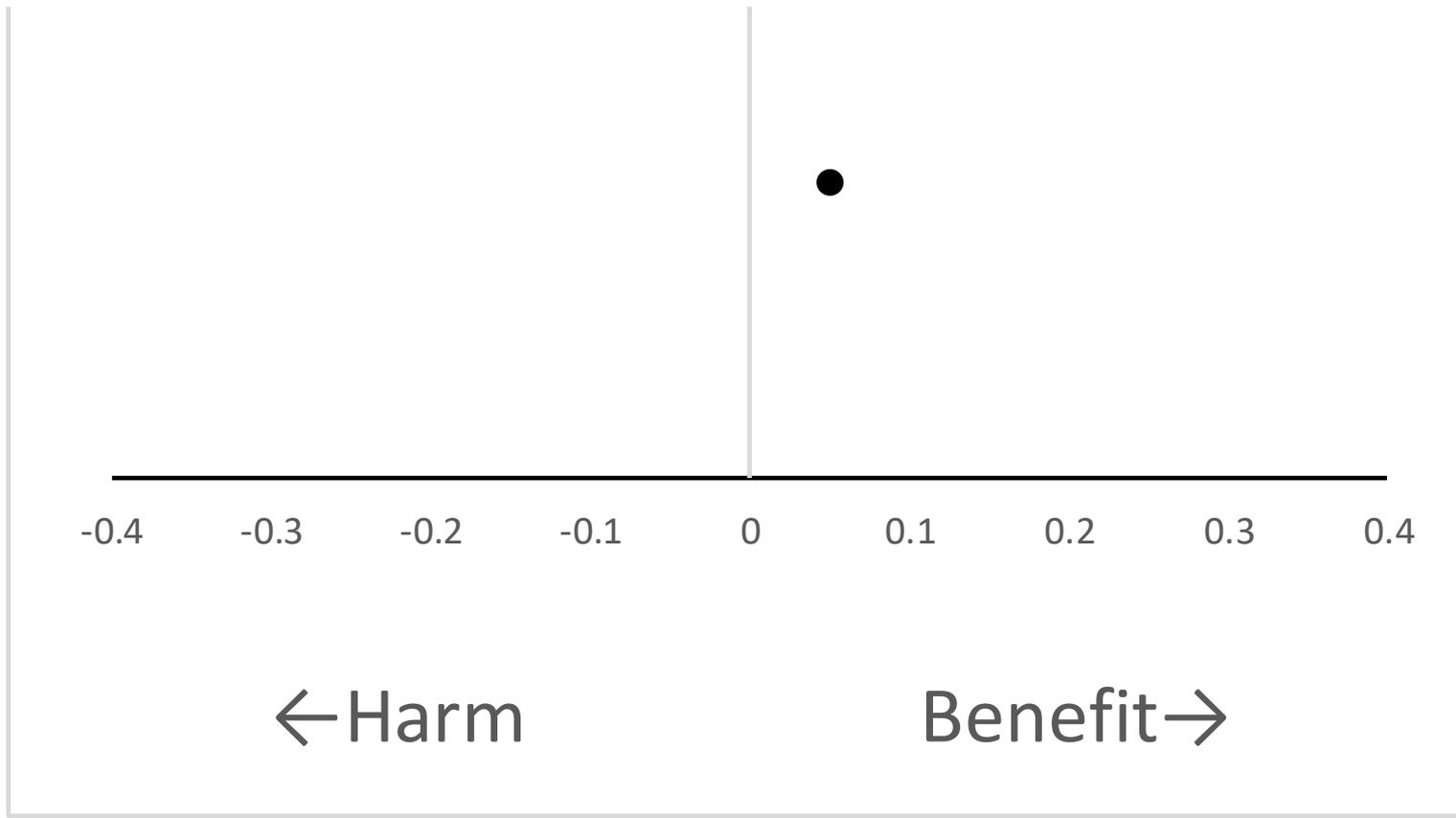
Effect Sizes from the EEF (UK)



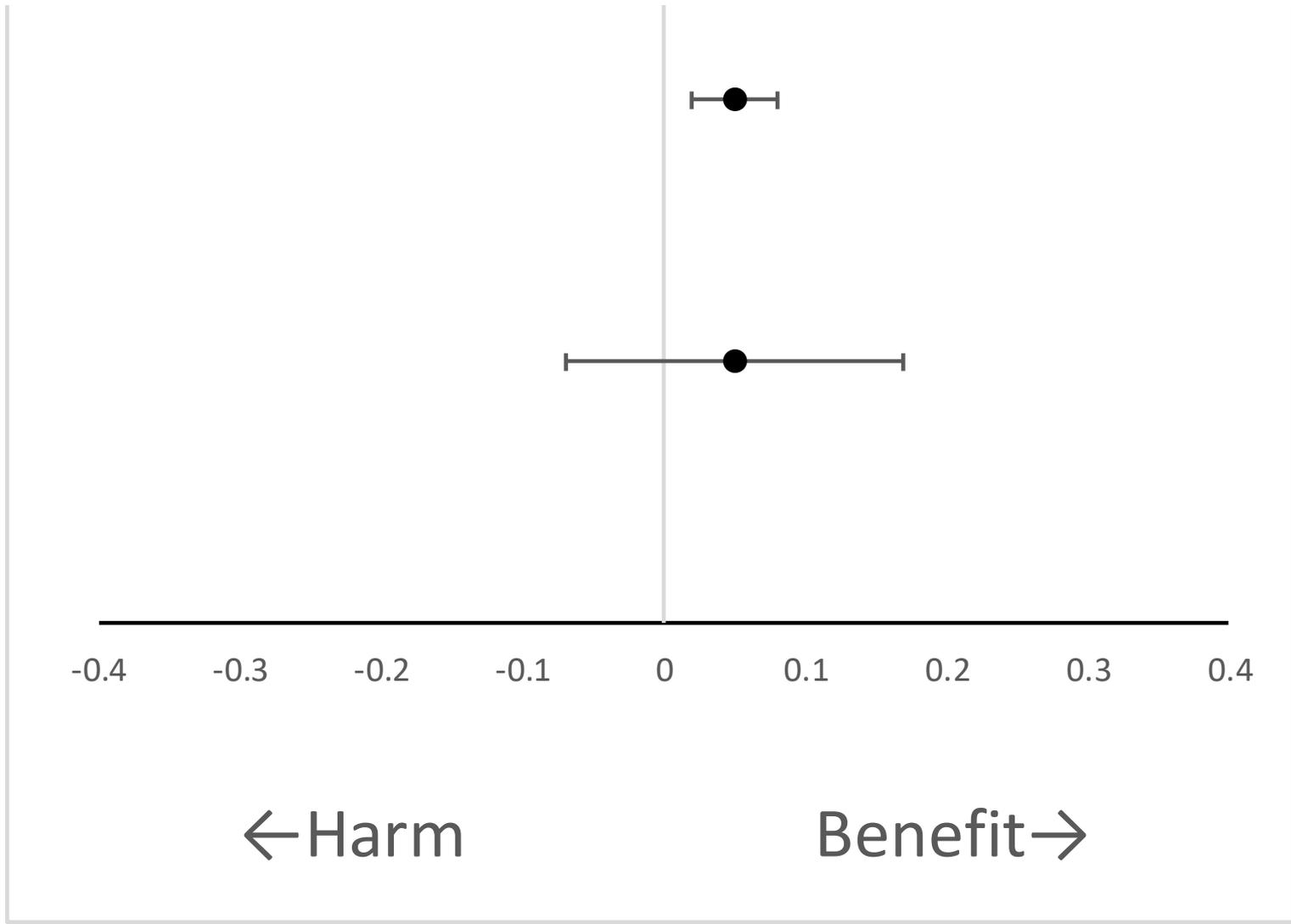
Effect Sizes from the IES (US)



Nothing works?

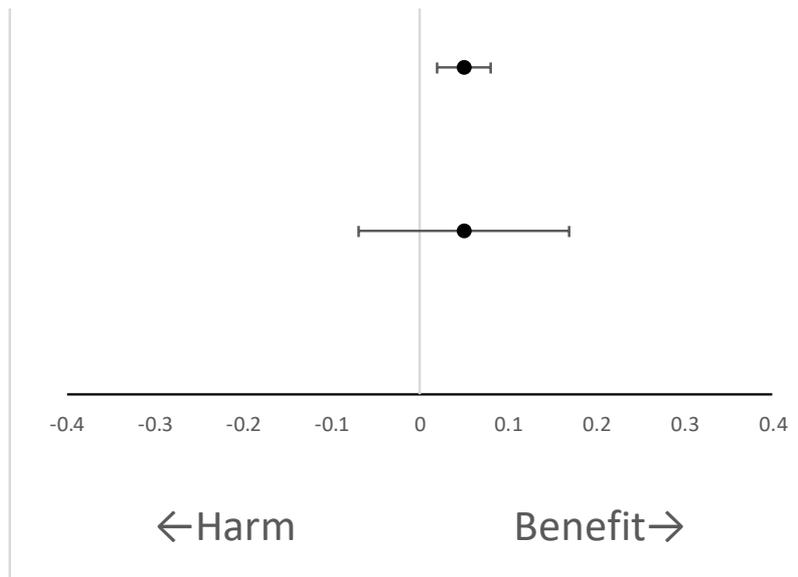


Precision of effect sizes



How large were confidence intervals?

- Average effect size: 0.06
- Average confidence interval width: 0.30



- Only 23% of trials were statistically significant

Are trials informative?

- **Statistical significance:**
 - Significant → Evidence of effectiveness
 - Not significant → Uninformative

- **Bayes factor:**
 - Evidence of effectiveness
 - **Uninformative**
 - **Evidence of ineffectiveness**

What is Bayes Factor?

- Ratio contrasting the probability of the data fitting under one hypothesis compared to another.

- $BF = \frac{P(Data | H_1)}{P(Data | H_0)}$

- We set up models for H_0 and H_1
 - H_0 : the true effect size in the population is 0
 - H_1 : the true effect size in the population comes from some positive distribution.

Bayes Factor

- What is a sensible model for H_1 ?

How Methodological Features Affect Effect Sizes in Education

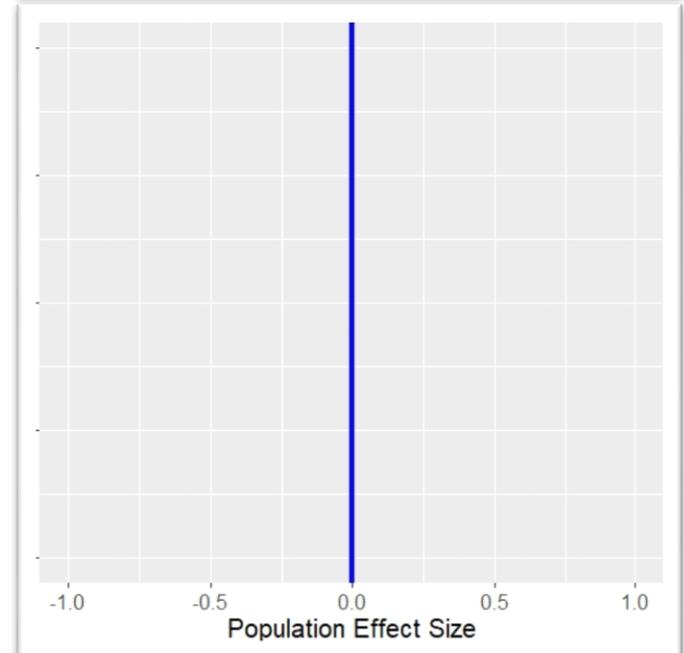
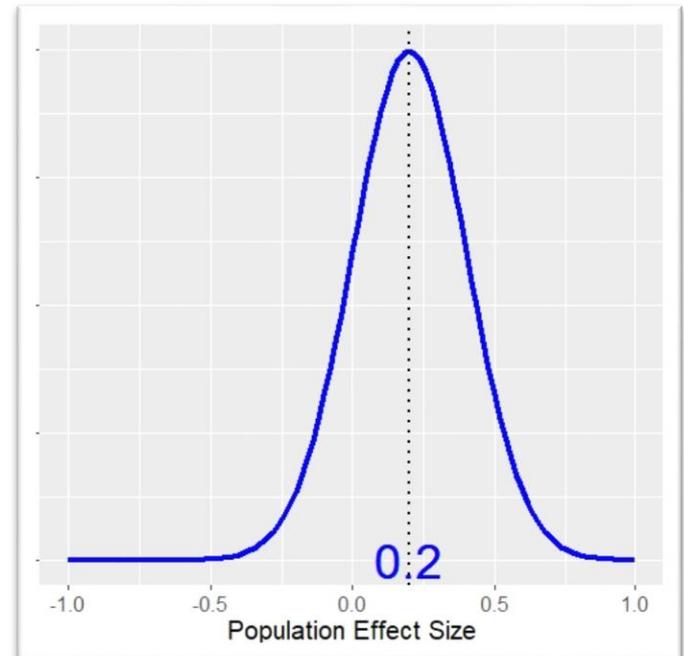
Alan C. K. Cheung¹ and Robert E. Slavin²

- Average effect size of randomized studies: 0.16
 - But lower in studies with larger samples
 - But lower in studies using independent outcome measures

Bayes Factor

- Model for H_1
 - Normal distribution
 - Mean: 0.2
 - SD: 0.2

- Model for H_0
 - Mean: 0



Bayes Factor

- $BF = \frac{P(Data|H_1)}{P(Data|H_0)}$
- We used conventional cut-offs:
 - BF > 3: Evidence of effectiveness
 - BF < $\frac{1}{3}$: Evidence of ineffectiveness
 - $\frac{1}{3} < BF < 3$: Uninformative

Bayes Factor

- Evidence of effectiveness: 18%
- Evidence of ineffectiveness: 45%
- Uninformative: 38%

Conclusion

- Trials often failed to provide evidence as to whether an intervention helped boost achievement or not.
- Why?
 1. The lab-based education literature is unreliable?
 2. The interventions are poorly implemented?
 3. RCTs are not adequately designed?

