Does the classroom level matter in the design of educational trials? A theoretical & empirical review.

DEMACK, Sean <http://orcid.org/0000-0002-2953-1337>

EEF Research Paper

Education Endowment Foundation

No. 003
May 2019

Does the classroom level matter in the design of educational trials?
A theoretical & empirical review.

Author:

Sean Demack

Author:
Sean Demack, Sheffield Institute of Education, Sheffield Hallam University
Email: S.Demack@shu.ac.uk

Does the classroom level matter in the design of educational trials? A theoretical & empirical review.

Sean Demack, Sheffield Institute of Education, Sheffield Hallam University.

Summary & Recommendations

This review was commissioned by EEF in 2018 in order to explore the implications of including or ignoring the clustering of pupil-level data within-schools at a class (or teacher or TA) level. The specific focus is on the theoretical and empirical implications for the design of Clustered Randomised Control Trials (CRTs) in educational settings within the English education system. The geographical location is important because of the heavy use of within-school academic selection in England . Setting/streaming creates within-school attainment clusters of pupils which act in conjuction with other things (such as 'the teacher') to make pupils within one class 'similar' to each other. This class-level clustering will be a structural reality in the data of many educational trials regardless of whether it is acknowledged in the research design. Specifically, this will relate to trials with 2+ classes of pupils per year group / school that aim to measure the impact of an education programme in terms of pupil-level gains in attainment.

From the examination of equations used to estimate the minimum effect size that could be detected with a specified level of statistical significance and power, it is shown that ignoring within-school clustering has clear methodolgical implications for the design of educational trials (class level clustering reduces statistical precision but this can be ofset through class level explanatory power). Some empirical patterns are examined but these need to be treated with some caution a because of the limited number of studies that have included a within school (class, teacher, TA) level in their design and analyses[1]. It seems likely that one key reason for the limited number of studies that have included a within-school level is the complexity and other practical problems that collecting class-level detail brings.

Simply ignoring class-level clustering does not make the problems disappear but will bring hidden bias into the design of educational trials and therefore increase the risk of making incorrect conclusions about the impact of educational interventions.

[1] See section 2. Amongst completed (published) trials, one Dfe funded and three EEF funded 3-level CRTS that randomised at the school level and included levels at class (or teacher/TA) and pupil levels are identified. A further three ongoing EEF funded maths trials were identified.

The review findings are summarised below. Whilst the theoretical implications are clear, the limited number of empirical studies means that the empirical summary points below are purposively general. More specific (and numeric) detail can be found in the body of the report.

- Theoretically, if within-school clustering exists but is not accounted for in the design and analyses of CRTs this can dramatically and negatively impact on trial sensitivity and statistical power.
- To distinguish a within-school level in a CRT design, there needs to be 2 or more within-school clusters (classes, teacher etc) per school. In single-form entry schools, the class and yeargroup clustering will be identical and possibly indistinguishable from the school level (if only one year group is involved).
- The strength of impact of within-school clustering on sensitivity and power depends on how strong this clustering is.
    - From the (very few) empirical studies and known patterns of 'ability' grouping in the wider English education system, class-level clustering is likely to be strongest in Secondary compared with Primary schools. Additionally, clustering is likely to be strongest in maths compared with other subject areas.
    - Whilst class-level clustering in Primary schools is generally weaker than Secondary schools, it should not be regarded as negligible (and therefore ignorable); particularly at KS2. In KS2 maths, Dracup (2012) found a quarter of Y5 and a third of Y6 maths lessons were taught in 'ability' groups. Whilst this is some way from Y11 maths in Secondary schools (74% taught in 'ability' groups), it does illustrate how clustering of attainment data is likely to be of impotance for trials used to evaluate interventions involving pupils between Y5 and Y11 in England.
    - In the discussion section of this report, some suggested class-level ICC values are provided for future cluster trials. ICC values for different school phases, key stages and subject areas are suggested. These will need to be revised as future 3-level CRTs are published.
- The use of pre-test covariates can notably improve statistical sensitivity and power, mitigates the effects of strong clustering. Given that 'ability' grouping policies of schools usually draw on pupil attainment, it seems likely that the explanatory power at the classroom level will be very high if attainment is the outcome (and pre-test). If it is reasonable to assume that the outcome will clustered at the class level, an outcome-only analysis that did not include any pre-test covariate explanatory power should be avoided.

- To avoid multicolinearity the pre-test covariates should be appropriately centred as specified by Hedges & Hedberg (2013): pupil data centred around their class-level mean; class means around around their school-level mean and school means around the overall (grand) mean.
- Collecting class/teacher identifiers at the start of a trial alone assumes that these pupil groupings will remain entact over the course of the trial. The 'Intention to Treat' (ITT) analyses will be based on this sample. There may be pupil movement (e.g. introduction of ability grouping; movement between groups) and teacher complexities (movement between classes, multiple teachers, use of TAs) that would not be captured with a single (trial start) data collection. By collecting this detail on multiple occasions, the integrity of the ITT sample can be examined[2]. Therefore, it seems important to at least collect class/teacher level detail at both start and end of trials (and possibly at time points inbetween).

The limited number of studies from which to view the empirical realities of class-level clustering (and how this is mitigated by covariate explanatory power) leads to the main two recommendation from this review:

- Research is needed to measure the clustering of attainment data at the class level in England across educational phases, year groups, key stages and subject areas - and how this changes over time / pupil cohorts.
- In the meantime …
    - More 3-level CRTs that include a within-school (class/teacher/TA) level of clustering need to be conducted. This is most critical for trials around Mathematics interventions but also seems warranted for other high status subjects such as English and Science. This is also more critical for trials within secondary compared with primary educational phases.

Ideally, a large-scale study would collect class level identifiers for pupils in classes over time for a large, randomly sampled number of Primary and Secondary schools. However, it seems likely that educational CRTs will be conducted regardless of such a reliable source. Further, 3-level CRTs can be used to test out processes for collecting this data that are robust and reliable whilst minimising burden on schools and teachers (which may result in making a larger scale study more feasible).
Three reasons for the need for more 3-level CRTs are summarised below:

[2] Extreme example, all 'mixed ability' classes at the start, two classes (and teachers of classes) selected per school. Setting/streaming introduced across all schools the week after randomisation - leaving original teachers with only a few of their original class pupils (and pupils dispersed across all setted classes).

- To develop methods for collecting detail on within-school clusters from schools that are accuracte, ethical and minimise the burden on schools.
- To build the empirical evidence base on the nature of within-school clustering of attainment (and other) data within the context of the English education system. This seems to be most important to support the design of trials in Secondary schools involving mathematics classes but more detail across educational phases, year groups, key stages and subject areas would also be valuable.
- To directly acknowledge the classroom (or teacher/TA) within trial design and analyses.

Undertaking more 3-level CRTs is likely to require notably more resources than if within-school clustering was ignored (i.e. 2-level CRTs). This is because schools are not regularly required to provide class/teacher level data for external view and so may not have systems that could easily provide it. Additionally, around the GDPR arrival there has been an increased attention and concern around data security and this may make schools less willing/happy to provide it. Further, teachers are currently rather hidden from external statistical scrutiny. Measures of attainment, destinations, exclusions etc. are commonly presented at a school level and can also be accessed (following an NPD request) at the pupil level. Attaching data to the classroom (and associated but not necessarily equivalent teacher) levels may make teachers feel vulnerable (and hence reticent to cooperate). Therefore it seems very important that resource is given to develop approaches for collecting details at a class and/or teacher / TA level that are robust, ethical and do not serve to burdon schools or teachers.

This first reason is not only about developing systems/practices to minimise school and teacher burdon, there are also a number of complexities at the class level that need to be thought through and addressed in future designs. For example, multiple teachers for one class; a single teacher across multiple classes; pupil and teacher movement between classes. These complexities serve to illustrate the distinction between the class and teacher (or TA) levels. The class and the teacher are clearly related but not exactly the same and this points towards a need for cross-classified multilevel research designs (Leckie, 2013) if the teacher and the class levels are to be disengangled.

Second, a greater number of 3-level CRTs are needed to build evidence of how strongly attainment (or other) data is clustered in different subject areas and schooling phases in England. These studies will also provide evidence of the strength of explanatory power for pre-test covariates included at different levels (school, class & pupil). It would also be of value to compare these findings with those found in other

countries to provide structural context and to aid the interpretation of other statistics (e.g. PISA, TIMMS). Tymms et al. (2015) provide an indication of clustering at the class level during the first year of schooling but more detail across the Y0 to Y11 education system is needed - ideally using a random sample of schools. In the meantime, educational trials can be used to help build the evidence base for future research and evaluation.

The final reason for needing more 3-level CRTs is to enable the the potential role of the classroom (and teacher/TA) to be foregrounded within the impact evaluation of educational trials. For example, through an implementation and process evaluation, data on how teachers engaged with an intervention might be collected through teacher surveys and interviews. In many cases, the theorised causal path for an intervention theorises some form of change at a teacher level which leads to a change at the pupil level, often within a classroom setting (For example, teacher Professional Development interventions). Without a class or teacher level in the trial design, to bring this data into follow-on impact analyses it is commonly aggregated (or averaged) to the school level. The two-level design therefore systematically excludes the possibility of exploring the role of the classroom / teacher in follow-on impact analyses. Including a class level allows this IPE data to be attached at the appropriate level and for designs to include within-school (teacher or class level) variation in fidelity to a programme in follow-on impact analyses.

