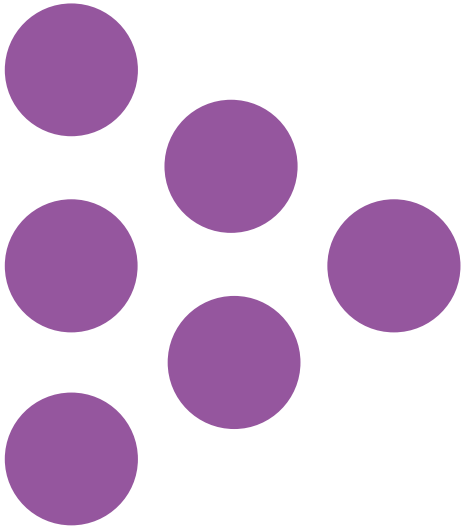
NFER
National Foundation for Educational Research

Experimental educational research: rethinking why, how and when to use random assignment

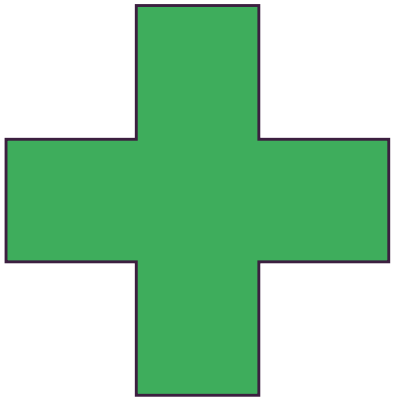
RCTs in the Social Sciences 2025 Valedictory Conference



Sims, S., Anders, J., Inglis, M., Lortie-Forgues, H., **Styles, B.**, & Weidmann, B. (2022). Experimental education research: rethinking why, how and when to use random assignment. cepeowp23-07r1.pdf (ucl.ac.uk)

13 years of English education RCTs

NFER
National Foundation for Educational Research



13 years of English RCTs

- Evaluation planned in
- Unbiased estimates
- Pre-specification
- Open science
- Grist for the toolkit mill

- Distal outcomes
- Effect sizes
- Months progress

- Low effect sizes
- Often uninformative
- Complex causal chains
- School randomised
- Recruitment
- Follow-up testing

The problem

Precision weighted mean effect size from experimental research on broad outcome measures is