Conclusion

- The basic research on which interventions are based is unreliable?
 - Publication bias
 - P-hacking
 - Replication crisis
- Solutions?
 - Improving basic research
 - Preregistration, data sharing, replication
 - Greater care when assessing basic research

Conclusion

- The interventions are poorly implemented?
- Solutions?
 - Encouraging greater collaboration researchers – practitioners.

Centre for Mathematical Cognition

Following a £6.6m grant from Research England, Loughborough University is seeking to appoint up to sixteen new academic, research and professional services staff as it establishes a new Centre for Mathematical Cognition (CMC).



Loughborough
University

Conclusion

- RCTs are not adequately designed?
- Solutions?
 - Methodological reform
 - Larger sample size (very unlikely)
 - More proximal measures

A Randomized Controlled Trial of Interleaved Mathematics Practice

Doug Rohrer, Robert F. Dedrick, Marissa K. Hartwig, and Chi-Ngai Cheung
University of South Florida

Reactions

What do we mean by uninformative?

Absurd! Near zero effect sizes do not mean the trial is a waste or inconclusive. Same in US. Most ideas do not work even with equipoise. Sorting out which are promising and which not is invaluable. What is this report on about. Better story - most things people think work don't

Schools Week @SchoolsWeek

The EEF should trial new programmes at a small scale before blowing £500,000 on each, say researchers schoolsweek.co.uk/most-eef-trial...

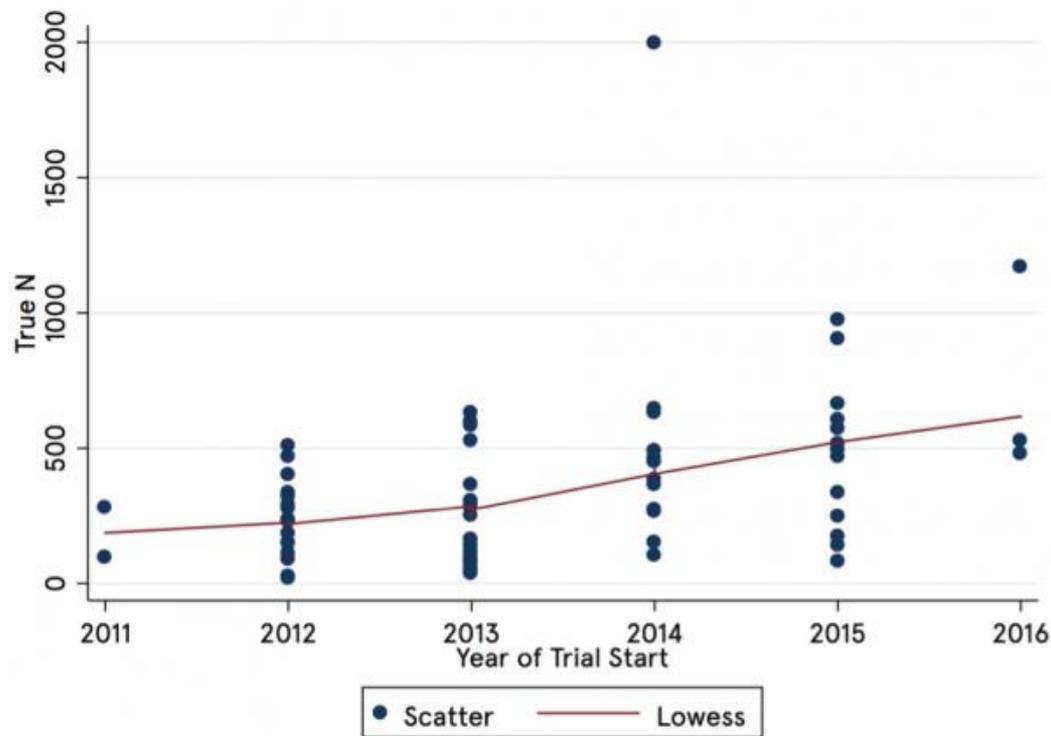
5:41 am - 16 Feb 2019

9 Retweets 27 Likes

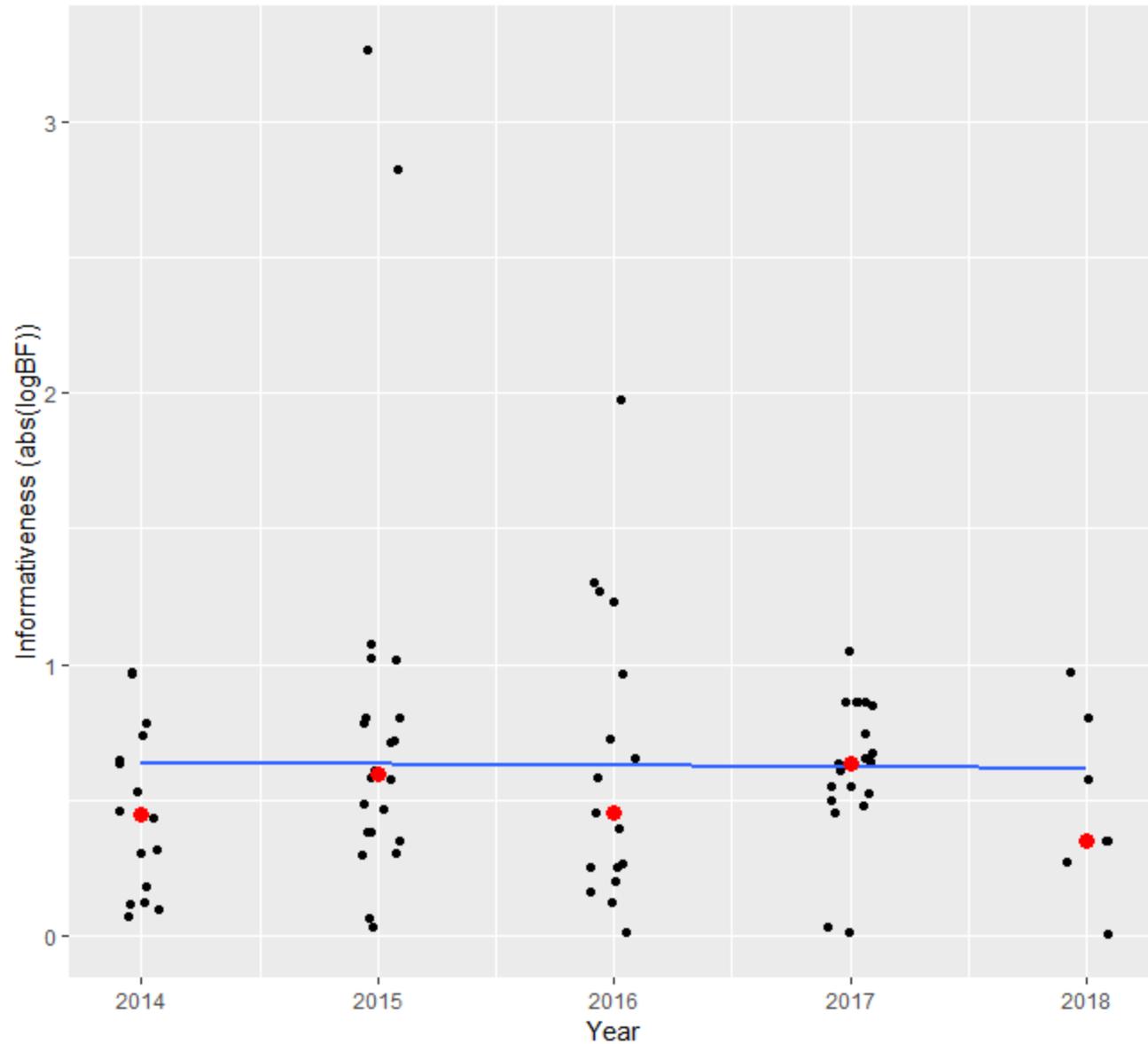


The challenges of being a trailblazer: Learning about learning

Dr Michael Sanders, 8 March 2019 - Blog posts



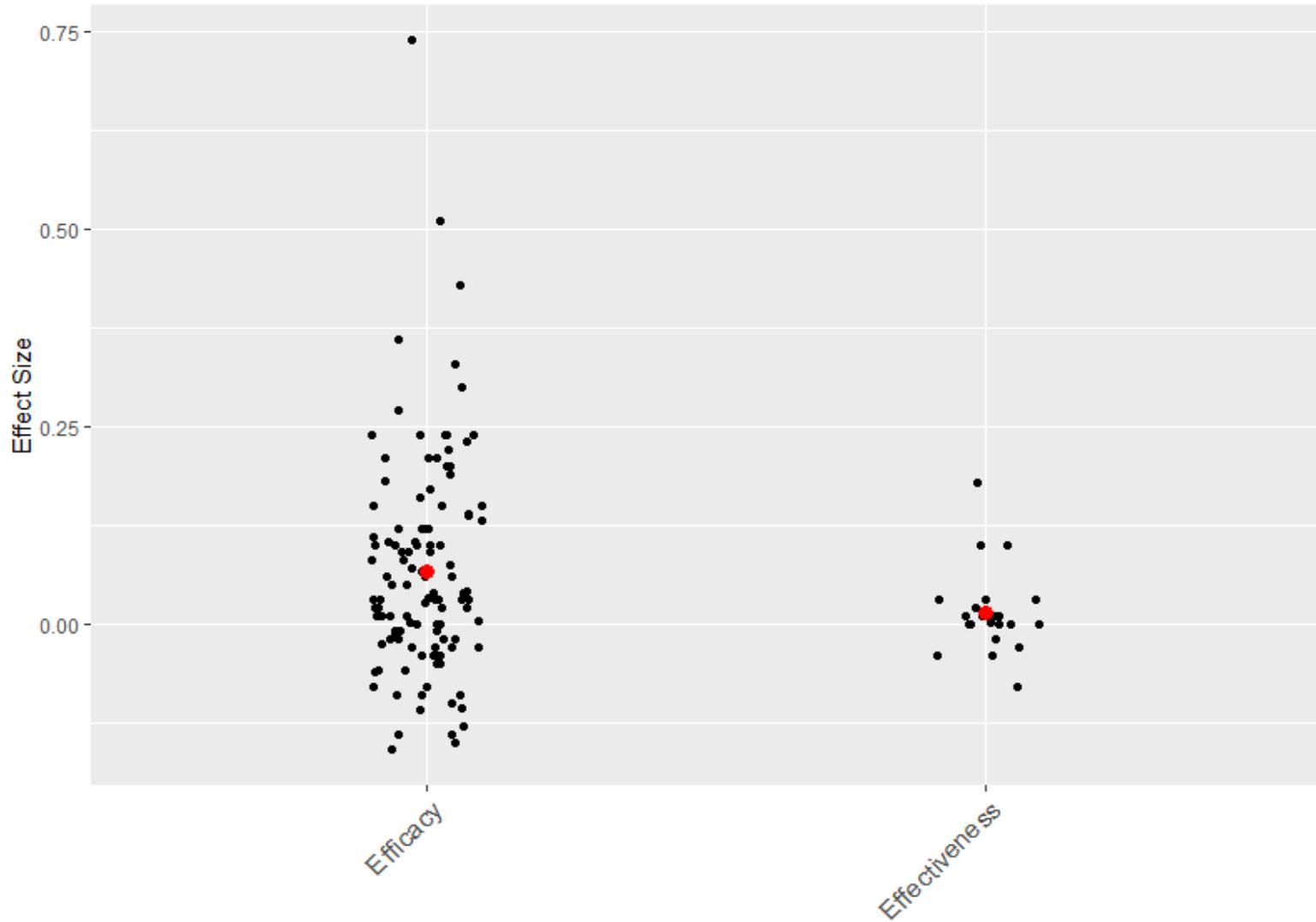
Are trials becoming more informative?