Finally, the practical experiences in collecting details on within-school clustering of attainment data might be drawn on for the design of larger scale observational studies that collected attainment data and class-level identifiers from a large scale random sample of schools. This future research might begin to record the nature of within-school clustering and how this changes over time which would be of value to the designers of educational trials. The widespread use of setting/streaming within the English education system has been highlighted by the OECD as an explanation for the relatively low levels of social mobility in England compared with other OECD countries (OECD, 2012). This makes the nature of class-level clustering in England of great sociological and educational interest; particularly if cluster patterns relating to socioeconomic background, ethnicity, gender, EAL and SEND were open to scrutiny.

Introduction

Cluster Randomized Trial (CRT) designs are inherently multilevel and reflect the hierarchical structure of schools and the wider education system. To capture this multilevel nature, CRTs are commonly analysed using multilevel (or hierarchical) linear models (Goldstein, 1987; Snijders & Bosker, 1999; Raudenbush & Bryk, A. S., 2002). It is fairly common for CRT designs and analyses to include a school and individual / pupil levels but the inclusion of a within-school level (e.g. class or teacher) is rare. This might reflect how complex and time consuming the collection of class level identifiers is in comparison with collecting detail at a school or pupil level. However, given the widespread use of setting and streaming policies in English schools (OECD, 2012; Dracup, 2014; Francis, 2017), it seems reasonable to expect sizable clustering of attainment data at a class-level. Looking at English, Maths and Science lessons in England, Dracup (2014) reported higher incidence of lessons taught in 'ability' groups in Secondary (43%) compared with Primary schools (12%) with higher use in maths in both Secondary (71% overall, 74% in Y11) and Primary (19% overall, 26% in Y5, 34% in Y6)[3]. In addition to setting/streaming it seems likely that clustering effects related to teachers themselves will exist although it is impotant not to consider class and teacher levels as one and the same. This is because of complexities such as teachers moving between classes (e.g. to teach a specific topic), use of multiple teachers and one teacher taking multiple classes. Therefore, for evaluations of education programmes that seek to cause positive change in pupil level attainment via some form of 'change' at the teacher and/or classroom level, the lack of a class-level in the CRT design seems like a problematic omission; particularly for maths evaluations in Secondary schools.

This paper presents an examination of the methodological and empirical implications of not including a class level within CRT designs and impact analyses, and is organised into five sections. The first section briefly introduces the research questions; the second section presents a summary of three level CRTs that included school, classroom and pupil levels; the third section focuses on EEF maths trials given that this subject has the greatest incidence of pupil segregation / setting; the fourth section examines the methodological theory of RCT and CRT designs; the fifth section reflects on the empirical data from the second and third sections from a methodological perspective before reflecting more widely about whether classroom level 'matters' in the design of educational CRTs.

[3] See Table 6 later for a summary of the OFSTED data reported by Dracup showing the percentage of observed lessons that grouped pupils by perceived or measured ability.

1. Research Questions

The first time that I observed class level clustering of attainment data was in 2014 within a DfE funded CRT-centred evaluation of a KS3 Multiplicative Reasoning Project (MRP) that involved 62 English secondary schools (Boylan et al., 2015). The MRP was a Professional Development (PD) programme for maths teachers; the aim was for this PD programme to lead to pupil gains in mathematics attainment during a single academic year (2013/14) across three pupil year groups (in Y7, Y8 and Y9). Whilst teachers received MRP PD training and materials away from the classroom, it was in classrooms where a new approach and materials were used in order to develop pupils mathematical understanding and this was theorised to lead to gains in maths attainment. Therefore, a class level seemed important and the impact of MRP on maths attainment[4] was evaluated using a 3-level Clustered Randomised Controlled Trial (CRT) design (schools<maths classes <pupils) with randomisation at the school level. In this 3-level CRT trial, at the design stage, class level clustering of maths attainment was assumed to be relatively weak compared with clustering at the school and pupil levels. In reality, class-level clustering was found to be much higher than clustering at either school or pupil levels. This had serious implications for the statistical sensitivity of the trial design Specifically, the Minimim Detectable Effect Size (MDES) estimates predicted prior to randomisation (that assumed a class level ICC of 0.05) ranged between 0.24 (Y7) and 0.26 (Y9) standard deviations. However, with the much stronger class level ICCs, the achieved MDES estimates actually ranged between 0.41 (Y7) and 0.48 (Y9) standard deviations (Boylan et al., 2014; Table 4, p28 and Table 5, p31.)

Since the MRP evaluation, I have designed two 3-level CRTs to evaluate the impact of Primary (Y6) school teacher PD programmes[5]. In both of these CRTs, class level clustering was found to be much weaker than was observed with the MRP evaluation. At the time of writing I am working on a new trial that is collecting school and class level data for a CRT centred evaluation of a Secondary maths teacher PD programme.

Given the extent of setting and streaming in English schools (Dracup, 2015; Francis, 2017), it is reasonable to assume that class level clustering will not be trivial. Setting/streaming are likely to result in engraining sizable attainment clustering at the classroom level; particularly in some subjects (Mathematics) and educational phases (Secondary).

[4] As measured by the GL Progress in Mathematics assessments.
[5] Jay, T. et al. (2017); Boylan et al., (2018)

At the same time, collecting data at the class and/or teacher level brings sizable practical challenges and costs and, therefore, perhaps it is best to just ignore clustering at the class level when designing educational CRTs. This paper examines the methodological implications of ignoring class level clustering and draws on findings from 3-level trials and EEF maths trials to provide empirical context in order to explore whether class-level clustering 'matters' in the design of educational trials.

Research questions

To examine whether the classroom-level 'matters' in the design of educational trials, four research questions are posed.

1. How many clustered trials in England have included a school and class level[6]?
2. How many EEF maths trials have been completed?
3. What are the methodological implications of ignoring class level in educational CRT trials?
4. What are the empirical realities of ignoring class level in educational CRT trials?

RQ1 and RQ2 draw on data collected from EEF and through published reports, protocols, statistical analysis plans along with email exchanges from evaluators to summarise key aspects of trials that included both school and class levels (RQ1) and all EEF maths trials to date (RQ2).

RQ1 focuses on summarising trials that have randomised at the school level but also included a class level in their design and analyses. A total of nine CRTs are identified, within which are seven 3-level CRTs that randomised at the school level and included a within-school class (or teacher or TA) level.

RQ2 focuses on RCTs used to evaluate maths programmes. Given experiences with MRP and the higher likelihood of schools using setting/streaming in maths compared with other subjects (Dracup, 2014), it seems likely that class level clustering will be a particular issue for maths trials. A total of 44 EEF trials are identified that have been used to evaluate education programmes with a maths focus and these are summarised in this section.

RQ3 draws on methodological theory (Bloom, 1995 & 2006; Bloom et al., 2007; Hedges & Hedberg, 2013; Kelsey et al, 2017 and Spybrook et al., 2016) for the design of RCTs and CRTs to examine the implications for ignoring class level clustering. First, equations are used to estimate minimum effect size that 2-level CRT and 3-level CRT designs could detect as statistically significant (α=0.05) with a specified statistical

[6] This focuses predominantly on EEF funded trials but also includes one DfE funded trial.

power (commonly 80%; 1-β=0.80). Following this, the equations are used to provide two numerical and visual perspectives on how class level clustering influences the statistical sensitivity and power of trial design.

RQ4 first reflects on the summary tables for RQ1 and RQ2 to discuss what is known about the empirical realities of class level clustering and the methodological implications (from RQ3) that these bring. Second, the findings are reflected on more broadly and discussed within the context of the English education system and complexities/challenges in collecting class level detail for educational trials.

To aid reading more technical aspects of the paper, a brief glossary of terms is provided.

Glossary

- **Effect Sizes:**

Effect sizes are standardised measures of the mean difference between two groups in units of standard deviations that are commonly used to report the impact of an intervention in educational trials. Standardising effect sizes is useful for comparing findings across many trials using the same scale/units. There are many different effect sizes but the one most commonly used in EEF trials is Hedges g (Rosnow & Rosenthal, 2003; Shagen & Elliot, 2004).

- **Minimum Detectable Effect Size (MDES)**

The MDES is the smallest effect size that, if true, can be detected with a specified level of statistical significance and statistical power (Bloom, 1995).

- **Statistical Significance**

Statistical significance is a measure of the probability of making a Type I (or false positive) error. Specifically, this is the probability of concluding that the observed difference between two groups (in a randomised trial or sample) is 'true' when in fact it is 'false'. This is commonly denoted as alpha (α) and set at a level of 5% or less ($\alpha \leq 0.05$). In this case, an estimated effect is declared statistically significant where the probability of observing an effect at least as large, under the null hypothesis, is at 5 per cent or less.

- **Statistical Power**

Statistical power is a measure of the probability of correctly concluding a difference between two groups and is measured by focusing on minimising the probability of making a Type II (or false negative) error. Specifically, this is the probability of concluding that the observed difference between two groups is 'false' when in fact it is 'true'. This (false negative) error is commonly denoted as beta (β) and set at a level of 20% or less ($\beta \leq 0.20$). Statistical power is calculated by subtracting this probability from 1 (Power = 1-β).

- **Clustering**

Clustering is when data are 'gathered' (or grouped) rather than being randomly distributed at the individual level. In educational trials, clustering might be at a school level. For example, the attainment of pupils within a school might be closer to the mean attainment for their school compared with the overall (grand) mean. One way of examining clustering is to look at how the variance in an outcome variable is decomposed at different levels. Variance in an outcome can be decomposed to be at the school level (between schools variance) or at the individual level (within school, pupil level variance). This paper examines clustering effects within schools at the class, teacher or TA levels. For example, within schools, the attainment of pupils in the same maths class might be closer to the mean for their class than the mean for their school.