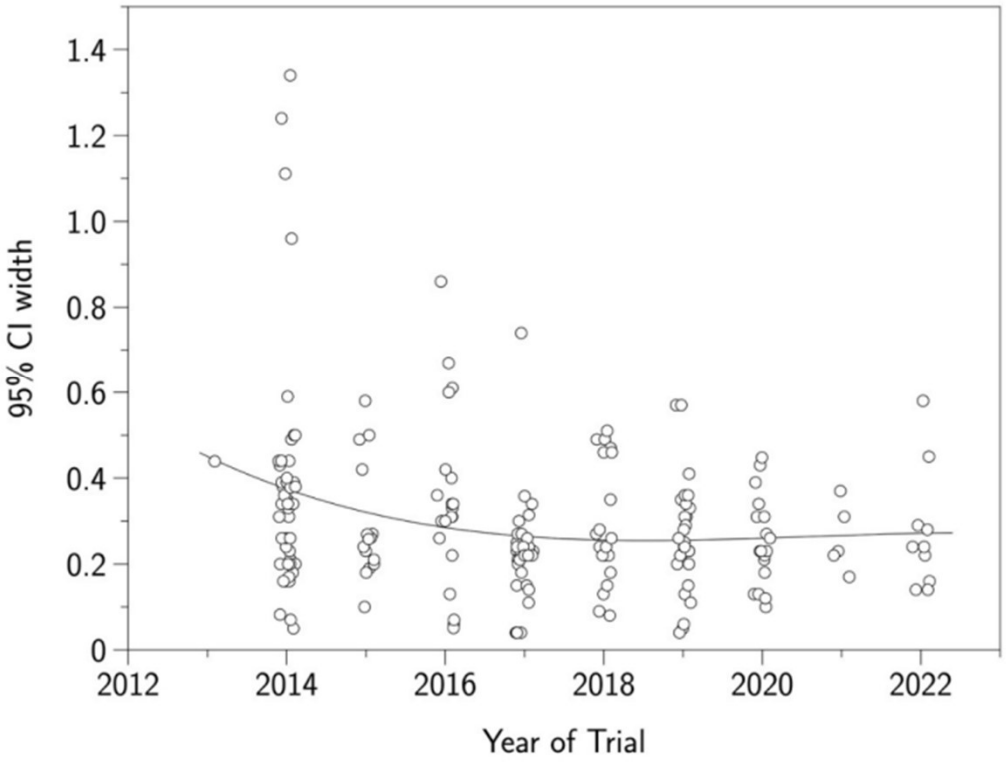
0.04

Average 95% confidence interval half-width is

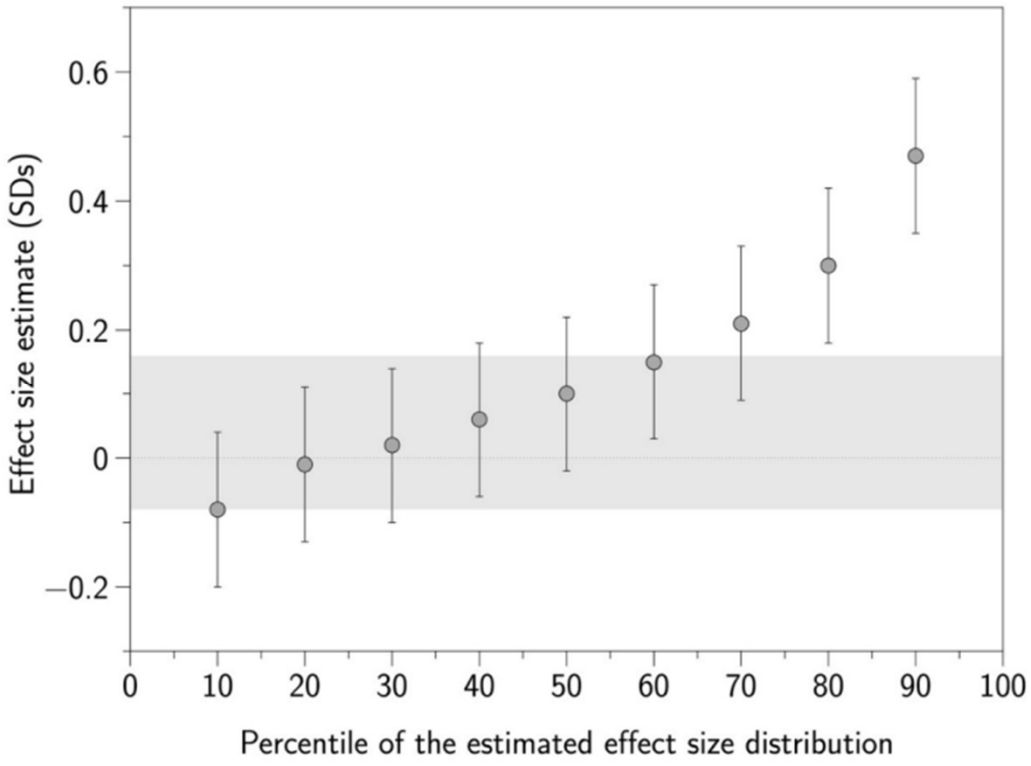
0.12

Lortie-Forgues and Inglis, 2019

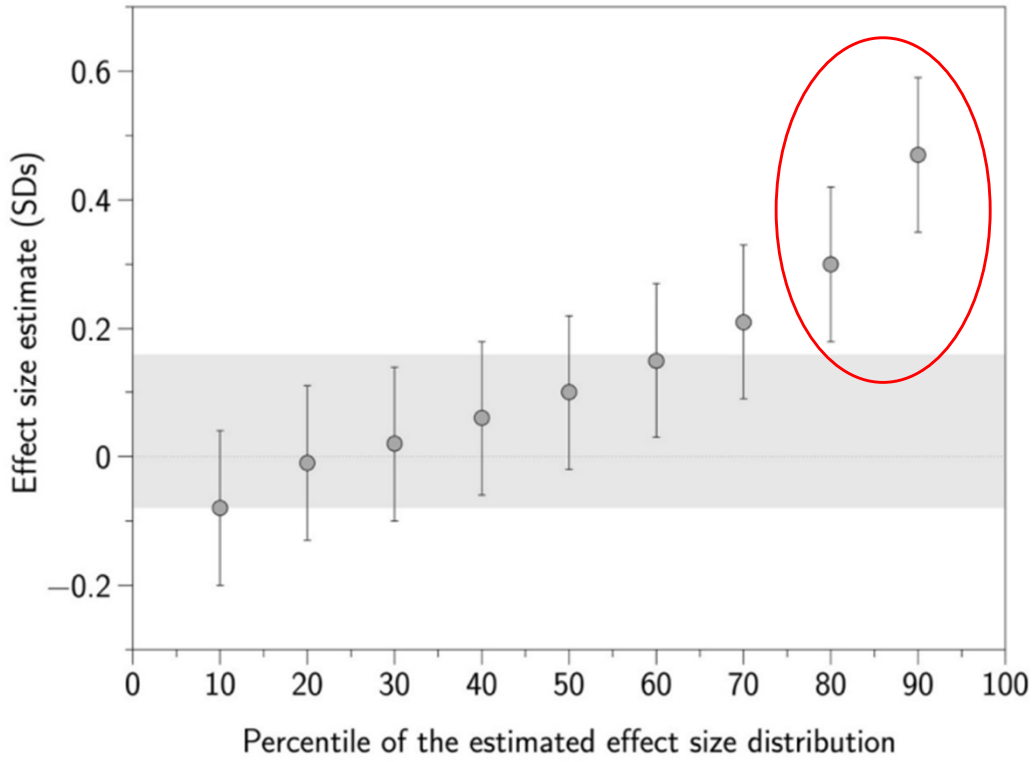
Precision


95% CI width
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
2012
2014
2016
2018
2020
2022
Year of Trial

Decision-informing?