$r = -0.01$

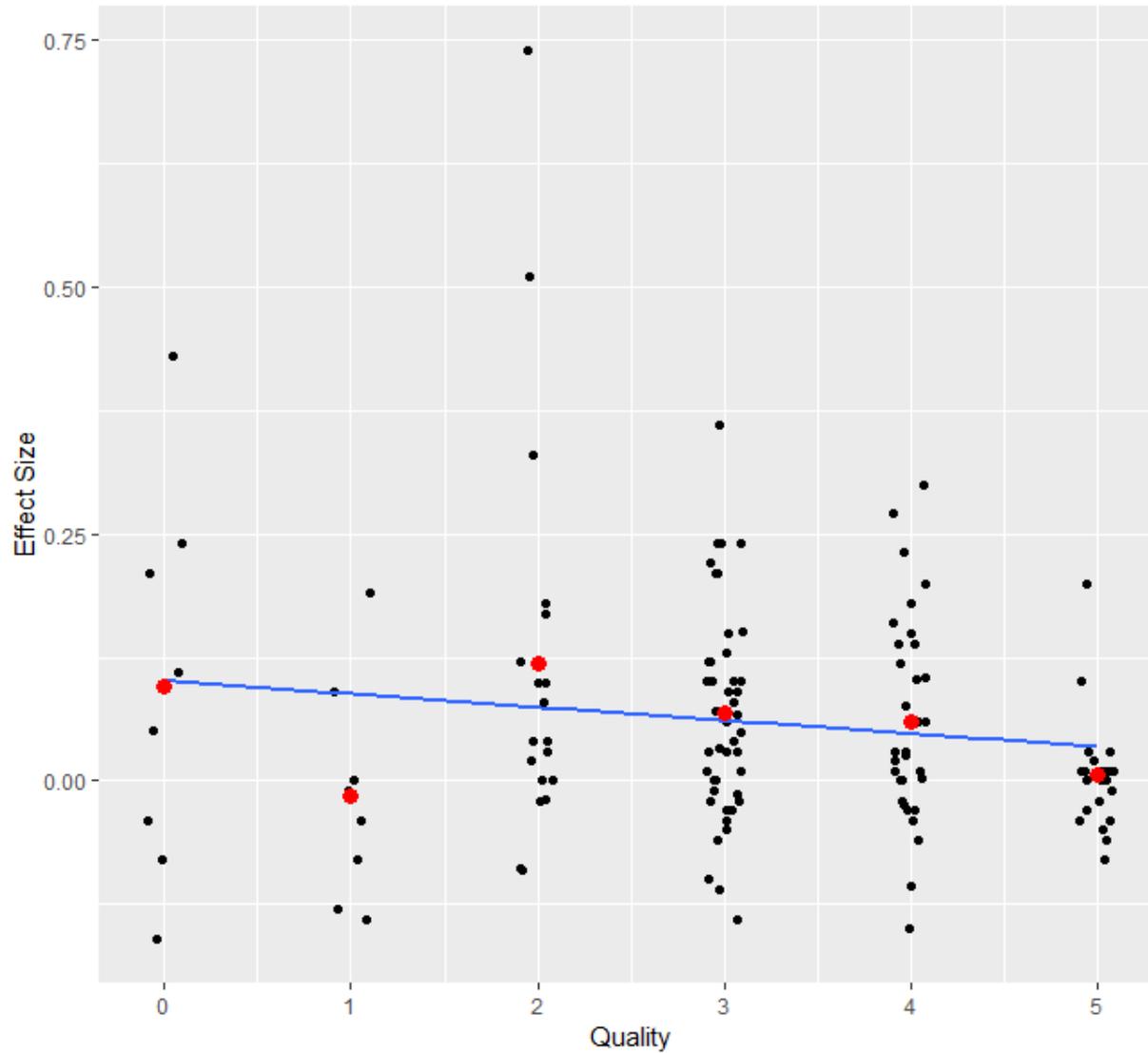
Thank you

Type of trials



Q(1) = 4.23, p = 0.04 (Sig)

Quality of the trial



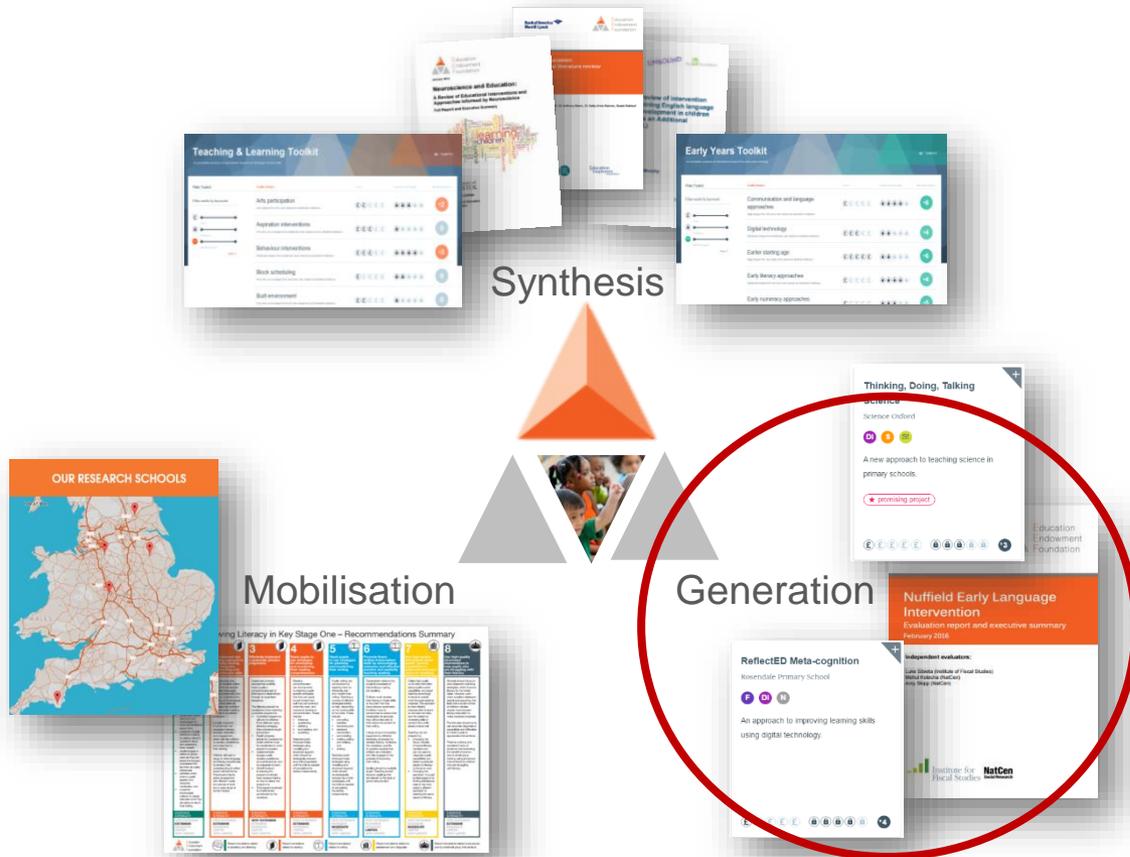
$r = -0.14$

**Reflections on 8 years of
commissioning
education RCTs:
What could we do
different?**



**Camilla Nevill
September 2019**

What we do...



155 RCTs (190 evaluations)	children and young people reached 1,300,000
£114 million total funding committed to date	13,000+ schools, nurseries, colleges involved

Three successes to celebrate...

- 1) **High-quality, independent RCTs of education programmes are possible at a grand scale.**
- 2) Together we have developed capacity to conduct education RCTs and raised standards.
- 3) We have learnt new things about what does and does not work.



Three successes to celebrate...

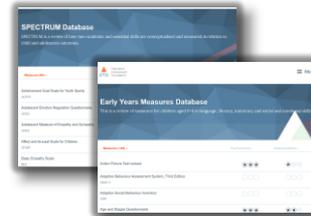
1) High-quality, independent RCTs of education programmes are possible at a grand scale.

2) Together we have developed capacity to conduct education RCTs and raised standards.

3) We have learnt new things about what does and does not work.



Developing capacity and standards has been a journey...



SAP template and analysis guidance #3



Independence standards

Analysis guidance #1

Padlocks

SPECTRUM and early years databases

2011

2013

2015

2017

2019

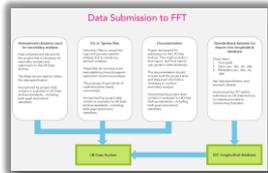
2012

2014

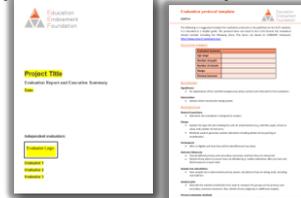
2016

2018

Data archive



Reporting and protocol templates



IPE Handbook



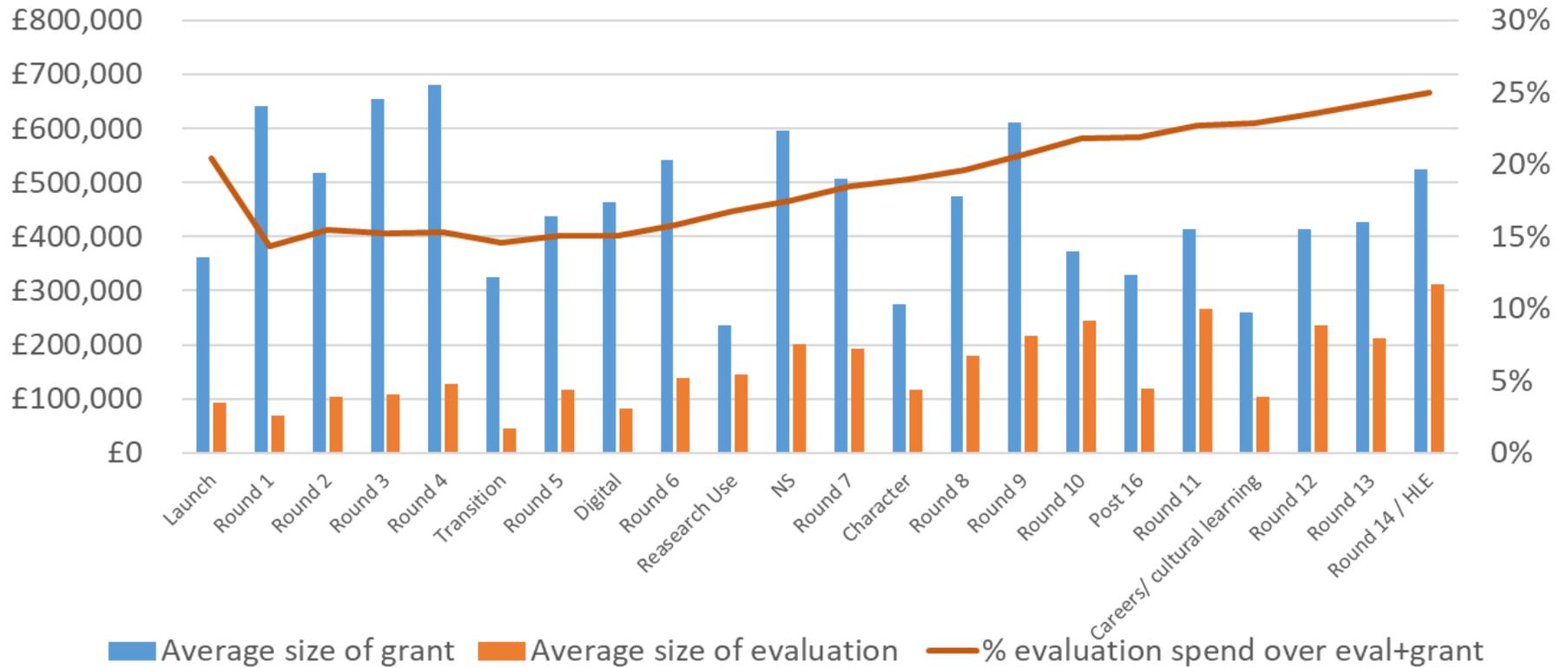
Cost guidance



IPE guidance, QED study plan, longitudinal analysis



The % cost of evaluation (cumulative) by funding round



EEF's 25 evaluation partners



EUROPE



Three successes to celebrate...

- 1) High-quality, independent RCTs of education programmes are possible at a grand scale.
- 2) Together we have developed capacity to conduct education RCTs and raised standards.
- 3) We have learnt new things about what does and does not work.



Three difficult things...

- 1) RCTs are not suited to answering some kinds of questions.
- 2) Few things work better on average than business as usual and few things scale well.
- 3) Funds and time are limited.