- **Intra-Cluster Correlation Coefficient (ICC)**

ICCs are a common way of measuring how the variation in an outcome variable is clustered at different levels. Within a 2-level design (school and pupil), the total variance in an outcome can be decomposed into two levels, the proportion of this total variance that is found at the upper (school) level is known as the school level ICC. ICCs can take a value between 0.00 (no clustering at the school level; all of the variance in an outcome is found at the student level) and 1.00 (all clustering is at the school level, none at the student level) and are sometimes expressed as a percentage. Within a 2-level design, the proportion of variance clustered at the individual pupil level can be calculated: $ICC_{pup} = 1 - ICC_{Sch}$. At one extreme, an ICC_{sch} value of 0.00 would mean that, in relation to a trial outcome, variation would only be at the individual pupil level (i.e. $ICC_{pup} = 1$) and all schools would have the same (aggregated pupil) mean outcome score (which would also be equal the overall, grand mean of the trial outcome). At the other extreme, an ICC_{sch} value of 1.00 would mean that, in relation to a trial outcome, variation would only be at the school level and that in each school, all pupils would have exactly the same outcome score. Within a 3-level design (school, class and pupil), the total variance in an outcome is decomposed into three levels and the proportion of this total variance that is found at each level are measured by ICC values. So, ICC_{Sch} + ICC_{class} + ICC_{pup}= 1 (sometimes written as 100%). The (relative) strength of clustering measured by ICCs at one level determines the maximum possible strength of clustering at other levels. For example, when ICC_{sch} =0.10, 10% of the variance is clustered at the school level and therefore the maximum proportion of variance that could be clustered at the class and/or individual level will be 0.90 (90%). (Konstantopoulos, 2007; Eldridge & Kerry, 2012).

- **Covariate Explanatory Power**

Covariate explanatory power is the proportion of variance in an outcome variable that is accounted for by variation in one or more covariates and is commonly denoted as R-square (R^2). Within a multilevel design, explanatory power can be estimated and measured at multiple levels: In a 2-level design, covariate explanatory power might be at the school (R^2_{Sch}) or individual pupil (R^2_{pup}) levels. In a 3-level design, explanatory power might be at the school (R^2_{Sch}), class/teacher/TA (R^2_{class}) or individual pupil (R^2_{pup}) levels.

2. How many clustered trials in England have included a school and class level? [RQ1]

Table 1 summarises six UK educational trials that acknowledged the clustering of outcome data at three levels; school, within-school (class, teacher or TA) and pupil level. Four of the six trials were used to evaluate the impact of interventions in primary schools[7] with three focusing on KS2. Only two trials evaluated the impact of interventions in secondary schools. Five of the six trials were funded by EEF with one trial funded by the Department for Education (DfE). Three trials had a clear English language / literacy focus, two with a maths focus, one had a maths and English and one focused across English, maths and science subject areas.

There are some examples of 3-level educational CRTs from the USA[8] but the focus here is solely UK trials in order to best ensure comparability in terms of the structural (education system) context.

As is apparent, there are very few 3-level trials that have been conducted in England to draw on to provide estimates for class level clustering and covariate explanatory power. Additionally, of the six 3-level trials summarised in Table 1, only four actually used a 3-level CRT design that included school level randomisation. The Grammar for Writing KS2 trial and Teacher Observation KS4 trial include a class (or teacher) level but in both cases this was within a larger, more complex, design. A further three 3-level CRTs with school level randomisation were identified from protocols and SAPs of EEF maths trials still in progress (see RQ2 below). This results in a total of seven 3-level CRTs with school-level randomisation. Therefore caution is needed when drawing on this small group of evaluations for future designs of educational trials.

Table 1 summarises the 3-level CRTs in terms of sample size, strength of clustering, covariate explanatory power and reported effect sizes. Where available, the strength of clustering is shown at three levels (school, class/teacher/TA and pupil levels) and four measures of covariate explanatory power (school, class/teacher/TA, pupil and total) are shown.

[7] This includes the Family Skills evaluation in Y0 (Early Years Foundation).

[8] Spybrook & Raudenbush (2009) identified 5 five 3-level CRTs with school level randomisation and included a class level and 11 multisite trials with school level randomisation and a class level (Table 2, p302) but do not identify the specific studies.

Table 1 Completed 3-level trials that included school, class and pupil level

Evaluator	Intervention name. Date (Funder)	Phase, year group & Subject Area	Outcome(s)	Schools / classes per school / pupils per class	ICC_{Sch}	ICC_{class}	ICC_{pup}	School R^2	Class R^2	Pupil R^2	Total R^2	Hedges G effect sizes (95% CIs)
NATCEN	Family Skills[9] 2016/17 (EEF)	EYF Y0, Literacy	CEM literacy	102 / 1.7 / 16	0.02	0.15	0.83	0.29	n/a	0.54	n/a	+0.01 (-0.03; +0.05)
York & Durham	Grammar for writing [10] 2012/13 (EEF)	Primary KS2, Y6 Literacy	GL PiE Raw GL PiE writing	50 / 2.0 / 20.0 50 / 2.0 / 20.0	0.21 0.26	0.27 0.32	0.52 0.42	n/a n/a	n/a n/a	n/a n/a	0.29 0.29	+0.10 (-0.09; +0.30) +0.10 (-0.09; +0.30)
SHU	Dialogic Teaching[11] 2015/16 (EEF)	Primary KS2, Y5 English, Science & Maths	GL PT English GL PT Maths GL PT Science	76 / 2.0 / 15.8 76 / 2.0 / 16.3 76 / 2.0 / 16.1	0.07 0.06 0.09	0.02 0.04 0.01	0.91 0.90 0.90	0.25 0.67 0.82	-0.05 0.33 -0.02	0.43 0.46 0.42	0.41 0.47 0.45	+0.15 (0.00; +0.30) +0.09 (-0.04; +0.20) +0.12 (+0.01; +0.23)
SHU	ScratchMaths[12] 2015/16 & 16/17 (EEF)	Primary KS2, Y5 & Y6 Maths	KS2 Maths CT Thinking	110 / 1.9 / 54.4 81/2.0/48.8	0.11 0.13	0.01 0.02	0.88 0.87	0.18 0.36	-0.40 -0.17	0.51 0.24	0.47 0.25	0.00 (-0.12; +0.12) +0.15 (0.00; +0.29)
SHU	Multiplicative Reasoning[13], 2014 (DfE)	Secondary, KS3, Y7, Y8 & Y9 Maths	GL PTiM12 PTiM13 PTiM14	55 / 2.3 / 18.7 56 / 2.0 / 18.9 52 / 2.1 / 18.9	0.21 0.27 0.02	0.42 0.47 0.70	0.37 0.26 0.28	0.93 0.88 0.76	0.96 0.88 0.76	0.37 0.16 0.09	0.74 0.69 0.58	-0.02 (-0.12; +0.08) -0.08 (-0.24; +0.07) -0.11 (-0.29; +0.06)

[9] The Family Skills evaluation used the Centre for Evaluation & Monitoring (CEM) BASE Reception Baseline Assessment at both baseline and outcome. See Hussain et al., 2018, https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/family-skills/ and www.cem.org

[10] The Grammar for Writing evaluation used the GL Progress in English (PiE) test as the outcome and predicted KS2 writing level as a baseline covariate. See Torgerson et al., 2014, https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/grammar-for-writing/ and www.gl-assessment.co.uk/products/progress-test-in-english-pte/).

[11] The Dialogic Teaching evaluation used GL Progress Test in English, Maths and Science as outcomes and KS1 test score as the baseline covariate. See Jay et al., 2017; https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/dialogic-teaching/ and www.gl-assessment.co.uk/products/progress-test-in-maths-ptm/

[12] The ScratchMaths evaluation used KS2 maths as the primary outcome (Y6 in 2016/17) and developed a measure of Computational Thinking for an interim secondary outcome (Y5 in 2015/16) with KS1 maths used as a baseline covariate for both. See Boylan et al., 2018; https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/scratch-programming/ and www.gl-assessment.co.uk/products/progress-test-in-maths-ptm/

[13] The Multiplicative Reasoning evaluation used the GL Progress Test in Mathematics (PTiM) for three pupil cohorts (Y7, Y8 & Y9) in the 2013/14 academic year and KS2 maths as the baseline covariate. See Boylan et al., 2015; www.gov.uk/government/publications/multiplicative-reasoning-professional-development-programme and www.gl-assessment.co.uk/products/progress-test-in-maths-ptm/

NFER	Teacher Observation[14], 2014/15 & 15/16 (EEF)	Secondary Maths & English, KS4 Y11	Maths GCSE English GCSE	82 / - / 14 82 / - / 14	n/a n/a	0.43 0.29	n/a n/a	n/a n/a	n/a n/a	n/a n/a	0.46 0.34	+0.10 (-0.49; +0.68) -0.02 (-0.50; +0.47)
University of Nottingham	Catch-up Numeracy Effectiveness re-grant[15] 2016/17 (EEF)	Primary KS2 Y3, Y4 & Y5 Maths	GLPTiM (8, 9 &10)	All - 142 / - / 10 FSM - 132/-/4	0.25 0.28	0.09 0.06	0.60 0.62	n/a n/a	n/a n/a	0.48 0.45	n/a n/a	-0.04 (-0.21; +0.13) +0.11 (-0.05; +0.27)

[14] The Teacher Observation evaluation used GCSE English and maths (Y11) and developed a KS3 test (Y10) as outcomes and KS2 scores as baseline covariates. See Worth et al., 2017; https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/teacher-observation/

[15] The Catch-up numeracy regrant uses a four rather than three level CRT design (Region<School<TA<Pupil) and so this needs to be taken into account when reading the ICC statistics. See https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/catch-up-numeracy-2015/

Family Skills was a literacy intervention targeted at parents of children for whom English was an additional language (EAL) evaluated by NatCen (Hussain et al., 2018). Using a 3-level CRT design, half of the 102 recruited schools were randomised to the intervention group and half to a business as usual control groups. Within intervention schools, parents of EAL reception-aged (Y0) children attended up to 11 weekly sessions with external tutors employed by the delivery partner. The efficacy trial ran in the 2016/17 academic year and did not find a statistically significant difference in literacy between the intervention and control group . The initial estimated ICC_{Sch}(=0.11 sds) was higher than the value observed (ICC_{Sch}=0.02) and the estimated ICC_{class}(=0.05) was lower than the value observed (ICC_{class}=0.15). Initial estimates for covariate explanatory power at the school (R^2_{sch}=0.20) and pupil (R^2_{pup}=0.54) levels are given but zero explanatory power was assumed at the class level (R^2_{class}=0.00). The observed explanatory power is not reported.