Decision-informing?



Sims, S., Anders, J., Inglis, M., & Lortie-Forgues, H. (2022). Quantifying “Promising Trials Bias” in Randomized Controlled Trials in Education. Journal of Research on Educational Effectiveness, 1-18.

Theory- or decision-informing?

NFER
National Foundation for Educational Research

	Experiments informing theory	Experiments directly informing decisions	
	Ideal	**Common compromise**	**Ideal**
Aims (*why*):			
Testing:	Deduced hypotheses	Interventions / Policies	Interventions / Policies
Quantities of interest:	Confidence interval / *p* value	Sample Average Treatment Effect	Population Average Treatment Effect
Methods (*how*):			
Sampling:	Purposeful	Convenience	Representative
Dep. Var.:	Relevant construct	Socially desirable	Consequential outcome
Instruments:	Maximally valid	Narrow/proximal	Broad standardized
Implementer:	Researcher	Developer	Educator

https://repec-cepeo.ucl.ac.uk/cepeow/cepeowp23-07r1.pdf

OUTCOMES

…Slavin (2019) recommends that the results on researcher designed outcome measures should never be emphasized in research reports. Citing these arguments, the Institute of Education Sciences (IES) recently announced its intention only to fund experiments that use standardized tests as outcome measures, on the grounds that “without common measures, we have little ability to look across interventions for what works and what is most cost effective” (Schneider, 2020, p.1)

Theory building example

Laboratory study of ‘self-explanation prompts’ to improve undergraduate understanding of mathematical proofs (Hodds et al, 2014)

Researcher-designed comprehension test

$d = 0.95$, $p < 0.001$

Decision informing example

Embedding Formative Assessment (Anders et al, 2022)

Attainment8

g = 0.10 (-0.01, 0.21)

Something in-between

Interleaving (Rohrer et al, 2019) of mathematics assignments

Assessment covered exact content in the blocked/interleaved sessions

$d = 0.83$, $p < 0.001$

Nuffield Early Language Intervention

Language Skills score derived from four tests closely aligned with the intervention

A feature of Early Years trials

d = 0.27 (0.07 – 0.46) (Sibieta et al, 2016)
d = 0.26 (0.17 – 0.35) (Dimova et al, 2020)

One of *just two* EEF trials to have shown evidence of impact in both an initial efficacy trial and a second effectiveness trial.

Helping Handwriting Shine

Fine motor control of handwriting

Theory of Change reasonably linear / straightforward

Broad assessment of general writing skills

6-7 year-olds g = -0.02, p = 0.77

9-10 year-olds g = 0.12, p = 0.16

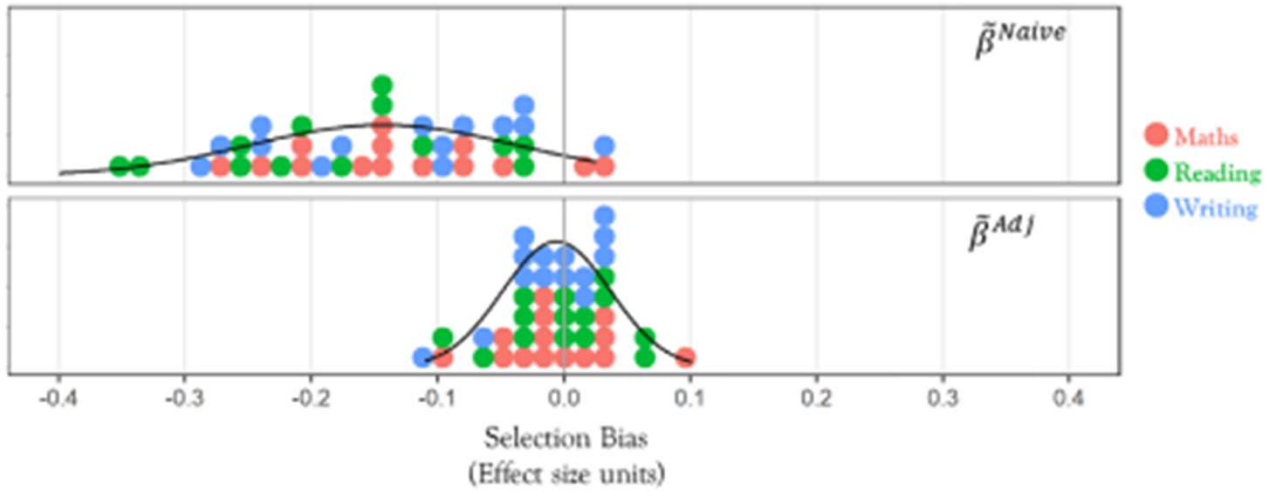
BUT

Handwriting Speed Test

PRECISION

QEDs can be unbiased

Weidmann and Miratrix (2020)