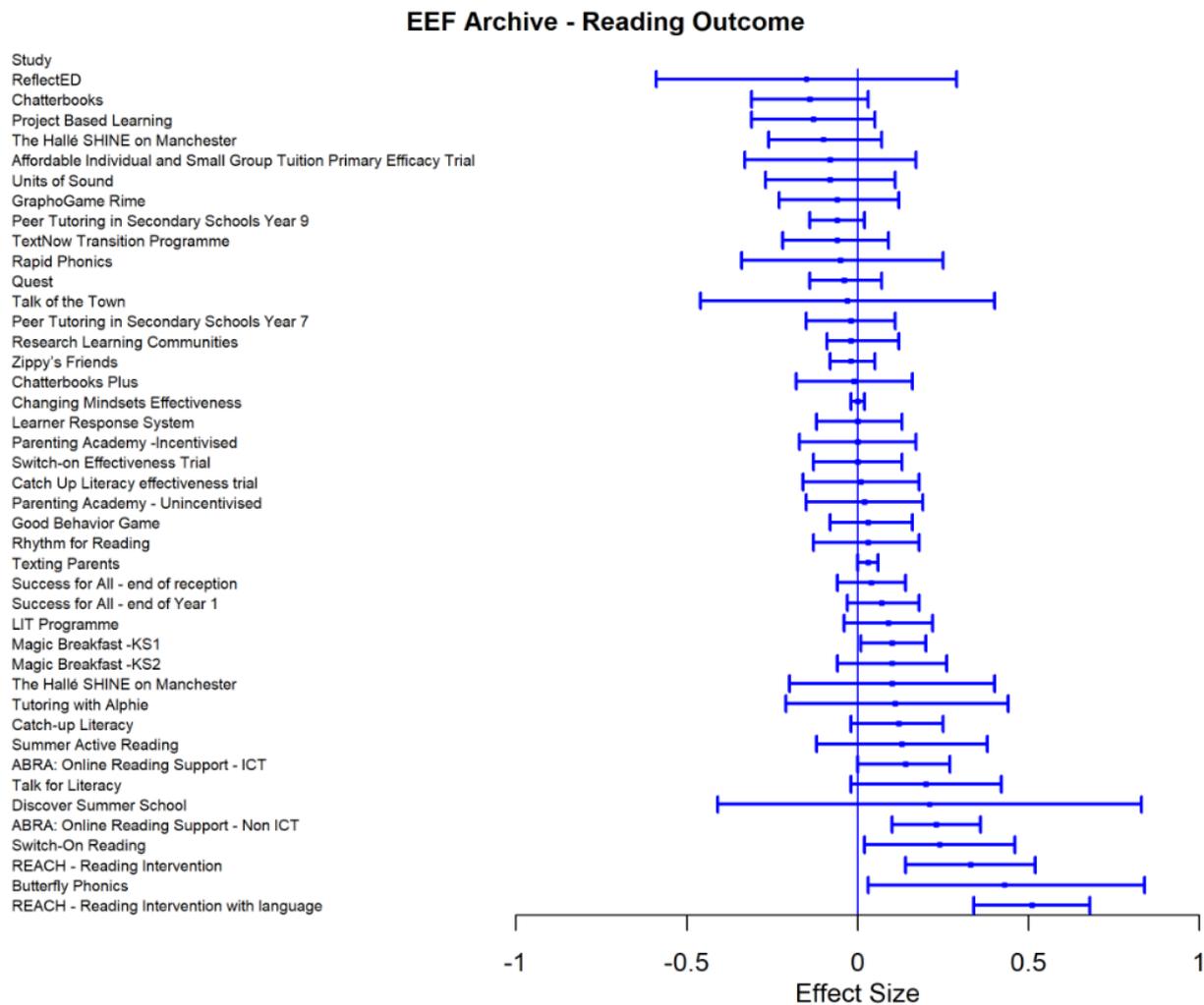


Three difficult things...

- 1) RCTs are not suited to answering some kinds of questions.
- 2) Few things work better on average than business as usual and few things scale well.
- 3) Funds and time are limited.



Effect sizes for reading in 43 archived EEF RCTs

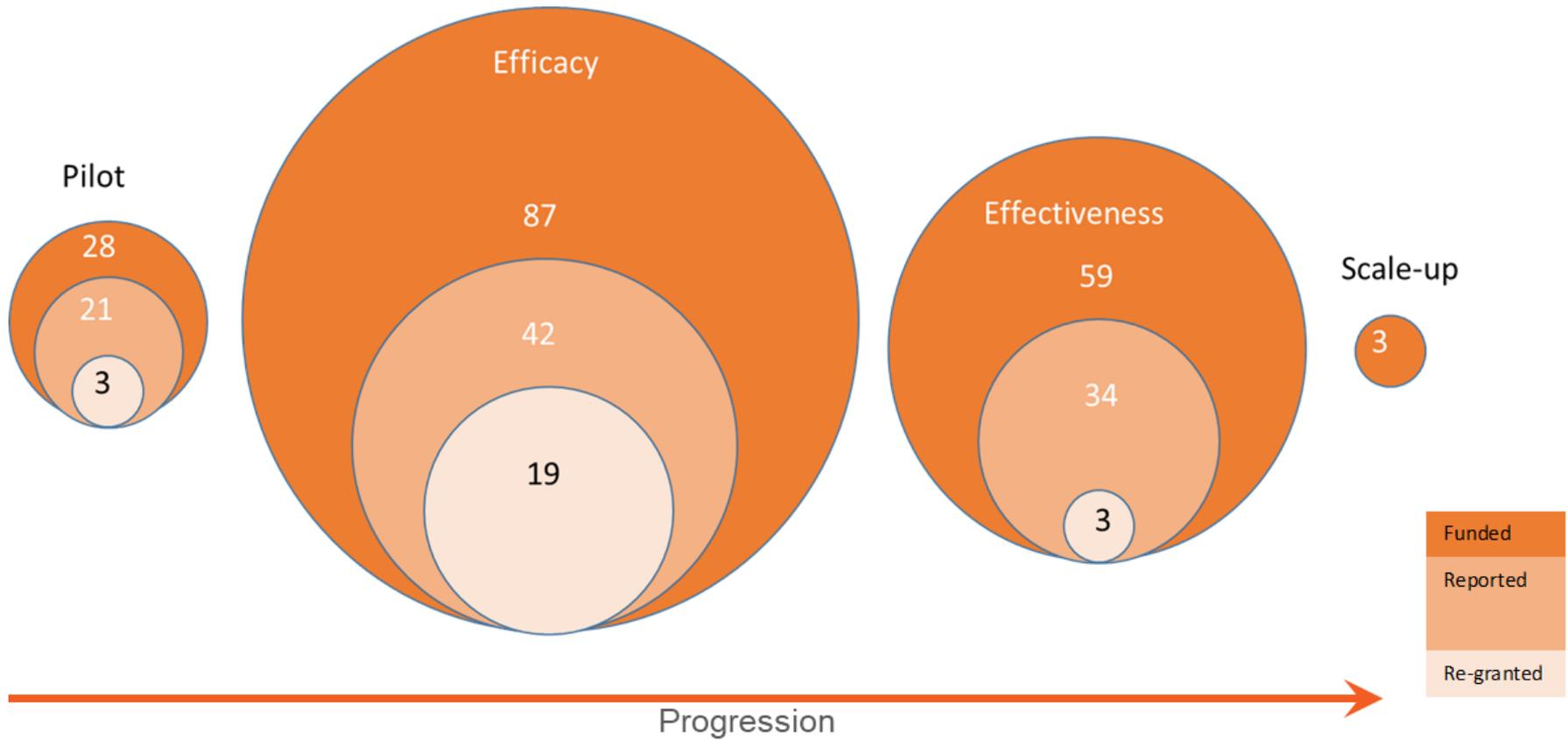


Why are we not seeing larger effects, including at scale?

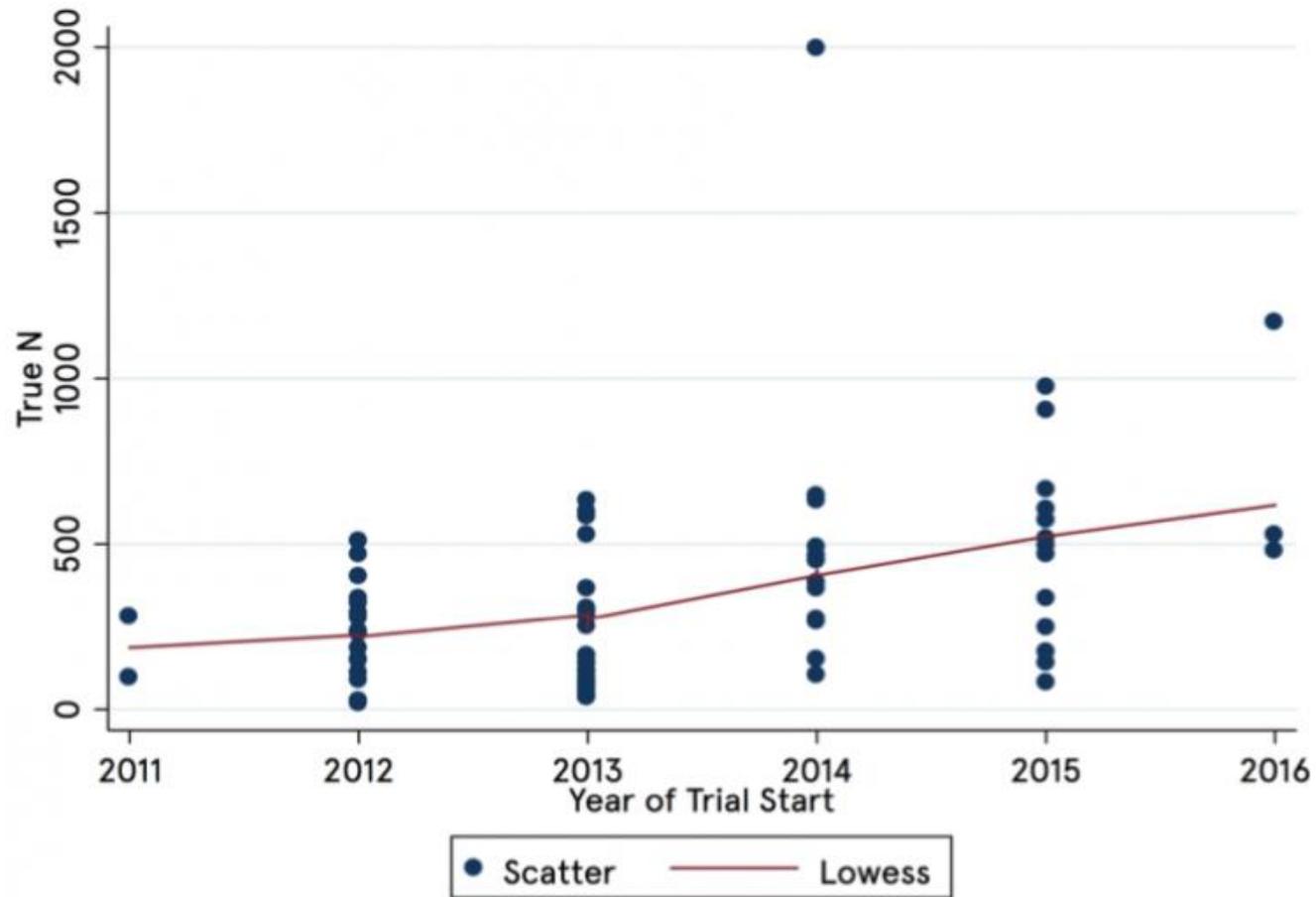
- 1) Highly active control group and optimised teaching profession?
- 2) RCTs tell us what works on average but often the answer depends.
- 3) Tension between testing what is in the (fragmented) system, versus testing what is new and theory-driven.
- 4) All the usual challenges of scaling.



The pipeline of EEF trials



The size of EEF trials has grown...



Sanders, M. (2019) *The Challenges of being a trailblazer*. What Works blog

Three difficult things...

- 1) RCTs are not suited to answering some kinds of questions.
- 2) Few things work better on average than business as usual and few things scale well.
- 3) **Funds and time are limited.**



What next?

1) What to test?

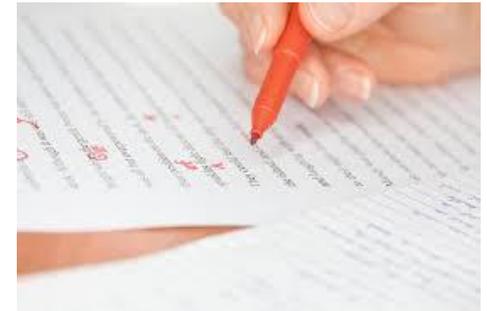
2) How to test it?

3) The power of RCT data!

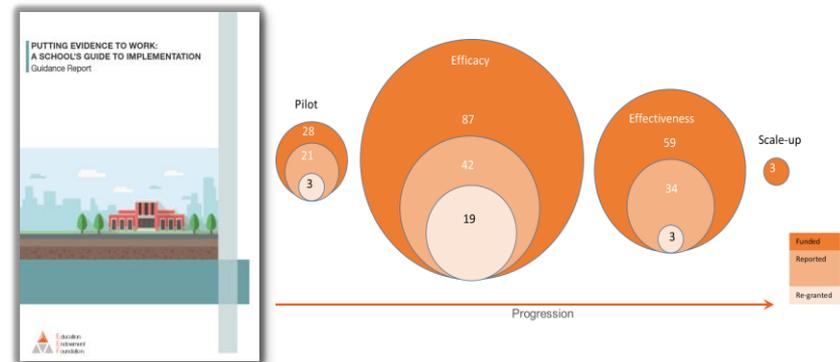
School Choices



Teacher Choices



Well-implemented, theory-driven programmes



What next?