Grammar for Writing was a teacher PD programme focusing on KS2 writing that was evaluated by York and Durham Universities (Torgerson et al., 2014). The impact of Grammar for Writing on English attainment was evaluated using a complex experimental design (partial split plot, 3-armed) that included randomisation at both class and pupil levels but not school level. However, the impact evaluation did include ICC estimates for the strength of clustering at both class and school levels. 55 schools were recruited to the trial and in each school two KS2 classes were randomly allocated to either the intervention group or business as usual control group. Within the 55 intervention classes, pupils were randomly selected to receive additional support (whole class plus small group work) or to receive the intervention just with the whole class. The efficacy trial ran in the 2012/13 academic year and did not find a statistically significant difference in writing between the intervention and control group. Initial estimated ICC_{Sch} and ICC_{class}(both=0.19) were slightly lower than the values observed (ICC_{Sch}=0.21; ICC_{class}=0.27). One estimate for predicted covariate explanatory power is provided (R_{All}=0.70, therefore R^2_{All}=0.49) but the observed value was slightly lower (R_{All}=0.54, therefore R^2_{All}=0.29). The Grammar for Writing intervention is currently being re-evaluated using an effectiveness trial[16] which uses a more

[16] See https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/grammar-for-writing-effectiveness-trial

conventional 2-level CRT design with school level randomisation involving 155 primary schools and is due to report soon.

Dialogic Teaching was a teacher PD programme evaluated by Sheffield Hallam University (Jay et al., 2017). The Dialogic Teaching intervention emphasised dialogue in the classroom with the theory that this would lead to pupil-level gains in language, mathematics and science attainment. Using a 3-level CRT design, half of the 80 recruited schools were randomly allocated to the intervention group and half to a business as usual control group. Within intervention schools, teachers received PD training and schools were provided with classroom resources and loaned video and audio equipment. The efficacy trial ran in the 2015/16 academic year amongst Y5 pupils. The evaluation found that, in Y5, pupils in intervention schools attained significantly higher in English and science compared with pupils in control schools. A positive impact was also observed for maths but this did not reach statistical significance. The effect sizes observed for English and science were very small (<0.10) and lower than the estimated MDES for the evaluation (0.25 sds) . Therefore whilst a positive and statistically significant effect was observed, the statistical power to detect an effect size of 0.10 sds was notably lower than 0.80[17]. Initial estimated ICC_{Sch} and ICC_{class}(both=0.10) were both slightly higher than the values observed (ICC_{Sch}=0.06-0.09; ICC_{class}=0.01-0.04). A single initial estimates for covariate explanatory power was provided (R^2_{All}= 0.49). The observed explanatory power can be calculated from the model tables in the report[18]. These are shown to differ for the three outcome variables (English, Maths & Science) and at the three levels. Interestingly, it is only in Maths where any class-level covariate explanatory power was observed (negative values were found for both English and Science). A recent EEF evaluation round included a re-grant for the Dialogic Teaching intervention using an effectiveness trial.

ScratchMaths was a teacher PD programme focusing on KS2 maths evaluated by Sheffield Hallam University (Boylan et al., 2018). Using a 3-level CRT design, half of the 110 recruited schools were randomised to the intervention group and half to a business as usual control

[17] To detect an MDES of 0.10 or higher, using the design parameters specified in Table 1 equation 3.0 can be used to show that this design had an estimated statistical power that ranged from 0.33 (English) and 0.60 (Science). Given that the observed effect sizes<0.10 sds, the actual statistical power for the observed effect sizes for Dialogic Teaching will be lower than these estimates.

[18] See Dialogic Teaching report, Appendix F pp59-60.

groups. Within intervention schools, teachers received PD training and schools were provided with ScratchMaths classroom resources. The efficacy trial ran in the 2015/16 and 2016/17 academic years. The primary outcome was KS2 maths attainment amongst Y6 pupils in 2016/17 and an interim secondary outcome was Computational Thinking measured at the end of 2015/16 when pupils were in Y5. The evaluation found no evidence that the ScratchMaths intervention resulted in pupil level gains in KS2 maths but did find a statistically significant positive impact for the Computational Thinking interim secondary outcome[19]. Initial estimated ICC_{Sch}(=0.13) and ICC_{class}(=0.07) were both slightly higher than the values observed (ICC_{Sch}=0.11; ICC_{class}=0.01-0.02). A single (school level) initial estimate for covariate explanatory power is provided (R^2_{Sch}= 0.59). Observed covariate explanatory power differed for the two outcome variables (KS2 Maths & Computational Thinking) and at the three levels. Interestingly, for both outcomes, a negative explanatory power was observed at the class level. This is perhaps counter-intuitive, but negative R-squares are a known phenomenon within multilevel analyses (Recchia, 2010; Lou & Azen, 2013). For example, when an explanatory variable acts in differing directions at different levels. Recchia (2010, p8) illustrates this using 'percieved support' as an explanatory variable and attainment as an outcome. At a class or school level, it may be that higher levels of (mean) percieved support is associated with higher attainment. However, at the pupil level it higher percieved support may be linked to lower attainment.

The Multiplicative Reasoning Project (MRP) was evaluated by Sheffield Hallam University (Boylan et al., 2015) using a 3-level CRT design. MRP was a teacher PD programme focusing on KS3 maths. Using a 3-level CRT design, half of the 58 recruited schools were randomised to the intervention group and half to a business as usual control group. Within intervention schools, teachers received PD training and schools were provided with MRP classroom resources. The efficacy trial ran in the 2013/14 academic year and involved pupils in Y7, Y8 and Y9. The evaluation found no evidence that the MRP intervention resulted in pupil level gains in KS3 maths. The initial ICC estimate for school-level clustering (ICC_{Sch}=0.10) was lower than the observed values for Y7 (ICC_{Sch}=0.21) and Y8 (ICC_{Sch}=0.27) but higher than was observed for Y9 (ICC_{Sch}=0.02). The initial estimate for class-level clustering ICC_{class}(=0.05)

[19] This statistically significant effect size was 0.13 sds, given the design parameters in Table 1, equation 3.0 can be used to estimate that the statistical power for detecting this effect size to be (1-β=0.44).

was much lower than the observed values for Y7 (ICC_{class}=0.42), Y8 (ICC_{class}=0.40) and Y9 (ICC_{class}=0.70). Initial estimates for (total) covariate explanatory power is provided for Y7 (R^2_{All}= 0.40), Y8 (R^2_{All}= 0.30) and Y9 (R^2_{All}= 0.20) which were all lower than the values observed at Y7 (R^2_{All}= 0.74), Y8 (R^2_{All}= 0.69) and Y9 (R^2_{All}= 0.58). Covariate explanatory power was observed to be much stronger at the school (R^2_{Sch}= between 0.76-0.93) and class (R^2_{class}= between 0.76-0.96) levels compared with the pupil level (R^2_{pup}= between 0.09-0.37). The differences between estimated and observed class-level clustering and covariate explanatory power led to a marked reduction in statistical sensitivity which is reflected on by the evaluators:

> " The structure of variation is somewhat different to what was anticipated at the design stage - with around 50% of the total variation in the (overall) PiM outcome variable located at the class-level and 20% located at the school level. This leaves 30% of the variation at the individual pupil level….
>
> …The notable class level ICC is an indication of how 'similar' pupils are (in terms of attainment) within each maths class - in other words, this is an ability grouping effect. Ability grouping brings together pupils with similar levels of mathematics into 'sets' or 'streams'. It is interesting to observe how this class-level ICC is much stronger for the Y9 sample (70% of the variation located at the class level) compared with Y8 (47%) and Y7 (42%). This may reflect a movement to universal or near universal setting by ability group between Y7 and Y9."
> Boylan et al., 2015 p30