- Selection into programme was at the school level
- Naïve comparison group compared with randomised control group
- Matched comparison group compared with randomised control group

National Tutoring Programme

- Teachers selected pupils with guidance to prioritise Pupil Premium

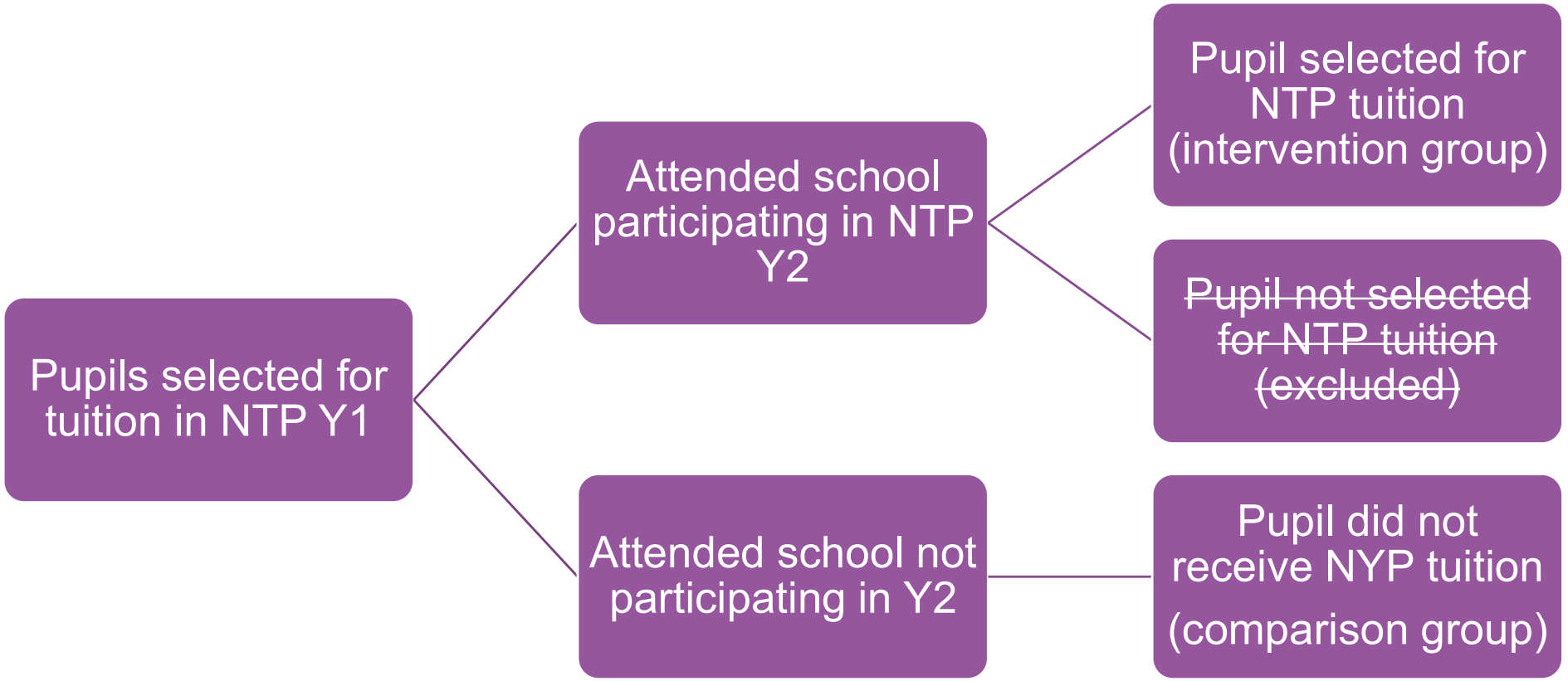
Which group to select to create an unbiased comparison between NTP and comparison schools?

- Pupil Premium?
- Prior Low Attainment?

All such comparisons resulted in effect dilution.

National Tutoring Programme

Reducing selection bias within pupil-level analysis



Selection bias still lingered.

Evaluation problem unsolved

- Weidmann and Miratrix (2020) needs to be repeated for pupil-level selection.
- It almost certainly won't work: most variance in education test scores is between pupils.
- Randomisation remains the only solution here.

Multi-site trials

- Treatment effect heterogeneity across schools can be large
- Only a tiny number of representative samples (US)

Pupil-randomised trials

- Interventions are targeted at individual pupils
- Much easier to power adequately
- Implementation is easier

RCTs with pupil-level randomisation had even higher weighted mean effect sizes compared with CRTs with school-level randomisation (Demack et al, 2021).