1) What to test?

2) How to test it?

3) The power of RCT data!

New guidance on IPE, cost and measures



Alternative designs



Finding a balance between innovation and comparability, transparency and pre-specification



What next?

1) What to test?

2) How to test it?

3) The power of RCT data!

The EEF's archive holds 100 RCTs linked to long-term outcomes and is powerful for understanding variation.



In EEF's next eight years...

What could we do different?

Thank you

Camilla.Nevill@eefoundation.org.uk

CELEBRATING

100
YEARS

of RCTs in education

100 years of education trials: no significant difference?



@TheNFER

@RoyalStatSoc



www.nfer.ac.uk

www.rss.org.uk

#EducationRCTs100



NFER

National Foundation for
Educational Research

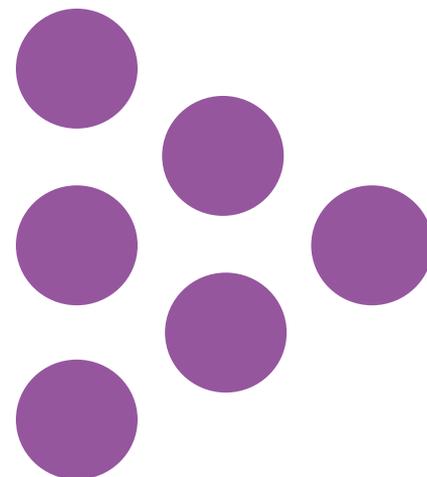


ROYAL
STATISTICAL
SOCIETY

DATA | EVIDENCE | DECISIONS

English education RCTs in 2019 – how far have we come?

Dr Ben Styles



A reminder of how things were

If you did this analysis, we suggest it is also included. If you did not, an acknowledgement that results could have been different had you chosen a different metric for fidelity would be appropriate...

A reminder of how things were

...we suggest that a statement is added about the bias that might result from comparing a subgroup of the intervention pupils who had performed well [in the intervention tasks]...to the whole of the control group.

A reminder of how things were

It seems...that independent samples t-tests were run on five post-test scores and one of them returned a p-value of 0.05 (the only non-blinded measure).

A reminder of how things were

In terms of other subgroups that you analyse...I wonder if you could acknowledge the exploratory nature of this further analysis and the need for another RCT on the specific subgroup(s) of interest.

RECENT IMPROVEMENTS

Replicability – progress made in England

Embraced	Partial adoption	Still to do
Large-scale collaborative research	Reproducibility practice (e.g. Statistical Analysis Plan)	Replication culture
Trial registration		More appropriate (usually more stringent) statistical thresholds
Standardisation of definitions and analyses		Training of the scientific workforce
Improvement in study design		

Ioannidis, J. P. A. 2014. “How to Make More Published Research True.” PLoS Medicine 11 (10): e1001747.doi:10.1371/journal.pmed.1001747.

Replicability

-
- independent evaluation
 - pre-specification of single primary outcome
 - data sharing
 - analysis code sharing

Independent evaluation

-
- Possibly unique to English education trials
 - Implies large-scale collaboration; helpful for replicability

Unexpected consequences:

- Developer recruits schools
- Incentives sometimes more concerned with randomised group than measurement of outcomes

DIFFERENCES

Governance of evaluation

EEF

Evaluation Advisory Board
Evaluation team

Nuffield

Programme Head
Advisory Group (independent; one
per trial)

NIHR

Trials Coordinating Centre (NETSCC);
Trial Steering Committee (independent,
independent chair; one per trial); Data
Monitoring and Ethics Committee
(independent; one per trial)

Proposal process

EEF

Three weeks plus one week for
EEF review
Set-up can take months

Nuffield

Months plus months for review by
independent researchers

NIHR

Months plus months for review by
independent researchers

FUTURE

Sample size

Outcome description	Outcome measure	Effect size (Hedges' g)	95% confidence interval (lower)	95% confidence interval (upper)	P	Number of intervention pupils in model	Number of control pupils in model
Primary	Reading, spelling and grammar (Short form of PiE)	0.36	0.19	0.52	<0.001	149	142
Primary (FSM)	Reading, spelling and grammar (Short form of PiE)	0.40	0.15	0.66	<0.01	67	57

Graduate Coaching Programme

Security rating	Cost
	£££££

Sample size

Outcome description	Outcome measure	Effect size	95% confidence interval	Number of control group pupils	Number of literacy group pupils	Number of numeracy group pupils	Intra-cluster correlation
Primary	Literacy score (PiE)	-0.05	-0.18–0.08	850	577	513	0.09
Primary	Mathematics score (PiM)	0.20	0.02–0.37	848	577	517	0.11

Improving Numeracy and Literacy

Evidence strength	Cost Rating*
	£
	£

Sample size – effect size

- Efficacy trial; often small effect
- Preparedness of intervention

Has a proper process of scientific enquiry got us to the point of embarking on an RCT? The hypothetico-deductive model (Cartwright, 2019)

Sample size – pre-post correlation and ICC

- Statutory tests
 - Changes to assessment
 - Non-statutory tests
 - School and pupil-level correlation
 - Is it cost effective to run a baseline?
- Papers to be written to fill these gaps.

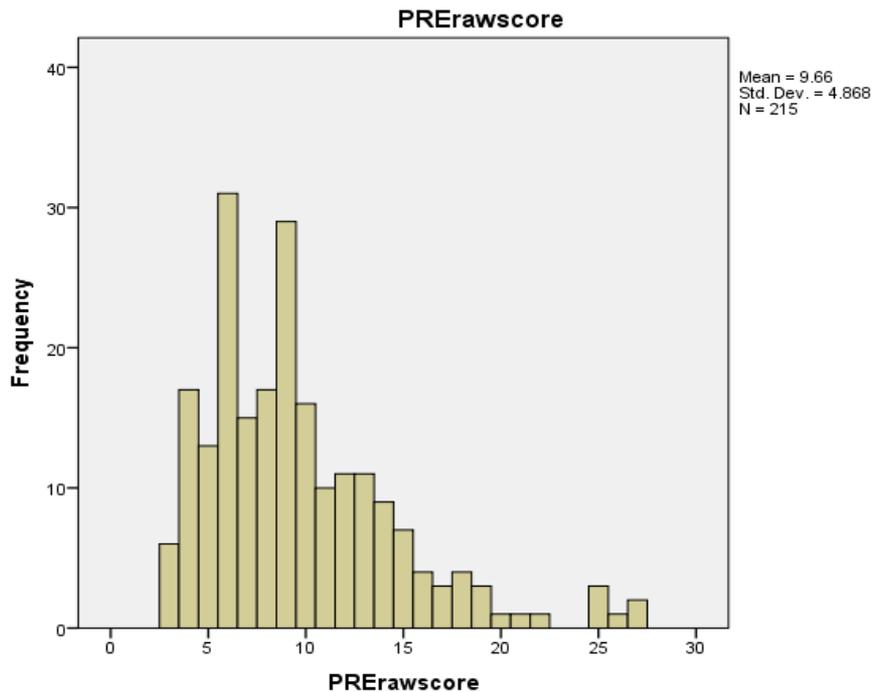
Measurement instruments and their development

- Reasonably strong infrastructure around test development (awaiting complete database; Early Years Outcomes covered)
- Good database of non-cognitive outcomes (SPECTRUM) but what happens if we need to develop a new measure?

Funding for outcome measure development and psychometrics:

- Validity
- Reliability
- Responsiveness
- Suitable for target population?

Pilot the measure



- 43 marks
- Every question has five options
- Expected score for random answers: 8.6
- Mean score at baseline: 9.7 (SD 4.9)

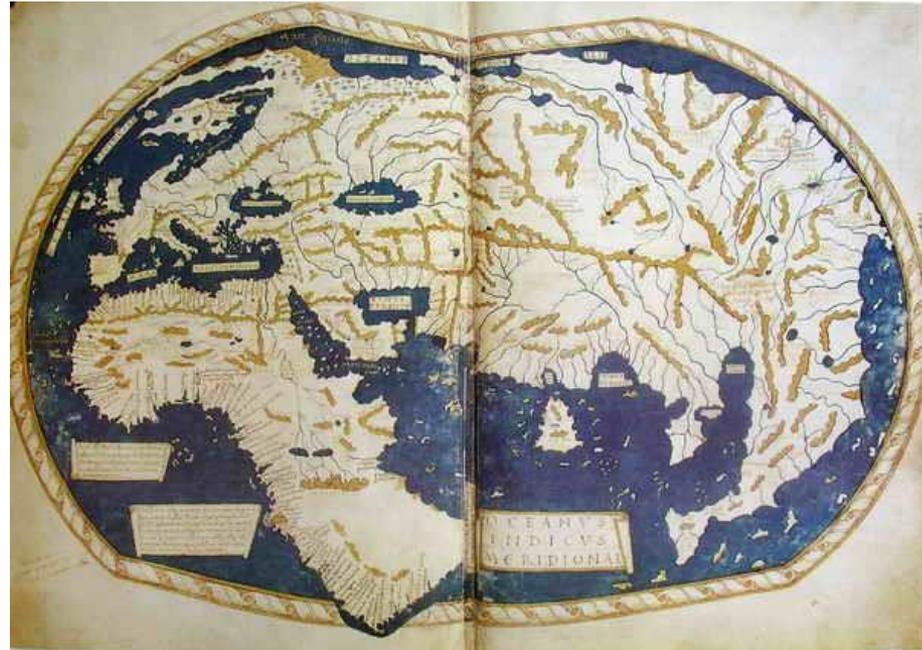
Key issues for the future

-
- Continue to prioritise the open science approach to improve replicability
 - Re-assess the process of scientific enquiry before an intervention gets subjected to an RCT
 - Allow time for piloting of the measure and/or design
 - Invest in trials infrastructure and methods research
 - Invest in training the scientific workforce

'Unzipping' the EEF toolkit: RCTs and the role of meta-analysis

Steve Higgins
School of Education
Durham University

s.e.higgins@durham.ac.uk
@profstig



**100 years of RCTs in education:
no significant difference?**

23 September 2019

Royal Statistical Society, London

Some premises

RCTs are sometimes necessary, but never sufficient to establish something 'works' (or has worked)

One trial, however rigorous and robust, will never be definitive

Meta-analysis is not replication (though at present is the best we have)

A threshold of 0.05 for 'statistical significance', in a developed field, is too low, too narrow and too confusing a bar

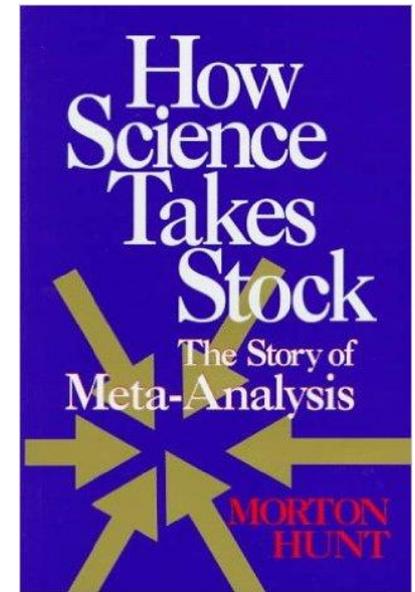
A short history of meta-analysis

Aggregating research findings to get a more definitive answer

- Pearson and Fisher
- Pratt and Rhine – a cautionary tale

Origins of meta-analysis and its early development

- Glass & Smith versus Hans Eysenck
- Rosenthal, Cohen, Hedges
- Elwood, Cochrane, Peto



Pearson, 1904

For example, taking the relation between deaths and recoveries, and presence and absence of vaccination scar in cases of small-pox, we have :²

				Correlation.
Metropolitan Asylums Board Returns,				
Epidemic 1893	0.595 ± 0.027
Epidemics for six towns	0.656 ± 0.009
Sheffield, 1887-8	0.769 ± 0.012
Homerton and Fulham, 1873-85	0.576 ± 0.009
London: Epidemic 1901	0.578 ± 0.031
Glasgow: Epidemic 1900-1	0.629 ± 0.030

We may safely say that the protective character of vaccination as against mortality after incurring small-pox is very substantial, and numerically it is represented by the value 0.6, which is fairly closely the actual result for the various epidemics which have at present been dealt with.