Teacher Observation was a teacher PD intervention that aimed to improve teacher effectiveness through structured peer observation evaluated by NFER (Worth et al., 2017). A complex experimental design was used to evaluate the impact of Teacher Observation on English and maths attainment in Y10 and Y11. The design included three experiments all with multiple levels; two 2-level CRT design were used for a school-level (school and pupil levels) and department-level (department and pupil levels) experiments and a 3-level CRT design was used for a teacher-level experiment (school, teacher and pupil levels). The Teacher Observation evaluation ran in the 2014/15 and 2015/16 academic years and

involved three pupil cohorts[20]. The primary outcome for this evaluation was drawn from the 2-level CRT design used to evaluate the school level experiment but some details on the 3-level CRT used for the teacher-level experiment are reported. The evaluation found no evidence that Teacher Observation led to a positive impact across all experiments. Whilst Worth et al. (2017) acknowledge that the analysis of the teacher-level experiment included a school, teacher and pupil level, estimates for how strongly the outcome variables were clustered at three levels are not reported. An initial estimate for (possibly teacher level or teacher and school level combined) clustering is provided (ICC_{Teach}=0.075) but the observed values were found to be larger for both the maths (ICC_{Teach}=0.43) and English (ICC_{Teach}=0.29) outcomes. A single initial estimate for covariate explanatory power is provided (R_{All}= 0.75; R^2_{All}= 0.56) which was higher than the values observed for the maths (R^2_{All}= 0.46) and English (R^2_{All}= 0.34). The differences between estimated and observed teacher-level clustering and covariate explanatory power led to reduced statistical sensitivity which is reflected on by the evaluators:

> "The discrepancy between the ICC at protocol and analysis stages is due to the prevalence of setting." (Worth et al., 2017 p27)

Comparing the three primary and single secondary school 3-level CRTs, class level clustering seems to be stronger at secondary. Whilst there are insufficient trials to draw any firm conclusions about patterns relating to the subject area for evaluations, the ICC estimates from the Dialogic Teaching trial were highest in (KS2) maths (ICC_{class}=0.04) and lowest in science (ICC_{class}=0.01).

The MRP trial provides some interesting patterns that seem to have face validity in the context of the English education system. As noted by Boylan et al. (2017), the increasing strength of class-level clustering might reflect how the use of 'ability' group setting becomes increasingly common in maths between Y7 and Y9. The link between setting and strength of clustering is also made by Worth et al., (2017).

[20] Cohort 1 completed Y11 in 2014/15 (GCSE maths outcome). Cohort 2 were tested at the end of Y10 in 2014/15 (using a customised test based on KS3 past maths papers) and at the end of Y11 in 2015/16 (GCSE maths). Cohort 3 were tested at the end of Y10 in 2015/16 (customised test based on KS3 past maths papers) - see Worth et al., 2017 Table 6 p20.

Class level clustering relates directly to between-class within-school differences (or variance) in an outcome variable. Therefore, in the context of the English education system and widespread use of policies that group pupils into sets or streams (Francis et al., 2017), it seems reasonable to assume relatively strong class level clustering and this appears to be echoed in the (admittedly sparse) number of trials that have collected data at a class or teacher level.

3. How many EEF maths trials have been completed? [RQ2]

Table 2 presents a summary of 44 EEF maths evaluations that included a RCT or CRT impact evaluation. Table 2 is constructed using the EEF website to identify trials under the mathematics 'big picture theme' and meta-analysis data provided by Lortie-Forgues (2017). The majority of maths trials were in Primary schools (n=28, 64%), particularly the Y3-Y6 KS2 year groups (n=20, 45%). 14 (32%) maths trials were in secondary schools, particularly the Y10 & Y11 KS4 year groups (n=10, 23%).

26 (59%) of the maths evaluations have completed and the reports are published on the EEF website, 13 (30%) are mid-trial or about to report and five (11%) are at the early set-up stages.

Details on the experimental design are available for 40 of the 44 evaluations[21]. The majority of these have a CRT design (n=34, 85%) but very few have a 3-level CRT design (n=6, 15%).

Of the six 3-level CRTs, one has reported, four are mid-trial and one is in the early set up stages. Dialogic Teaching is the first 3-level EEF maths trial to report[22]. At the time of this paper, ScratchMaths was classed as mid-trial but reported during the review period of the paper[23]. There are three further mid-trial 3-level CRTs; a re-grant evaluation for Catch-Up Numeracy[24] , an evaluation of 'Maths in Context[25]' and an evaluation of 'Diagnostic Questions[26]'. Finally, I am aware of a 3-level CRT being used to evaluate a secondary (KS3) maths intervention but at the time of writing, the protocol for this evaluation has not been published[27].

[21] At the time of writing, details were available for all trials except for four of the five trials at the early (set-up) stages.
[22] The Teacher Observation trial discussed above has been classed as a 2-level CRT in Table 2 to reflect the design of the first experiment.
[23] see https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/scratch-maths/
[24] See https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/catch-up-numeracy-2015/
[25] See https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/maths-in-context/
[26] See https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/diagnostic-questions/
[27] See https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/realistic-maths-education/

Table 2: Summary of EEF Maths Evaluations that included an RCT or CRT impact evaluation to date[28]

	Total N=	Key Stage						Impact Evaluation Design			
		EY	KS1	KS2	KS3	KS4	Post	RCT with within school, individual (pupil) level	2-level CRT designs		3-level CRT design. School, class/teacher & pupil levels with school level randomisation
									school level randomisation	Class level randomisation	
Early/start up trials (no protocol)	5	0	1	1	1	2	0	-	-	-	1
Mid trial (protocol / SAP but no final report)	13	0	1	5	1	4	2	0	8	1	4
Number with final report	26	1	6	14	2	4	0	6	17	2	1
..Reported a statistically significant impact	7	0	3	2	1	1	0	1	5	1	0
…EEF identified Promising Projects	8	0	2	5	0	1	0	1	3	2	1
…Trials with 4 or 5 EEF padlock rating	14	0	4	8	1	1	0	3	10	1	0
Total number with protocol, SAP or final report	44	1	8	20	4	10	2	6	25	3	5

[28] Date of writing = October 2018

Catch-up numeracy is a PD programme that trains Teaching Assistants (TAs) to deliver one-to-one sessions with KS2 pupils identified as struggling with numeracy. The evaluation is being undertaken by Nottingham University using a 3-level CRT design with school, teaching assistant and pupil levels and randomisation at the school level (Atkins, 2017). Half of the 150 recruited schools were randomly allocated to the intervention group and half to the control group. The trial took place in the 2016/17 academic year and is due to report in 2018. The Protocol and Statistical Analysis Plan (SAP) for the Catch-up numeracy re-grant evaluation provides detail on estimates for assumed clustering and explanatory power at the design stage. Clustering at the school-level (ICC_{Sch}=0.10) is estimated to be twice as strong as at the TA-level (ICC_{TA}=0.05). The school-level explanatory power[29] is estimated at (R^2_{Sch}= 0.64). Allowing the number of pupils-per-TA to vary between 2 and 6, this leads to MDES estimates between 0.17 and 0.25 sds.

Maths in Context is a teacher PD programme focusing on improving financial capability[30] amongst KS4 pupils. The evaluation is being undertaken by Nottingham University using a 3-level CRT design with school, class/teacher and pupil levels and randomisation at the school level (Atkins, 2018). Half of the 130 recruited schools were randomly allocated to the intervention group and half to the control group. The efficacy trial is taking place in the 2017/18 and 2018/19 academic years and is due to report in 2020. The Protocol and Statistical Analysis Plan (SAP) for the Maths in Context evaluation provides detail on assumptions made for the effects of clustering and explanatory power at the design stage. Clustering at the school-level (ICC_{Sch}=0.165) is assumed to be over three times as strong as at the class/teacher level (ICC_{class}=0.05). The school-level explanatory power[31] is assumed at (R^2_{Sch}= 0.49). With 4 classes of 25 pupils per school, this leads to an MDES estimate of 0.17 sds.

[29] This relates to a test, re-test correlation for GL Progress Test in Maths

[30] This is further specified as financial knowledge and understanding applied numeracy and problem solving skills.

[31] This relates to the correlation between KS2 maths and GCSE maths at the school level.

Diagnostic Questions (DQ)[32] is maths software based intervention that focuses on providing formative feedback to KS4 pupils and teachers. The evaluation is being undertaken by Alpha Plus and Manchester Met University (Seymour & Morris, 2018) using a 3-level CRT design with school, class and pupil levels and randomisation at the school level. Half of the 180 recruited schools were randomly allocated to the intervention group and half to the control group. The trial is taking place in the 2018/19 and 2019/20 academic years and is due to report in 2021. From the DQ Protocol, clustering at the school-level (ICC_{Sch}=0.20) is estimated to be four times as strong as clustering at the class level (ICC_{class}=0.05). Explanatory power[33] is only assumed at the school (R^2_{Sch}= 0.25) and pupil (R^2_{pup}= 0.50) levels with zero class-level explanatory power (R^2_{class}= 0.00). With 6 classes of 25 pupils per school, this leads to an MDES estimate of 0.17 sds.

From Table 1, four 3-level CRTs with school level randomisation and include a class (or teacher) level were identified and with the additional three mid-trial 3-level CRTs, a total of seven have been identified. Six of these seven trials are evaluating teacher (or TA) PD programmes that aim to cause pupil-level gains in attainment. The Diagnostic Questions intervention is distinct from all 3-level CRT trials because the intervention (software) is targeted directly at pupils (although training of teachers in the best use of the software is required). DQ also targets teachers and parents but the direct link to pupils makes the intervention distinctive amongst the 3-level CRT trials. Amongst the 25 2-level maths CRTs with school level randomisation, 88% (n=22) evaluated Teacher/TA PD programmes.

[32] From the protocol, the Diagnostic Questions (DQ) intervention has been renamed Eedi but is still named as Diagnostic Questions on the EEF website.
[33] This relates to the correlation between KS2 maths and GCSE maths at the school level.