Effect size by level of randomisation primary ITT attainment outcomes

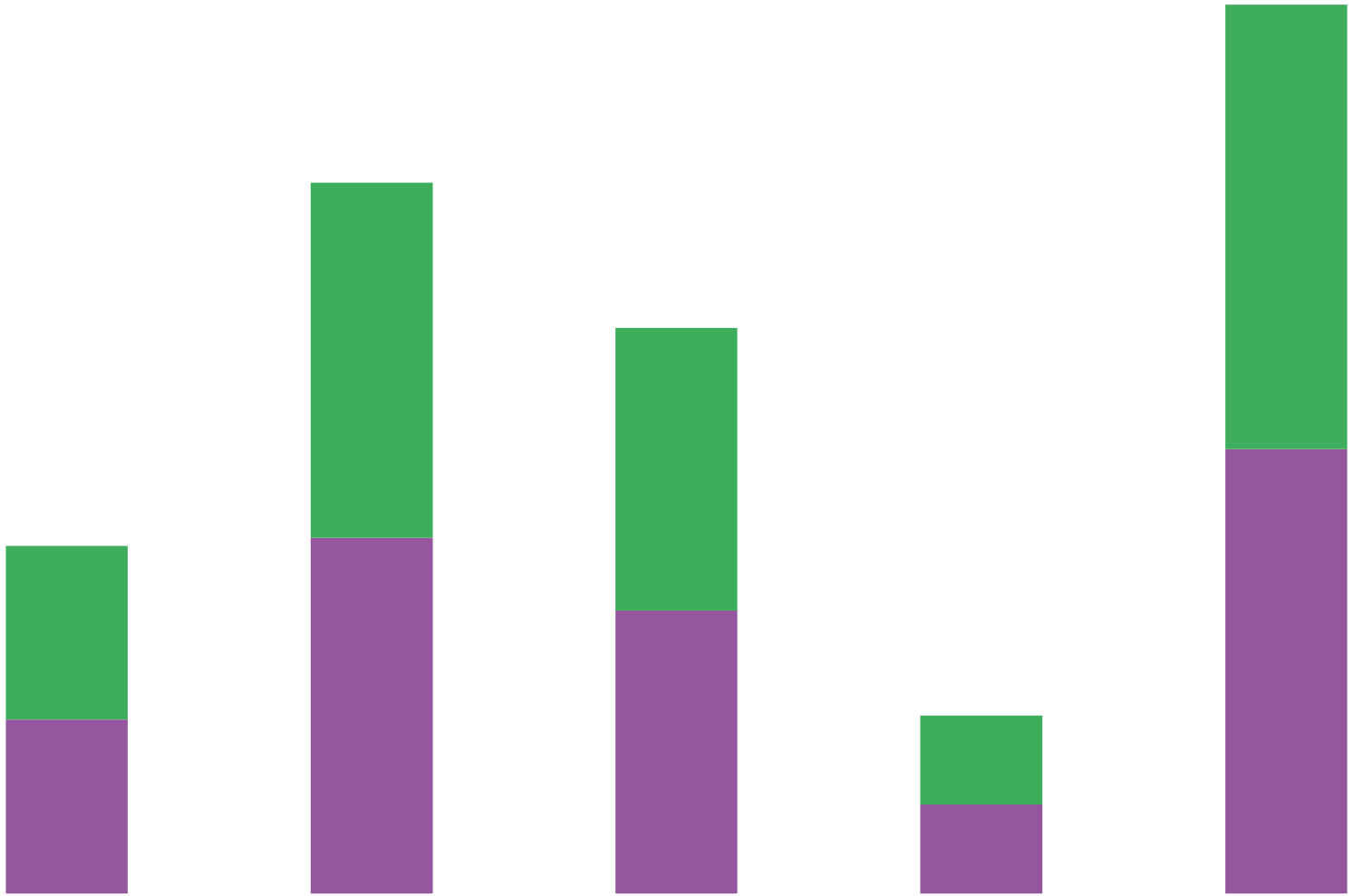
Pupil n=34 +0.12 (+0.06, +0.17)