Understand the extent of the difference

Explain variation in effects

To improve the effectiveness of inoculation

The following table gives the results of calculating the correlation coefficients of the tables in Appendix B :

INOCULATION AGAINST ENTERIC FEVER:				
<i>Correlation between Immunity and Inoculation.</i>				
I. Hospital Staffs	+	0.373 ± 0.021
II. Ladysmith Garrison	+	0.445 ± 0.017
III. Methuen's Column	+	0.191 ± 0.026
IV. Single Regiments	+	0.021 ± 0.033
V. Army in India	+	0.100 ± 0.013
Mean value	+	0.226
<i>Correlation between Mortality and Inoculation.</i>				
VI. Hospital Staffs	+	0.307 ± 0.128
VII. Ladysmith Garrison	-	0.010 ± 0.081
VIII. Single Regiments	+	0.300 ± 0.093
IX. Special Hospitals	+	0.119 ± 0.022
X. Various military Hospitals	+	0.194 ± 0.022
XI. Army in India	+	0.248 ± 0.050
Mean value	+	0.193

If we except IV and VII, the values of the correlations are at least twice (in the very sparse data of VI) and generally four, five, or more times their probable errors. From this standpoint we might say that they are all significant, but we are at once struck with the extreme irregularity and the lowness of the values reached. They are absolutely incomparable with the fairly steady and large values of the vaccination correlations obtained for different epidemics and towns. The effect of enteric inoculation is evidently largely influenced by difference of environment or of treatment.

A cautionary tale

Pratt and Rhine (1940) conducted a systematic review and 'meta-analysis' of 145 ESP studies conducted between 1882 and 1939

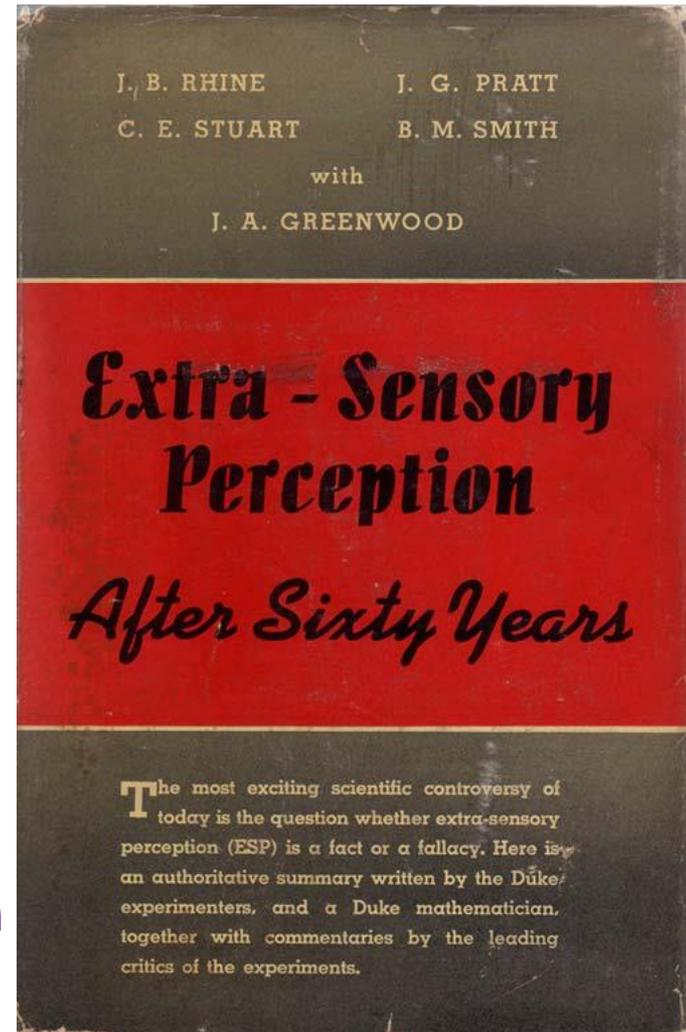
Reviewed experimental errors and clustered results from similar experiments for sub-group analysis

Conclusion: ESP works!

Conclusions from aggregation depend on internal validity of underlying studies

Highlights risks from publication bias

Highlights problems from lack of replication



Meta-analysis and meta-synthesis

Gene Glass, in his presidential address to the American Educational Research Association in 1976, introduced the term 'meta-analysis' to denote statistical synthesis of the results of similar studies.

Fraser, Walberg & Hattie (1987)

Testing Walberg's educational productivity model by reviewing the relative extent of effects across 220 meta-analyses (134 meta-analyses of achievement outcomes and 92 meta-analyses of attitude outcomes) using correlations

Hattie (1992)

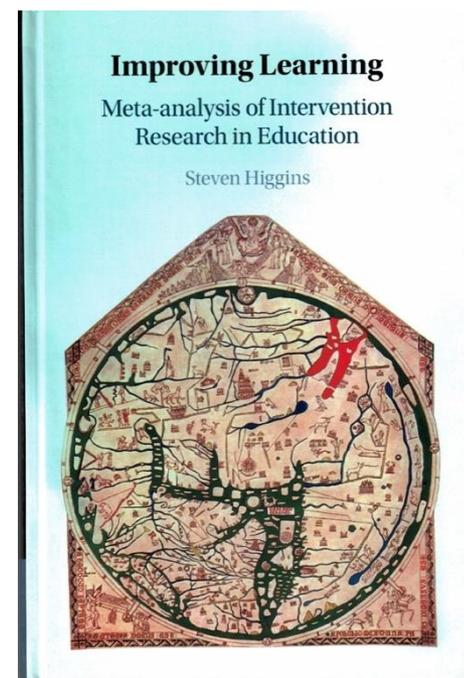
Used standardised mean differences (from Bloom)

Sipe & Curlette (1996)

Develop rigour of methodology – systematic review and synthesis

Marzano (1998)

Big questions: '*A Theory-based Meta-analysis of Research on Instruction*'



Diverse terminology

Meta-meta-analysis (Kazrin et al., 1979)

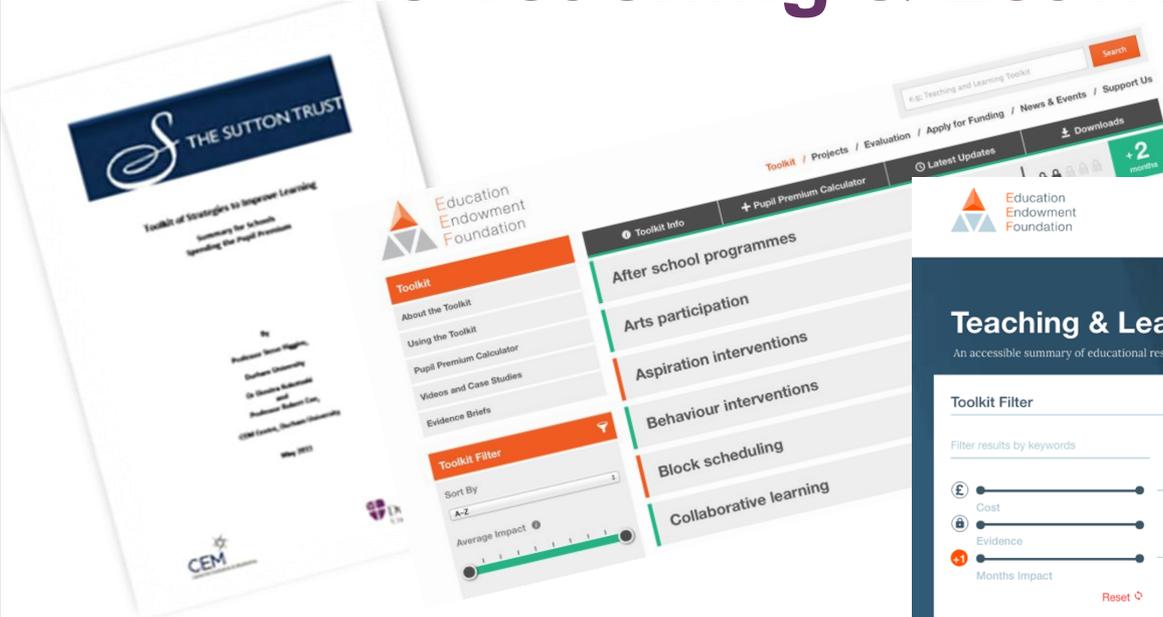
Mega-analysis (Smith, 1982)

Super-analysis (Dillon, 1982)

Super-synthesis (Sipe & Curlette, 1996)

Meta-synthesis (Sipe & Curlette, 1996)

The Teaching & Learning Toolkit



Best 'buys' on average from research

Key messages for Pupil Premium spending

Currently consulted by 70% of schools in England

Toolkit Filter	Toolkit Strand	Cost	Evidence Strength	Months Impact
Filter results by keywords	Feedback	£ £ £ £ £	🔒 🔒 🔒 🔒 🔒	+8
Cost	Meta-cognition and self-regulation	£ £ £ £ £	🔒 🔒 🔒 🔒 🔒	+8
Evidence	Peer tutoring	£ £ £ £ £	🔒 🔒 🔒 🔒 🔒	+5
Months Impact	Early years intervention	£ £ £ £ £	🔒 🔒 🔒 🔒 🔒	+5
Reset	One to one tuition	£ £ £ £ £	🔒 🔒 🔒 🔒 🔒	+5
	Homework (Secondary)	£ £ £ £ £	🔒 🔒 🔒 🔒 🔒	+5
	Collaborative learning	£ £ £ £ £	🔒 🔒 🔒 🔒 🔒	+5
	Oral language interventions	£ £ £ £ £	🔒 🔒 🔒 🔒 🔒	+5

What we tried to do

Summarise the evidence from meta-analysis about the impact of different strategies on learning (*tested attainment*) – series of integrated ‘umbrella’ reviews

- As found in research studies
- Averages

Apply criteria to evaluations: rigorous designs for causal inference

Estimate the *size* of the effect

- Standardised Mean Difference = ‘Months of gain’
- On tested attainment only

Estimate the *costs* of adopting

- Information rarely available

35 strands (main Toolkit) Over 200 meta-analyses About 8,000 studies

Early years version
Evidence for Learning in Australia
Plataforma de Prácticas Educativas Efectivas for Latin America & Caribbean (Spanish/ Portuguese)
Scottish Attainment Challenge: Learning & Teaching Toolkit
EduCaixa, Spain

Teaching and Learning Toolkit

All an overall summary of the international evidence on teaching, 150 practice

Filter Toolkit	Toolkit Strand	Cost	Evidence Strength	Impact Score
High impact for Australia	Feedback High impact for very low cost, based on moderate evidence.	●●●●●	●●●●●	+8
High impact for Australia	Metacognition and self-regulation High impact for very low cost, based on moderate evidence.	●●●●●	●●●●●	+7
High impact for Australia	Reading comprehension strategies High impact for very low cost, based on moderate evidence.	●●●●●	●●●●●	+6
Moderate impact	Homework (Secondary) Moderate impact for very low cost, based on moderate evidence.	●●●●●	●●●●●	+5
Moderate impact	Mastery learning Moderate impact for very low cost, based on moderate evidence.	●●●●●	●●●●●	+5
Moderate impact	Peer tutoring Moderate impact for very low cost, based on moderate evidence.	●●●●●	●●●●●	+5
Moderate impact	Oral language interventions Moderate impact for very low cost, based on moderate evidence.	●●●●●	●●●●●	+5
Moderate impact	Collaborative learning Moderate impact for very low cost, based on moderate evidence.	●●●●●	●●●●●	+5
Moderate impact	Early years intervention Moderate impact for very high cost, based on moderate evidence.	●●●●●	●●●●●	+5
Moderate impact	One to one tuition Moderate impact for high cost, based on moderate evidence.	●●●●●	●●●●●	+5
Moderate impact	Phonics Moderate impact for very low cost, based on moderate evidence.	●●●●●	●●●●●	+4
Moderate impact	Social and emotional learning Moderate impact for moderate cost, based on moderate evidence.	●●●●●	●●●●●	+4
Moderate impact	Outdoor adventure learning Moderate impact for moderate cost, based on moderate evidence.	●●●●●	●●●●●	+4
Moderate impact	Small group tuition Moderate impact for moderate cost, based on moderate evidence.	●●●●●	●●●●●	+4
Moderate impact	Digital technology Moderate impact for moderate cost, based on moderate evidence.	●●●●●	●●●●●	+4
Moderate impact	Behaviour interventions Moderate impact for moderate cost, based on moderate evidence.	●●●●●	●●●●●	+3
Moderate impact	Parental engagement Moderate impact for moderate cost, based on moderate evidence.	●●●●●	●●●●●	+3
Moderate impact	Individualised instruction Moderate impact for very low cost, based on moderate evidence.	●●●●●	●●●●●	+3
Moderate impact	Reducing class size Moderate impact for high cost, based on moderate evidence.	●●●●●	●●●●●	+3
Moderate impact	Summer schools Moderate impact for moderate cost, based on moderate evidence.	●●●●●	●●●●●	+2
Moderate impact	Sports participation Low impact for moderate cost, based on limited evidence.	●●●●●	●●●●●	+2
Moderate impact	Arts participation Low impact for moderate cost, based on limited evidence.	●●●●●	●●●●●	+2
Moderate impact	Learning styles Low impact for very low cost, based on limited evidence.	●●●●●	●●●●●	+2
Moderate impact	Extending school time Low impact for moderate cost, based on moderate evidence.	●●●●●	●●●●●	+2
Moderate impact	Homework (Primary) Low impact for very low cost, based on limited evidence.	●●●●●	●●●●●	+2
Moderate impact	Teaching assistants Low impact for high cost, based on limited evidence.	●●●●●	●●●●●	+1
Moderate impact	Performance pay Low impact for low cost, based on limited evidence.	●●●●●	●●●●●	+1
Moderate impact	Aspiration interventions Very low or no impact for moderate cost, based on very limited evidence.	●●●●●	●●●●●	0
Moderate impact	Block scheduling Very low or no impact for very low cost, based on limited evidence.	●●●●●	●●●●●	0
Moderate impact	Built environment Very low or no impact for low cost, based on very limited evidence.	●●●●●	●●●●●	0
Moderate impact	Mentoring Very low or no impact for moderate cost, based on moderate evidence.	●●●●●	●●●●●	0
Moderate impact	School uniform Very low or no impact for very low cost, based on very limited evidence.	●●●●●	●●●●●	0
Moderate impact	Setting or streaming Negative impact for very low cost, based on moderate evidence.	●●●●●	●●●●●	-1
Moderate impact	Repeating a year Negative impact for very high cost, based on moderate evidence.	●●●●●	●●●●●	-4

Early Years Toolkit

All 11 approaches in this Toolkit are listed here. You can filter by average cost, evidence strength or overall impact. These values are approximations to explore further.