4. What are the methodological implications of ignoring class level in educational CRT trials? [RQ3]

This section begins by briefly considering literature that has discussed effects of ignoring a (within-school) level of clustering within multilevel designs. Following this, the section focuses on equations for calculating the MDES for RCT, 2-level CRT and 3-level CRT designs to help explore how clustering, covariate explanatory power and sample size impact on statistical sensitivity.

Over two decades ago, Hill & Rowe (1996) reflected on findings from the first 15 years of multilevel research in education to highlight how variation between classes appeared to be far more significant than variation between schools. Opdenakker & Van Damme (2000) used data from the " Longitudinaal Onderzoek Secundair Onderwijs (LOSO)" for an observational sample of 2,680 pupils, nested in 150 classes, 81 mathematics teachers and 46 secondary schools in Flanders, Belgium to illustrate the importance of reflecting the complete (4-level; School<Teacher<Class<Pupil) hierarchy in the data for accuracy in estimating fixed coefficients, variance components and standard errors. They warn that ignoring an important level can lead to different research conclusions. Van den Noortgate et al. (2005) provide a reanalysis of the LOSO data used by Opdenakker & Van Damme but with a restricted 'balanced' sample (of 2 teachers per school, 2 classes per teacher, and 10 pupils per class) and highlight how ignoring a level has substantial impact on the conclusions of multilevel analysis. Van den Noortgate et al. (2005) conclude that correctly modelling the hierarchical structure is vital if variance decomposition is of interest or standard errors are used to make inferential conclusions but is of less importance if interest is only in estimates of fixed effects. Education CRTs for efficacy trials do focus on estimating coefficients for fixed effects models but also draw on standard errors to assess the level of statistical significance and power. Trammer & Steel (2001) discuss the theoretical importance of recognising various levels within spatial analyses using 1991 UK census data. Trammer & Steel (ibid) compare variance decomposition statistics for 3-level (Ward, enumeration district & individual) and two level (Ward & Individual; enumeration district & individual) models using a range of outcomes to empirically illustrate how variation across three levels is redistributed to two levels. Hutchison & Healy (2001) use educational

(secondary maths) data to illustrate the redistribution of variance from a 3-level (school, class & pupil) to a 2-level (school & pupil) model and note that "the estimated error variance of higher level means tend to be too small when the presence of a hierarchy at lower levels is ignored" (Hutchison & Healy; 2001 p5). Moerbeek (2004) compare 3-level and 2-level models to illustrate redistribution of the variance of the ignored level and how test statistics and statistical power are influenced if a level of nesting is ignored. Moerbeek (ibid) found that ignoring a level of clustering in an analysis had an effect on the variance components and that standard errors of regression coefficients may be overestimated which leads to lower statistical power to detect a specified effect. This literature all related to observational studies, perhaps reflecting how rare educational RCT designs were until relatively recently. In a critical note about Moerbeek's 2004 paper, Van Landeghem et al. (2005) highlight that the findings relating to observational studies will also be relevant to experimental studies with random assignment at higher levels in the data but primary outcomes measured at lower levels. In their final reflection Van Landeghem et al. (Ibid) emphasise practical over theoretical issues

> "With the knowledge available at this moment, it seems more worthwhile to put effort into the removal of obstacles preventing the inclusion of a relevant level of clustering than to speculate about the consequences of excluding the level from the analysis."Van Landeghem et al (2005, p433)

In terms of experimental designs, a key paper from Konstantopoulos (2008) provides methods for computing power in 3-level CRT designs and clearly states the importance of accurately reflecting data structure in the design and analyses of experiments. :

> "The appropriate power computations of three-level data structures need to include nesting effects at all levels. Similarly, the appropriate analyses of three-level data need to take into account this multilevel structure, because otherwise the standard errors of estimates and statistical tests of such analyses are incorrect. Specifically, the standard errors of treatment effect estimates (incorrectly) ignoring nesting are typically smaller, which translates to higher values of t tests and higher probabilities of finding a significant effect." Konstantopoulos, 2008, p85.

Konstantopoulos concluded that; sample size at higher cluster levels have a greater impact on power[34] and that ignoring a level of nesting results in an overestimation of statistical power "unless the intraclass correlation of the omitted level is exactly zero" (Konstantopoulos, 2008, p22).

In designing RCTs and CRTs it is important to ensure that the design has adequate statistical sensitivity to detect the 'effect' of an intervention. Statistical sensitivity relates to a number of things but here the focus is predominantly on statistical power (the probability of detecting an effect if one genuinely exists; which is commonly set to be 0.80 or 80% or higher) and minimum detectable effect sizes (MDES, the smallest true effect size that can be detected as statistically significant with a specified statistical power; Bloom, 1995).
This section now explores how clustering, covariate explanatory power and sample size impact on statistical sensitivity through an examination of equations used to estimate the MDES for different trial designs.

Minimum Detectable Effect Size (MDES) Equations

For comparability it is useful to standardise effect sizes into units of standard deviations (Bloom, 1995). Below are standardised MDES equations for a two armed trial for an RCT and two CRT designs. A number of methodological papers discuss 3-level CRT designs (Heo & Leon, 2008; Hedges & Hedberg, 2013; Spybrook et al., 2016) but two are drawn on directly for the MDES equations 1.0 to 3.0. The MDES equation for the RCT (1.0) and 2-level CRT (2.0) designs are drawn from Bloom (2006) and for the 3-level CRT (3.0) design the equation is from Dong & Maynard (2013) and Kelsey et al, (2017).

1.0 ***RCT design***: $$MDES_{RCT} = M_{n\prime-L\prime-2}\sqrt{\frac{1}{P(1-P)}}\sqrt{\left(\frac{(1-R^2_{pup})}{\mathrm{n}'}\right)}$$

2.0 **2 *level CRT design***: $$MDES_{2LCRT} = M_{K-L-2}\sqrt{\frac{1}{P(1-P)}}\sqrt{\left(\frac{\mathrm{ICC}_{sch}(1-R^2_{sch})}{\mathrm{K}}\right)+\left(\frac{(1-\mathrm{ICC}_{sch})(1-R^2_{pup})}{\mathrm{Km}}\right)}$$

3.0 **3 *level CRT design***: $$MDES_{3LCRT}\sim M_{K-L-2}\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{ICC_{sch}(1-R^2_{sch})}{K}+\frac{ICC_{class}(1-R^2_{class})}{JK}+\frac{(1-ICC_{sch}-ICC_{class})(1-R^2_{pup})}{nJK}}$$

[34] e.g. the number of classes has a greater impact on power compared with the number of pupils; the number of schools has a greater impact than the number of classes.

The RCT equation contains five variables and within these are three design parameters (Hedges, 2013); total number of individuals/pupils (n'); the proportion of pupils allocated to an intervention (P) and covariate explanatory power (R^2_{pup}).

MDES equations are shown for a 2-level and 3-level CRT design. The 2-level CRT design contains eight variables which include six design parameters; number of clusters (schools) (K); number of individuals/pupils per school (m[35]); the proportion of pupils allocated to an intervention (P); the school-level Intra-Cluster Correlation coefficient (ICC_{sch}) and covariate explanatory power at individual (R^2_{pup}) and school (R^2_{Sch}) levels. The 3-level CRT design contains 11 variables which include 9 design parameters; number of clusters (schools) (K); number of classes per school (J), number of individuals/pupils per class (n); the proportion of pupils allocated to an intervention (P); the school-level and class level ICCs (ICC_{sch} & ICC_{class}) and covariate explanatory power at individual (R^2_{pup}), class (R^2_{class}) and school (R^2_{Sch}) levels.

In RCT designs, the number of participants (n', sample size) and individual-level covariate explanatory power influence the MDES estimates. This is also the case for CRTs but with the additional complexity that sample sizes and covariate explanatory power are at multiple levels (individual & cluster levels). Additionally, with CRTs, the clustering of outcome data at school and class level is also influential.

Drawing on Hedges & Rhoads (2010), it can be useful to reorganise the terms for the 2-level and 3-level MDES equations as shown below in equations 2.2 and 3.2.

2.2
$$MDES_{2LCRT} \sim M_{K-L-2}\sqrt{\frac{1}{P(1-P)Km}}\sqrt{1+(m-1)ICC_{sch}-\left[R^2_{pup}+\left(mR^2_{sch}-R^2_{pup}\right)ICC_{Sch}\right]}$$

3.2
$$MDES_{3LCRT} \sim M_{K-L-2}\sqrt{\frac{1}{P(1-P)KJn}}\sqrt{1+(Jn-1)ICC_{sch}+(n-1)ICC_{class}-\left[R^2_{pup}+\left(JnR^2_{sch}-R^2_{pup}\right)ICC_{Sch}+\left(nR^2_{Class}-R^2_{pup}\right)ICC_{Class}\right]}$$

The t-distribution multiplier is determined by the level of statistical significance (α), statistical power (1-β) and the degrees of freedom calculated from the number of schools

[35] This assumes that m does not vary greatly across schools in a study. If there is great variation in the size of clusters across schools it can impact on statistical power. In these cases, it is suggested that the harmonic mean rather than the arithmetic mean is a more precise estimate for 'm' (Lauer et al., 2015).