Terminology

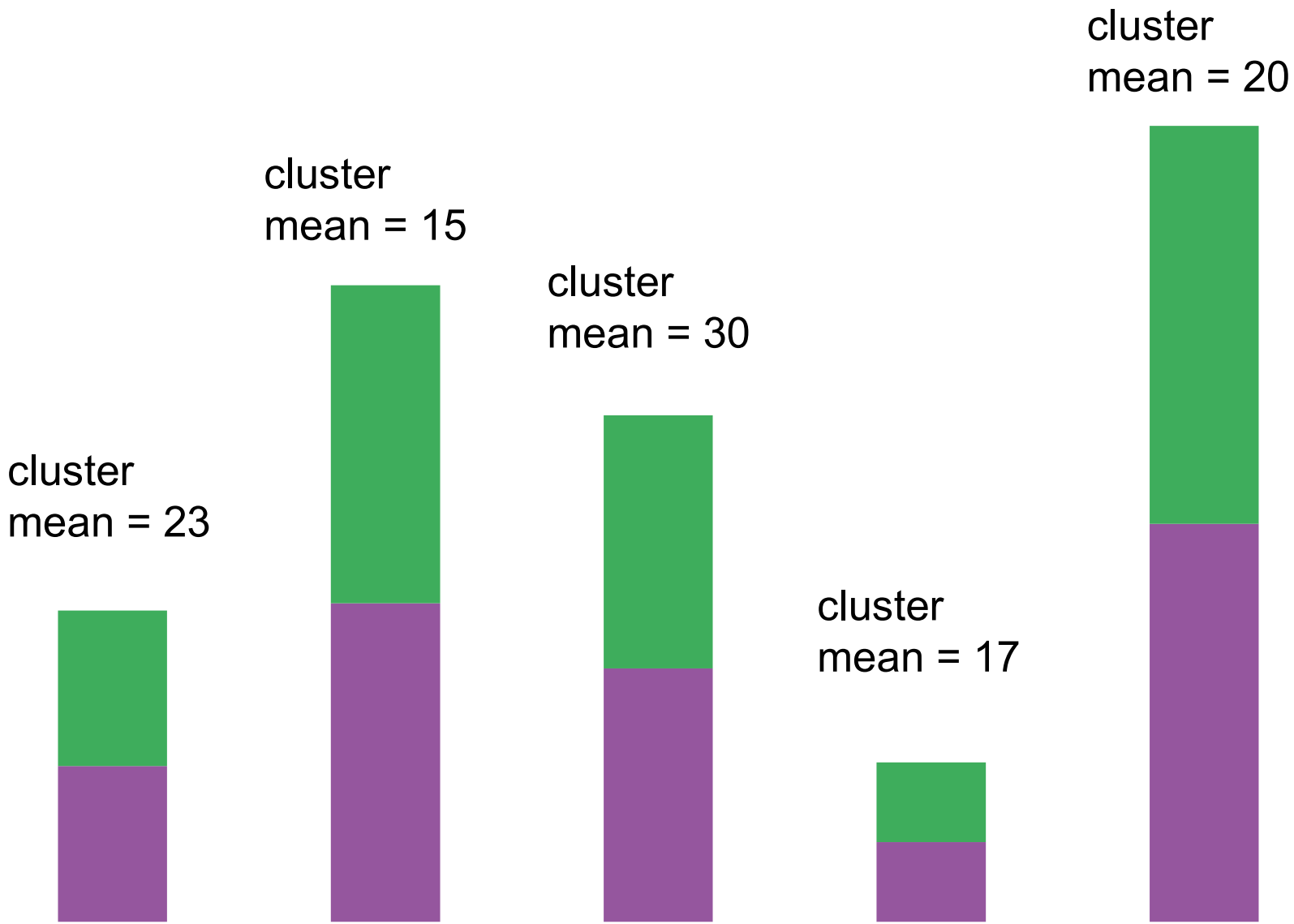
Multi-site = randomised block = pupil randomised across several schools