Filter Toolkit	Toolkit Strand	Cost	Evidence Strength	Impact Score
High impact for Australia	Communication and language	●●●●●	●●●●●	+3
High impact for Australia	Digital technology	●●●●●	●●●●●	+3
High impact for Australia	Early learning at age 5	●●●●●	●●●●●	+3
High impact for Australia	Early learning experiences	●●●●●	●●●●●	+3
High impact for Australia	Early learning opportunities	●●●●●	●●●●●	+3
High impact for Australia	Early learning settings	●●●●●	●●●●●	+3
High impact for Australia	Early learning environments	●●●●●	●●●●●	+3
High impact for Australia	Parental engagement	●●●●●	●●●●●	+3

All Approaches - Full Toolkit

All 11 approaches in this Toolkit are listed here. You can filter by average cost, evidence strength or overall impact. These values are approximations to explore further.

Filter Toolkit	Toolkit Strand	Cost	Evidence Strength	Impact Score
High impact for Australia	Arts participation	●●●●●	●●●●●	+2
High impact for Australia	Aspiration interventions	●●●●●	●●●●●	+2
High impact for Australia	Behaviour interventions	●●●●●	●●●●●	+2
High impact for Australia	Block scheduling	●●●●●	●●●●●	+2
High impact for Australia	Collaborative learning	●●●●●	●●●●●	+2
High impact for Australia	Digital technology	●●●●●	●●●●●	+2

Scottish Attainment Challenge: Learning & Teaching Toolkit

This is the 150 practice of the international evidence on teaching, all from 150 practice studies.

Filter Toolkit	Toolkit Strand	Cost	Evidence Strength	Impact Score
High impact for Australia	Arts participation	●●●●●	●●●●●	+2
High impact for Australia	Aspiration interventions	●●●●●	●●●●●	+2
High impact for Australia	Behaviour interventions	●●●●●	●●●●●	+2
High impact for Australia	Block scheduling	●●●●●	●●●●●	+2
High impact for Australia	Collaborative learning	●●●●●	●●●●●	+2
High impact for Australia	Digital technology	●●●●●	●●●●●	+2
High impact for Australia	Early years intervention	●●●●●	●●●●●	+2

SUMMA

Summa is a platform for sharing and accessing evidence-based practice. It provides a central hub for educators to find and share high-quality resources and research.

Filter Toolkit	Toolkit Strand	Cost	Evidence Strength	Impact Score
High impact for Australia	Más efectiva	●●●●●	●●●●●	+2
High impact for Australia	Aplicación de prácticas locales	●●●●●	●●●●●	+2
High impact for Australia	Aprendizaje en el aula	●●●●●	●●●●●	+2
High impact for Australia	Aprendizaje Colaborativo	●●●●●	●●●●●	+2
High impact for Australia	Aprendizaje socioemocional	●●●●●	●●●●●	+2
High impact for Australia	Auxiliar de clase	●●●●●	●●●●●	+2

EduCaixa

EduCaixa is a platform for sharing and accessing evidence-based practice. It provides a central hub for educators to find and share high-quality resources and research.

Filter Toolkit	Toolkit Strand	Cost	Evidence Strength	Impact Score
High impact for Australia	REPETIR CURSO	●●●●●	●●●●●	-1
High impact for Australia	REMUNERACIÓN POR DESEMPEÑO	●●●●●	●●●●●	-1
High impact for Australia	APRENDIZAJE SOCIAL Y EMOCIONAL	●●●●●	●●●●●	+2
High impact for Australia	COMPRESIÓN LECTORA	●●●●●	●●●●●	+2



<http://educationendowmentfoundation.org.uk/toolkit/>

What did we learn?

Comparative messages from meta-synthesis are welcomed by policy makers and practitioners

Not everything works as well as people think

Banarama – the within-strand differences are larger than those between strands

Lack of randomization may not introduce as much bias as we suspect

RCTs provide the ‘surveying pegs’ in the educational landscape

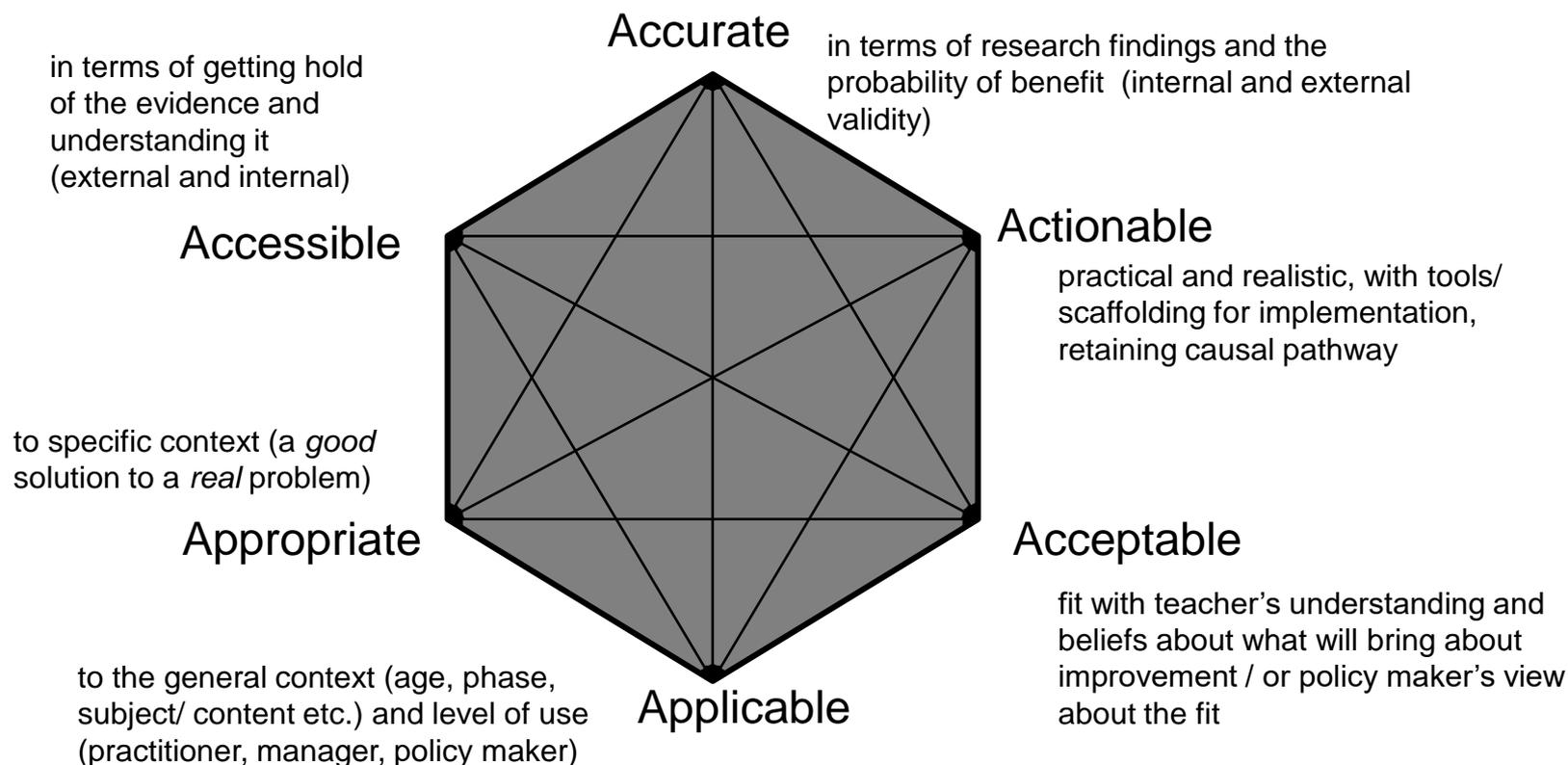
Some of the variation in effect sizes can be explained by aspects of design and measurement

Effect size is a problematic measure



Research and evidence challenges

Research/ provider responsibilities



Current Toolkit challenges

Separate meta-analyses

Different inclusion criteria

Inconsistent quality

Limited to:

- fixed effect averages
- qualitative analysis of moderators
- poor granularity

Technical issues

Effect size comparability (e.g. study designs and measures)

Most moderator analyses (meta-regression) underpowered

Systematic variation associated with: e.g.

- Sample size

- Sample type (e.g. restricted range)

- Age of pupils

- Test type

- Intervention length, intensity, *etc.*, *etc.*

‘Unzipping’ the toolkit

To create a database of impact studies in education so as to:

- develop the EEF Toolkit in terms of its accuracy, applicability and accessibility
- support the wider work of the Education Endowment Foundation
 - Guidance reports
 - Work with international partners

The EEF Education Evidence Database

Phase 1:
'Unzipping' the
meta-analyses

Creation of common inclusion criteria
and data extraction tools
Retrieving/screening studies
Data extraction

Phase 2:
'Back-filling' the
database

Piloting automated screening tools
Identifying similar studies across the
Toolkit strands
Adding new eligible studies to the database

Phase 3:
Creating a
sustainable
source for 'living
reviews'

Automating search and screening tools
Retrieving/screening studies
Data extraction
Semi-automated meta-analytic analysis

Progress to date

28 Toolkit strands 'unzipped'

7,200 reports of studies identified

6,000 studies screened (title and abstract)

4,300 full texts retrieved and screened

1,200 studies coded in EPPI-Reviewer 4

Mapping work to the Microsoft Academic Database
started

Exploratory meta-analyses undertaken

Goals

More comparable meta-analyses for each Toolkit strand

Analyses by school phase and by subject (English, maths, science)

Consistent moderator exploration (pedagogical factors)

Methodological exploration of variation in effect size estimates

	Effect size	Confidence interval	Number of studies	Heterogeneity (I ²)
Peer tutoring	0.39	0.33 to 0.45	128	73.2%
Tutees	0.39	0.28 to 0.50	38	72.6%
Tutors	0.39	0.26 to 0.52	35	79.6%
Reciprocal	0.39	0.30 to 0.47	44	50.6%
Literacy	0.37	0.29 0.45	73	71.2%
Mathematics	0.43	0.33 to 0.53	47	78.6%
Science	0.24	0.01 0.48	6	0.0%

“What works” or “what’s worked”?

Internal validity necessary for external – did it actually work there?

Defining “approaches” or “interventions” – unit of description and causal model

Problematic ‘populations’ – what inference for whom?

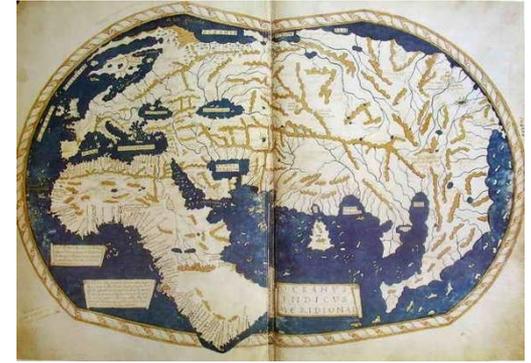
Importance of knowing what hasn’t worked (on average)

Mean or range – “on average” or better estimates of probability?

Sample averages, sub-groups or individuals?

Generalisability or predictability?

How do we improve the ‘mappa mundi’?