(K) and number of school-level covariates (L). Given that this is determined only by school level factors, this part of the equations would be identical for a 2-level or 3-level CRT design. Assuming a balanced design and that the number of pupils per school (m) is the same as the number of pupils per class (n) multiplied by the number of classes per school (J) [i.e. m=Jn], the $\sqrt{\frac{4}{Km}}$ term in the 2-level design would also be identical to the ($\sqrt{\frac{4}{KJn}}$).term in a 3-level design. Therefore, it is in the last component where the distinction between 2-level and 3-level CRT designs lies. Looking more closely at the terms in this third component for 2-level and 3-level designs, they both can be seen to contain two sub-components. The first is directly determined from school and/or class level clustering; with increasing strengths of clustering leading to higher MDES estimates (and hence lower statistical sensitivity). The second sub-component accounts for the use of covariate explanatory power (in relation to school and/or class level clustering). With increasing explanatory power, the impact of school/class level clustering on trial sensitivity can be seen to decrease. This moderating effect of covariate explanatory power on how clustering impacts on statistical sensitivity is seen most clearly when explanatory power is assumed to be equal at all levels (i.e. $R^2_{ALL} = R^2_{sch} = R^2_{class} = R^2_{pup}$).

	2-level	**3-level**
3. The impact of clustering on MDES	$1 + (m-1)ICC_{sch}$	$1 + (Jn-1)ICC_{sch} + (n-1)ICC_{class}$
4…mitigation of this impact through covariate explanatory power	$-\left[R^2_{pup} + \left(mR^2_{sch} - R^2_{pup}\right) ICC_{Sch}\right]$	$-\left[R^2_{pup} + \left(JnR^2_{sch} - R^2_{pup}\right) ICC_{Sch} + \left(nR^2_{Class} - R^2_{pup}\right) ICC_{Class}\right]$
$IF\ R^2_{ALL} = R^2_{sch} = R^2_{class} = R^2_{pup}$	$-R^2_{ALL}[1 + (m-1)\ ICC_{Sch}]$	$-R^2_{ALL}[1 + (Jn-1)\ ICC_{Sch} + (n-1)\ ICC_{Class}]$

When $R^2_{ALL}$is zero, the fourth component would cancel completely. With no covariate explanatory power, the impact of clustering (school and class) is at a maximum.

By comparing component 3 for the 2-level and 3-level CRT designs, it can be seen that school level clustering has a weighting of $(m-1)$ or $(Jn-1)$ compared with class level clustering $(n-1)$. If J>1, m>n and with zero explanatory power, clustering at the school level is given a greater weighting (in determining the MDES) compared with clustering at the class level.

This usually means that trial sensitivity is more dependent on clustering at the school compared with class levels and thus the size of the trial in terms of number of schools randomised.

When $R^2_{ALL} = 0.50$, the fourth sub-component would equate to half of the third sub-component. Therefore, the overall potential impact of class/school level clustering on trial sensitivity is reduced by half. For a given level of significance, power and sample size, as explanatory power increases, MDES estimates decrease (i.e. statistical sensitivity increases). As R^2_{ALL} approaches 1.00, the effect of clustering at school/class level (along with the MDES estimate) will approach zero.

Equations 1.0 to 3.0 can be rearranged so that statistical power or sample size rather than MDES is the subject. For example, for calculating the statistical power to detect a specified MDES given other fixed design parameters (sample size, clustering of outcome data, covariate explanatory power, balance of design) or to calculate the sample size required to detect a specified MDES statistic with a given statistical power and other design parameters (clustering of outcome data, covariate explanatory power, balance of design).

From literature and examining equations, it does seem clear that, theoretically, ignoring a class level 'matters' if class level clustering of a trial outcome variable is present. The impact on sensitivity brought by class level clustering is usually not as great as the impact brought by school level clustering but ignoring class level clustering does have implications. Specifically, assuming a 2-level design can result in underestimating MDES statistics (and hence falsely overstating the statistical sensitivity of a trial). The equations also illustrate that the impact of class level clustering on trial sensitivity can be mitigated through the use of covariate explanatory power and/or by increasing the number of schools and classes per school in the trial.

Answers provided by examining equations can be a little abstract to help fully illuminate whether class level clustering 'matters' in the design of educational trials. This paper now proceeds to use the equations to provide some numerical and visual illustrations to help

explore the inter-relationships between MDES estimates, sample size, statistical power, clustering at school and class levels and covariate explanatory power.

Visualising the impact of class-level clustering and covariate explanatory power on statistical sensitivity, power and required sample size.

Two perspectives are taken; first with a fixed number of (100) schools and second, with a fixed Minimum Detectable Effect Sise (MDES) of 0.20 sds.

Perspective 1 - a fixed number of 100 schools:

The first perspective has a fixed number of (K=100) schools but allows some variation at the class (2 to 6 per school) and pupil (10-25 per class; 20-150 per school) levels. The relationship between MDES and class level clustering is examined numerically and visually for different strengths of school level clustering and covariate explanatory power. For simplicity, this is first done by assuming the same explanatory power of covariates at all three levels (i.e. $R^2_{ALL} = R^2_{sch} = R^2_{class} = R^2_{pup}$). Following this, the covariate explanatory power is fixed at school and pupil levels s (i.e.$R^2_{sch} = R^2_{pup}$). but allowed to vary at the class level (R^2_{class}).

Perspective 2 - a fixed MDES of 0.20 sds[36]

The second perspective focuses on detecting a specified effect size (0.20 sds or higher) as statistically significant ($p<0.05$) with a power of 80% or higher. 2-level CRT designs which ignore class-level clustering are compared with 3-level CRT designs that include a class level. The impact of class-level clustering is considered in terms of statistical power, the actual MDES that could be with a statistical power of 80% or higher and the increase in sample size needed in order to detect an MDES of 0.20.

[36] To increase precision, the MDES used here was 0.2049 sds which rounds to 0.20 sds

Perspective 1: 100 schools;

Relationship between class level clustering and trial sensitivity

Perspective one focuses on 2-armed 3-level CRTs with school level randomisation and balanced (p=0.5) designs. The total number of schools is fixed at (K=) 100 but the number of classes per school and pupils per class has been allowed to vary according to three within-school sample scenarios summarised below:

Sample Scenario	*Pupils per Class*	*Classes per School*	*Pupils per School*	
1	***10***	***2***	***20***	e.g. Evaluations of targetted interventions (e.g. poor readers) or to analyses of pupil subgroups (e.g. gender, FSM) within whole-class interventions
2	***25***	***2***	***50***	e.g. Evaluations of whole class interventions in two classes of a year group - which may be an entire year group for a 2-form entry Primary school or a sample of classes in a larger Primary or Secondary school
3	***25***	***6***	***150***	e.g. This might relate to evaluations of whole class interventions across an entire year group (6 classes) in Secondary school

Table 3 presents a range of MDES estimates for the three sample scenarios. In calculating the MDES estimates, a two-tailed statistical significance is set at 0.05 ($\alpha/2$=0.025) and statistical power at 0.80 (β=0.20). A single primary outcome is assumed. The number of school-level covariates (L) is fixed at two[37] which results in a t-distribution multiplier with 96 degrees of freedom.

The MDES estimates in Table 3 were calculated allowing ICC values at the school level to vary between 0.00 and 0.20 and at the class level between 0.00 and 0.50. Explanatory power of covariates at school, class and pupil levels is allowed to vary between 0.00 and

[37] The school level group (intervention or control) identifier and a covariate that correlates with a trial outcome.

0.81. For simplicity, the same explanatory power of covariates is assumed for all three levels in Table 3.

Table 3: MDES estimates for sample scenarios 1 to 3 for different strengths of school & class level clustering and covariate explanatory power

- *zero covariate explanatory power [R_{school}^2 / R_{class}^2 / R_{pupil}^2 = 0]*

School level ICC	**Scenario 1** *2 classes of 10 pupils per school*	**Scenario 2** *2 classes of 25 pupils per school*	**Scenario 3** *6 classes of 25 pupils per school*
	Class level ICCs ***zero (0.00)****;low (0.05); high (0.20); v high (0.50)*		
0.00	**0.13**; 0.15; 0.21; 0.30	**0.08**; 0.12; 0.19; 0.29	**0.05**; 0.07; 0.11; 0.17
0.05	**0.18**; 0.20; 0.25; 0.32	**0.15**; 0.17; 0.23; 0.32	**0.13**; 0.14; 0.17; 0.21
0.10	**0.22**; 0.23; 0.27; 0.34	**0.19**; 0.21; 0.26; 0.34	**0.18**; 0.19; 0.21; 0.24
0.20	**0.28**; 0.29; 0.33; 0.39	**0.26**; 0.28; 0.32; 0.38	**0.26**; 0.26; 0.28; 0.30

- *moderate covariate explanatory power [R_{school}^2 / R_{class}^2 / R_{pupil}^2 = 0.49]*

School level ICC	**Scenario 1** *2 classes of 10 pupils per school*	**Scenario 2** *2 classes of 25 pupils per school*	**Scenario 3** *6 classes of 25 pupils per school*
	Class level ICCs ***zero (0.00)****;low (0.05); high (0.20); v high (0.50)*		
0.00	**0.09**; 0.11; 0.15; 0.21	**0.06**; 0.08; 0.14; 0.21	**0.03**; 0.05; 0.08; 0.12
0.05	**0.13**; 0.14; 0.18; 0.23	**0.11**; 0.12; 0.16; 0.22	**0.10**; 0.10; 0.12; 0.15
0.10	**0.15**; 0.17; 0.20; 0.25	**0.14**; 0.15; 0.19; 0.24	**0.13**; 0.14; 0.15; 0.17
0.20	**0.20**; 0.21; 0.23; 0.28	**0.19**; 0.20; 0.23; 0.27	**0.18**; 0.19; 0.20; 0.22