The best design – if no contamination

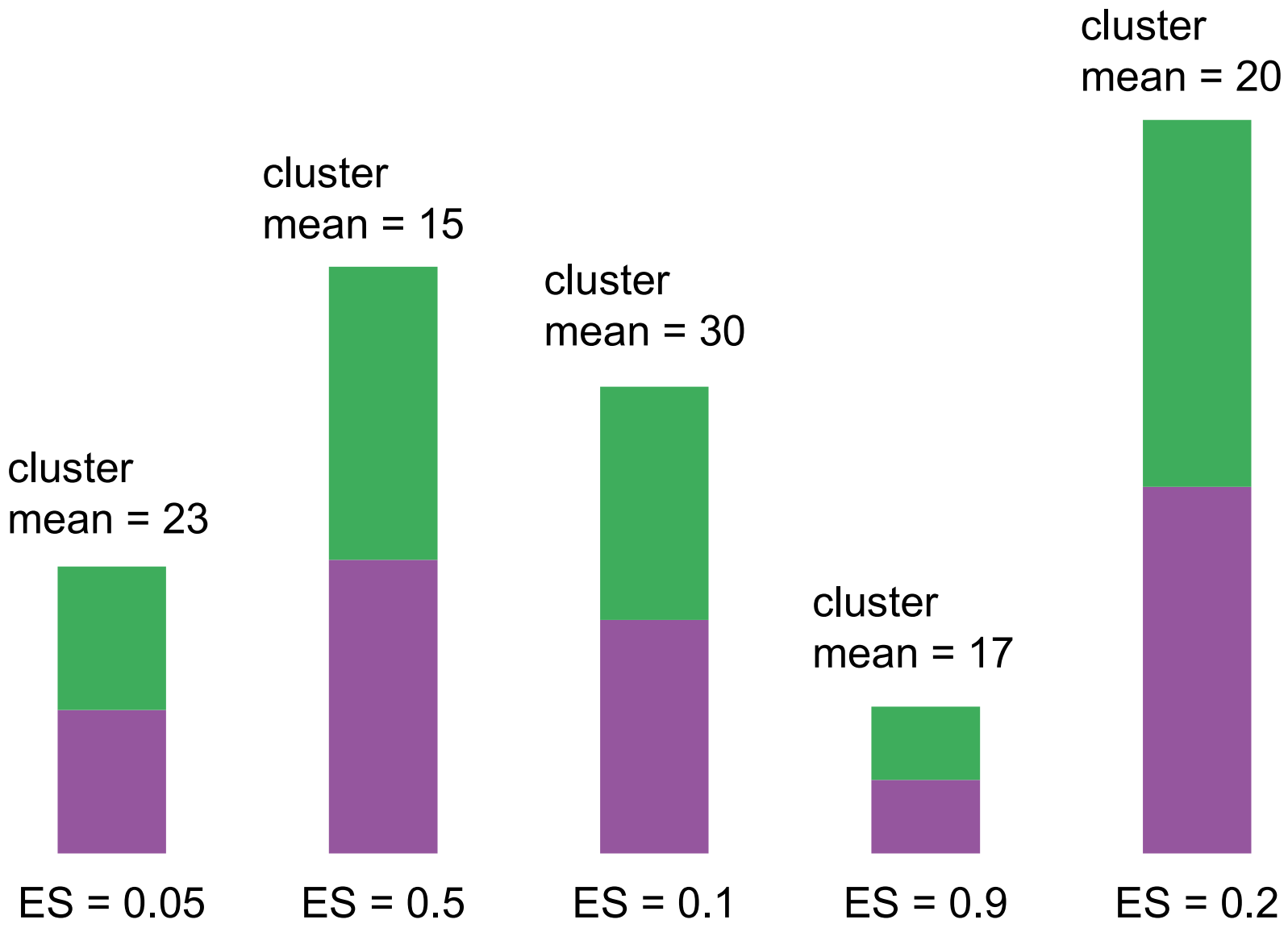
Randomised block design



Randomised block design



Randomised block design



Random site-by-treatment interaction – sample size

Coefficient for intervention in multi-level model is random at cluster-level

Hedges' ω see: http://ies.ed.gov/ncser/pubs/20103006/pdf/20103006.pdf

$$w = \frac{variance\ due\ to\ treatment\ by\ site\ interaction}{total\ variance\ due\ to\ site}$$

Modelling the site-by-treatment interaction reduces power

Effectiveness Trial of 1stClass@Number1

Extract from Statistical Analysis Plan:

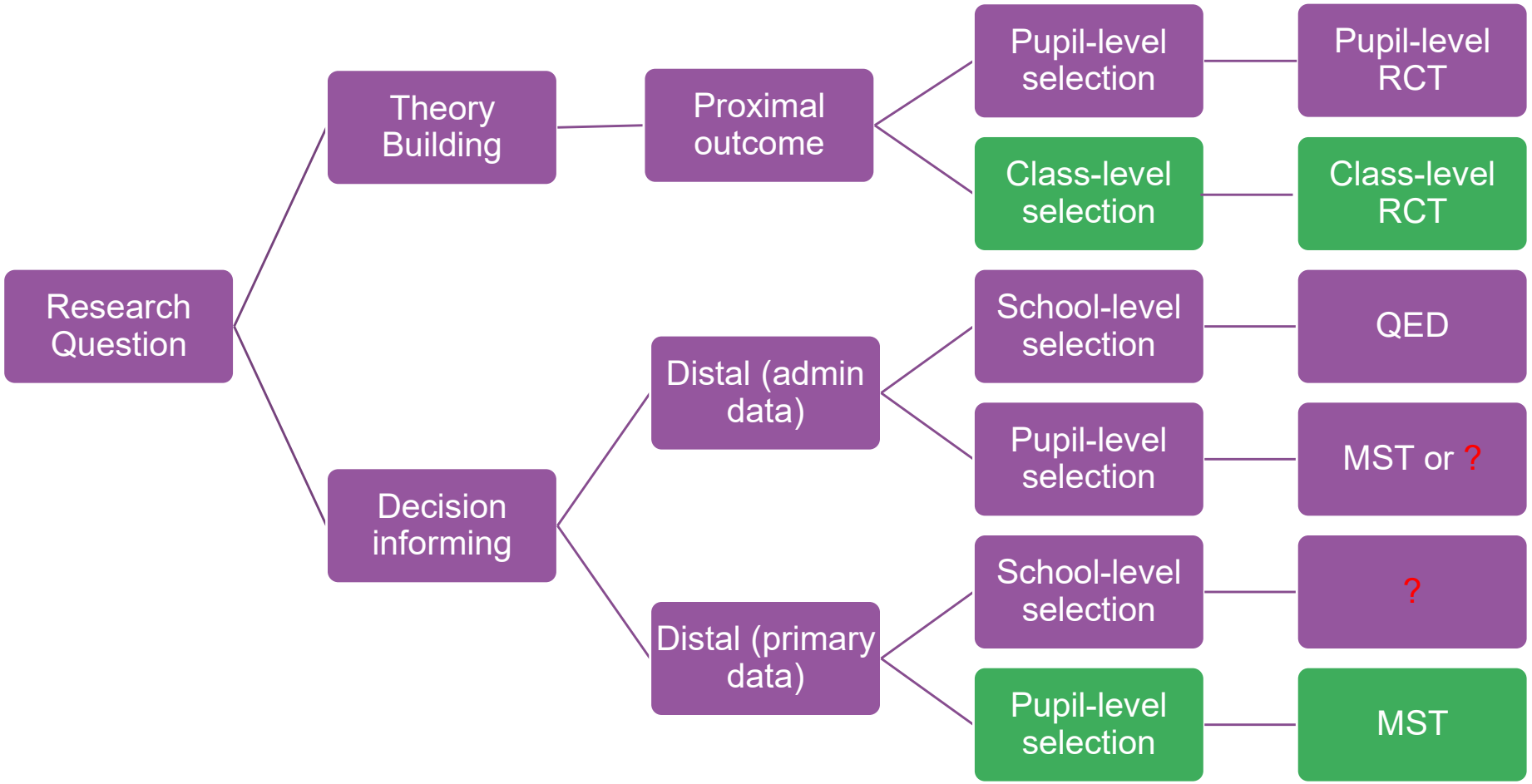
A limitation of this approach is that it assumes the impact of the intervention is the same for all schools (Feaster, Mikulich-Gilbertson and Brincks, 2011). If the assumption is not met, results may not be generalisable to schools outside the trial. To assess the sensitivity of the primary analysis to this assumption, the following model will be calculated:

$$QRT_{ij} = \beta_0 + \beta_1 intervention_{ij} + \beta_2 QRT\_BL_{ij} + \beta_3 FSM_{ij} + b_{0j} + b_{1j} intervention_{ij} + \epsilon_{ij}$$

This adds a random slope $\mathbf{b1j}$ to the primary analysis model, representing the effect of the intervention varying from school to school.

SUMMARY

How to evaluate impact



How Teacher Choices helps

- Easy to recruit to (pre-pandemic)
- Cost-neutral
- Shorter causal chain
- Randomisation below level of school
- Follow-up testing can be easier
- Proximal outcomes
- Positive results?

Acknowledgements

- Co-authors: Sam Sims, Jake Anders, Matthew Inglis, Hugo Lortie-Forgues & Ben Weidmann
- Education Endowment Foundation
- Nuffield Foundation
- NFER Research Operations
- NFER Statisticians
- NFER Researchers

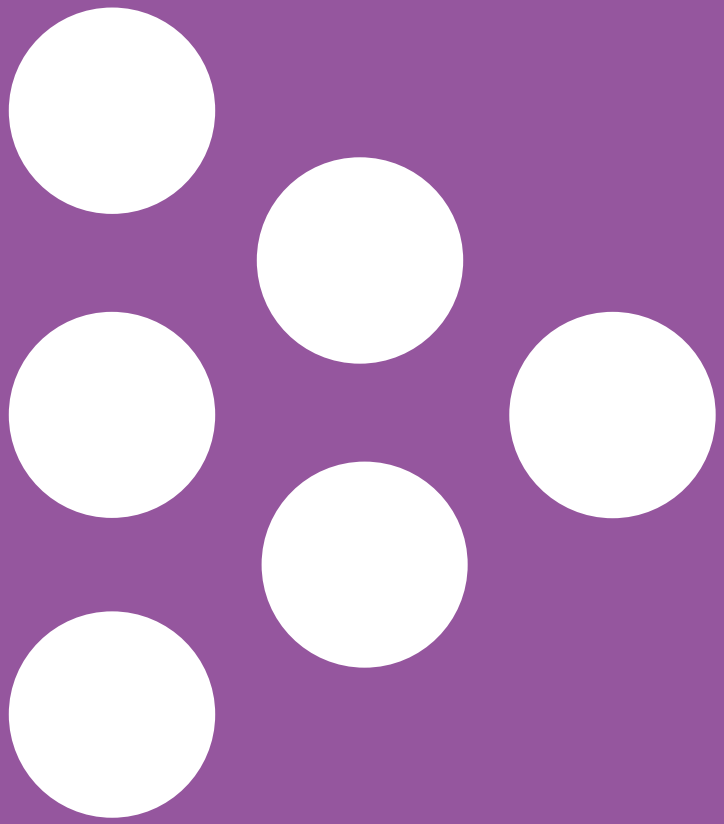
….and the RCTs in the Social Sciences conference!

Society for Research on Educational Effectiveness

NFER
National Foundation for Educational Research

- European Affinity Group
- Speak to me during the conference or email b.styles@nfer.ac.uk

NFER
National Foundation for Educational Research



Evidence for excellence in education

The Mere, Upton Park, Slough, Berks SL1 2DQ
T: +44 (0)1753 574123 • **F:** +44 (0)1753 691632 • enquiries@nfer.ac.uk

www.nfer.ac.uk