Replication, replication, replication

More clearly defined counterfactual

Include other measures of uncertainty (e.g. measurement, attrition, missing data)

Higher bar in areas of effective practice (marginal gains)

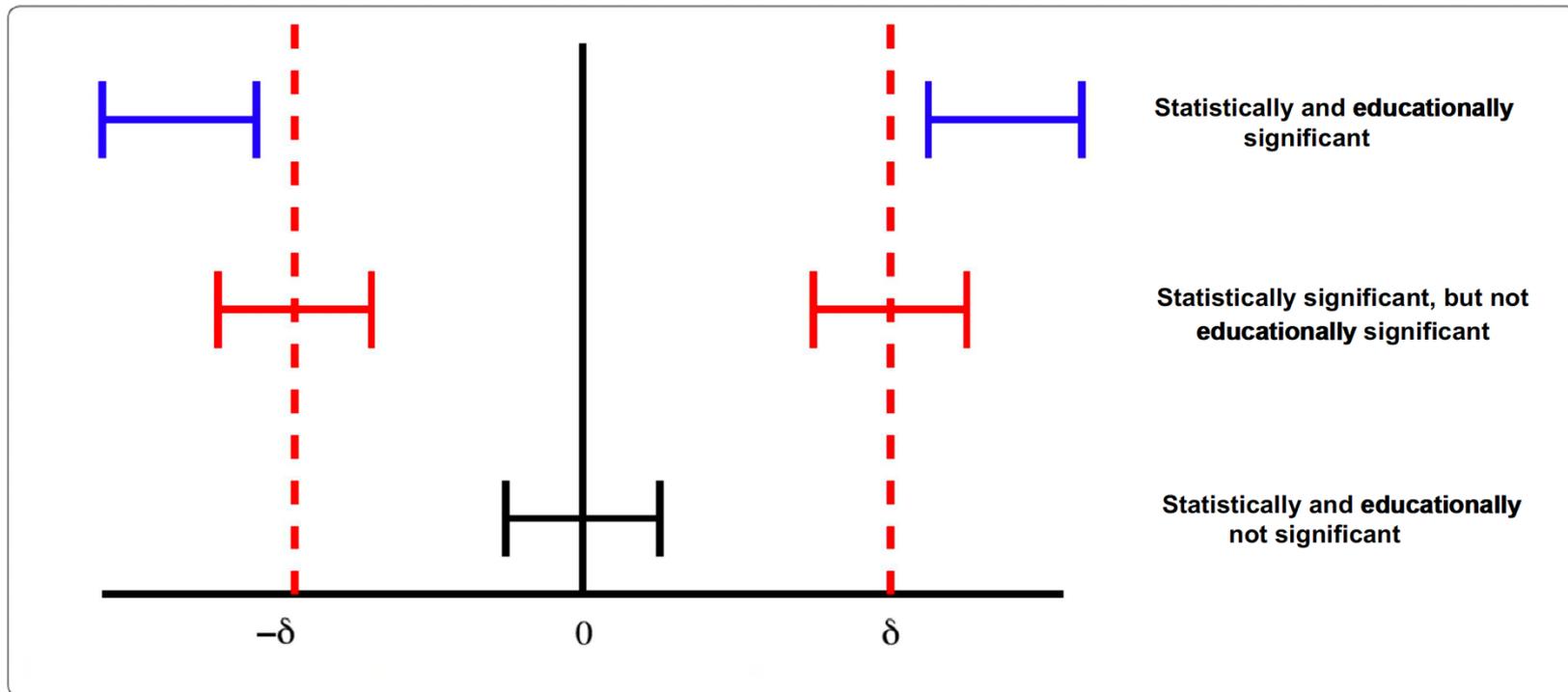
Non-inferiority

Superiority

Equivalence

A higher bar?

Statistical and educational significance evaluated using confidence intervals

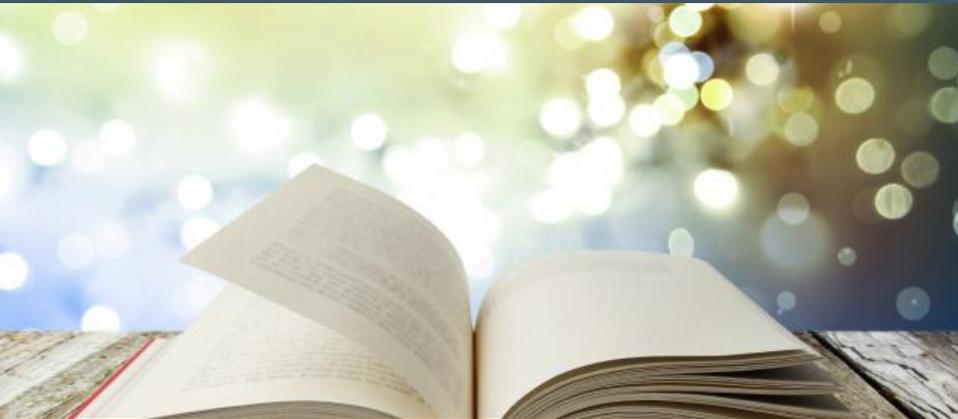


Adapted from Bigirumurame & Kasim 2017

**For every complex
problem there is a
solution that is simple,
neat...
and WRONG!**

H.L. Mencken 1880-1956

‘RCTs – What do they mean for teachers and school leaders?’



**Alex Quigley,
September 2019**

RCTs for school leaders & teachers



Key Conclusions

The following conclusions summarise the project outcome

1. Adopting PBL had no clear impact on either literacy (as measured by the Progress in English assessment) or student engagement with school and learning.
2. The impact evaluation indicated that PBL may have had a negative impact on the literacy attainment of pupils entitled to free school meals. However, as no negative impact was found for low-attaining pupils, considerable caution should be applied to this finding.
3. The amount of data lost from the project (schools dropping out and lost to follow-up) particularly from the intervention schools, as well as the adoption of PBL or similar approaches by a number of control group schools, further limits the strength of any impact finding.
4. From our observations and feedback from schools, we found that PBL was considered to be worthwhile and may enhance pupils' skills including oracy, communication, teamwork, and self-directed study skills.
5. PBL was generally delivered with fidelity but requires substantial management support and organisational change. The Innovation Unit training and support programme for teachers and school leadership was found to be effective in supporting this intervention.



Project-Based Learning

The Innovation Unit

Testing the impact of project-based learning in secondary schools.

Independent Evaluator

Durham University, The York Trials



Unit

Schools	Grant
24	£906,000

Themes

Developing effective learners

Organising your school

Language and literacy



-2

Resources

Evaluation Conclusions

1. The project found no evidence that this version of Lesson Study improves maths and reading attainment at KS2.
2. There is evidence that some control schools implemented similar approaches to Lesson Study, such as teacher observation. This trial might, therefore, underestimate the impact of Lesson Study when introduced in schools with no similar activity. If that is the case, the results suggest that this version of Lesson Study had no impact over and above elements of the Lesson Study approach that were already widely used.
3. Teachers felt Lesson Study was useful professional development, valued the opportunity to collaborate with colleagues in a structured way, and reported several changes to their practice as a result of the programme.
4. Schools generally implemented the programme as the developers intended. Attendance at training was high and most schools implemented one Lesson Study cycle each term.

Lesson Study

University

Lesson Study is a form of collaborative professional development that originated in Japan.

Lead Evaluator

School of Economics



Schools

181

Grant

£543,425

Employment & development



0

Aiding 'best bets'

Texting Parents

This project involved text messages being sent to parents using school communications systems, such as Schoolcomms. Texts informed parents about dates of upcoming tests, whether homework was submitted on time, and what their children were learning at school.



EEF Summary

We funded this project because **existing evidence** suggests that engaging parents in their children's education can have a positive impact on pupil outcomes. A study in the United States found evidence that texting information to parents about children's attendance and homework submission records was successful in increasing their attainment.

This evaluation found a small positive impact on mathematics attainment and on decreasing absenteeism. While this result was small, the cost of sending texts parents is very low (a maximum of around £6 per pupil per year averaged over three years) making the intervention highly cost-effective.

Texting Parents
Bristol University and Harvard University

★ promising project

Using text message prompts to improve parental engagement and pupil attainment.

Independent Evaluator
Queen's University Belfast

Pupils	Schools	Grant
15697	34	£532,620

Themes

- Parental engagement
- Organising your school

Parent Academy

The Parent Academy was a series of classes for pupils' parents, designed to improve the English and mathematics attainment of pupils in Years 3 to 6 in English primary schools. Parents were offered the opportunity to participate in 12 Parent Academy classes, 6 on English and 6 on mathematics, delivered fortnightly by tutors with teaching qualifications and experience of teaching adults. The programme also included an educational family trip.

The evaluation used a two-arm randomised controlled trial to test the efficacy of two versions of the intervention. In the first version, parents were incentivised to attend with a payment of £30 per session and in the second version they were not. Children of both groups of parents were compared with a similar group whose parents were not offered Parent Academy. Sixteen schools in two urban local authorities took part in the trial. A total of 2,593 children were involved. The project also included a process evaluation which assessed how the intervention was delivered and reported on its perceived benefits. The intervention was developed by the University of Chicago. It was not manualised and involved the development of a new adult learning course. The intervention and evaluation were funded by the Education Endowment Foundation and the KPMG Foundation. The trial took place between September 2014 and July 2015 with classes delivered between November 2014 and June 2015.

Parent Academy
Chicago University

A programme which equips parents with the skills to support their children to learn.

Independent Evaluator
NatCen

Schools	Grant
14	£991,400

Themes

- Parental engagement
- Organising your school

Resources

- Executive Summary**
25th July, 2018 - Project/EEF_parenting-...

Necessary conditions for successful implementation

Based on the information gathered in the process evaluation observations, interviews and survey process questions, those schools who appear to have had the most successful local implementation of the RISE programme had in place the following key conditions:

A Research Lead who had developed strong relationships within the school, both with senior leaders and 'ordinary' teachers. Those with these pre-existing positive relationships commanded more respect and had a stronger platform for asking for colleagues to consider research evidence and try changes to their practices.

Active and visible support of the Headteacher for the principles behind the RISE programme. When such support was in place, Research Leads were given a higher profile to challenge normative practices and to carry out local innovation and evaluation. In these situations, greater resources were also present for the Research Lead's role, e.g. paid time away from teaching responsibilities, a budget for evaluation etc.

Following on from this, **additional ring-fenced time to undertake the role of Research Lead** considerably improved the ability to turn the RISE training into local action.

As well as support from the school Headteacher, **it was beneficial when there was a strong link between the Research Lead and the school's Teaching and Learning Co-ordinator** – such co-operation allowed for greater research lead impact and input with CPD.

To enact local change, Research Leads found it easier when they had **a solid understanding of their school attainment data**. Those that were familiar with interpreting such data and had a global overview, found it easier to construct local research-based strategies to target issues of greatest local concern.

A lever for change

UCL INSTITUTE OF EDUCATION



Dos and Don'ts of attainment grouping

To cite this resource: Francis, B., Taylor, B., Hodgen, J.,
Tereshchenko, A. & Archer, L. (2018).
Dos and don'ts of attainment grouping.
London: UCL Institute of Education.

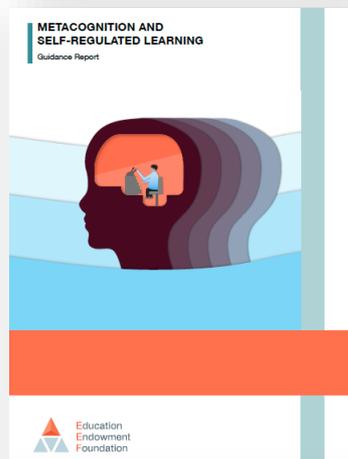
This work was supported by a grant from the Education
Endowment Foundation

You can download a copy of this resource from our website:
www.ucl.ac.uk/ioe-groupingstudents

Enquiries regarding this resource should be sent to:



RCTs – part of a rich evidence picture



Step-by-step – Quiz Starters

1 Prepare a bank of quiz questions in advance, for use across the unit.

Tip... Questions can be re-used so a bank of 30-50 should last the whole unit.

2 Questions can be focused on facts, vocabulary, or ideas.

What is the opposite of a galaxy?

Write a sentence including "system", "moon" and "planet".

Which is bigger? The sun or the moon?

3 At the start of each lesson, present students with 5-10 questions.

4 Students should answer questions individually, without checking their notes.

Key Idea Questions should require students to recall key knowledge that you want them to remember.

Planning for better - informed use of evidence...

“Research can never replace professional experience and teachers’ understanding of their schools and students. But it can be a powerful supplement to these important skills. Used intelligently, evidence is the teacher’s friend.”



Sir Kevan Collins, EEF



Thank you

Alex.quigley@eefoundation.org.uk

www.educationendowmentfoundation.org.uk

CELEBRATING

100
YEARS

of RCTs in education

PANEL

Improving the quality of education RCTs



@TheNFER

@RoyalStatSoc



www.nfer.ac.uk

www.rss.org.uk

#EducationRCTs100



NFER

National Foundation for
Educational Research



ROYAL
STATISTICAL
SOCIETY

DATA | EVIDENCE | DECISIONS