- *very high covariate explanatory power [R_{school}^2 / R_{class}^2 / R_{pupil}^2 = 0.81]*

School level ICC	**Scenario 1** *2 classes of 10 pupils per school*	**Scenario 2** *2 classes of 25 pupils per school*	**Scenario 3** *6 classes of 25 pupils per school*
	Class level ICCs ***zero (0.00)****;low (0.05); high (0.20); v high (0.50)*		
0.00	**0.06**; 0.07; 0.09; 0.13	**0.03**; 0.05; 0.08; 0.13	**0.02**; 0.05; 0.05; 0.08
0.05	**0.08**; 0.09; 0.11; 0.14	**0.06**; 0.08; 0.10; 0.14	**0.06**; 0.06; 0.07; 0.09
0.10	**0.09**; 0.10; 0.12; 0.15	**0.08**; 0.09; 0.11; 0.15	**0.08**; 0.08; 0.09; 0.11
0.20	**0.12**; 0.13; 0.14; 0.17	**0.11**; 0.12; 0.14; 0.17	**0.11**; 0.11; 0.12; 0.13

With 100 schools and 2 classes of 25 pupils per school (scenario 2), assuming zero covariate explanatory power and ICC_{Sch} = 0.10, when ICC_{class} = 0.00, the estimated MDES is 0.19 sds (highlighted in Table 3). This MDES estimate is seen to increase with increasing class level clustering; from 0.20 sds when ICC_{class} = 0.01 to 0.26 sds when ICC_{class} = 0.20 and to 0.34 when ICC_{class} = 0.50. In absolute terms, an MDES of 0.19 is an underestimate of between 0.004 (when ICC_{class} =0.01) and 0.144 sds (ICC_{class} =0.50). In relative terms, an MDES of 0.19 sd is an underestimate of between 2.0% and 42.6%.

With moderate covariate explanatory power (R_{ALL}^2 =0.49), the estimated MDES is 0.14 sds when ICC_{class} =0.00. An MDES of 0.14 sds remains until ICC_{class} =0.03 (MDES=0.15 sds) which increases to 0.24 sds when ICC_{class} =0.50. In absolute terms, this equates to an underestimate of between 0.003 to 0.103 sds; in relative terms, an underestimate of between 2.0% and 42.6%.

When covariate explanatory power is very high (R_{ALL}^2 =0.81), when ICC_{class} =0.00, the estimated MDES is 0.08 sds but as class level clustering increases from 0.01 to 0.20, MDES estimates increase from 0.09 (when ICC_{class}. = 0.01) to 0.15 sds (when ICC_{class}. = 0.50). In absolute terms, a difference/underestimate of between 0.002 to 0.063 sds; in relative terms, an underestimate of between 2.0% and 42.6%.

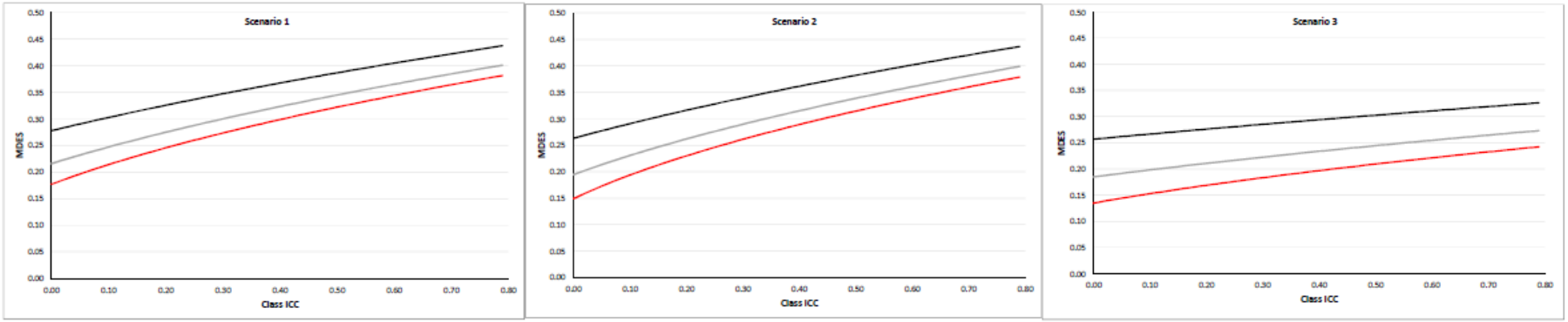
Figure 1a provides a visual representation of Table 3. MDES estimates (Y axis) are plotted against Class ICC values (X axis). Three line graphs are shown for each of the three sample scenarios (for three levels of covariate exploratory power; R_{ALL}^2 = 0.00; 0.49 & 0.81). On each line graph, three lines are plotted that illustrate the relationship between class level clustering and MDES estimates for three strengths of school level clustering (School ICC=0.05, 0.10 and 0.20).

Figure 1a highlights the following:

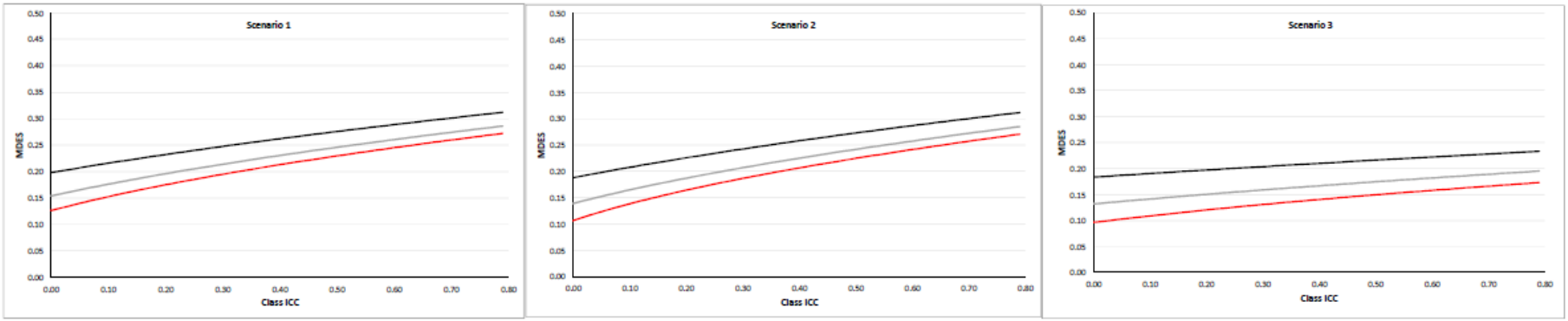
- Ignoring class level clustering when it exists results in underestimating MDES values (and hence overstating the statistical sensitivity of a trial).
- Use of covariates with high explanatory power mitigates the potential impact of class level clustering on MDES estimates.
- The similarity of the graphs for scenarios 1 and 2 and difference with scenario 3 highlight the greater influence of the number of classes per school compared with pupils per class on statistical sensitivity / MDES

Figure 1a: Class ICC (X) v MDES Estimate (Y) by School-level ICC (0.05; 0.10; 0.20)

School, Class & Pupil R-square=0.00; 0.49; 0.81
Number of schools=100 / alpha=0.05 / 1-Beta=0.8
...school, class & individual R-square=0



...school, class & individual R-square=0.49



...school, class & individual R-square=0.81

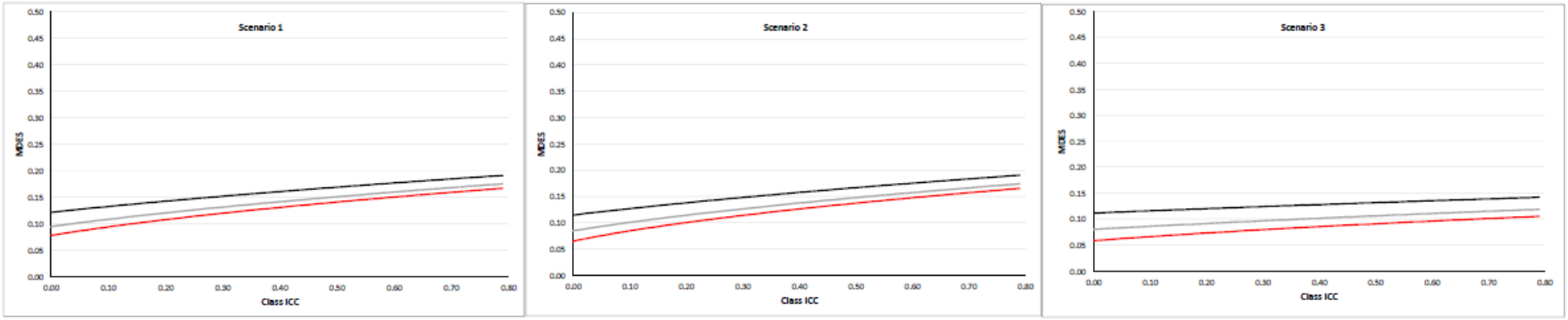


Table 3 and Figure 1a provide an examination of the relationship between class-level clustering and statistical sensitivity (MDES). Table 3 and Figure 1a also show how covariate explanatory power mitigates the negative influence of class-level clustering on MDES estimates. However, explanatory power is assumed to be equal at all levels (i.e. $R^2_{ALL} = R^2_{sch} = R^2_{class} = R^2_{pup}$). Figure 1b provides a second visualisation for perspective one and fixes explanatory power at school and pupil levels (i.e. $R^2_{sch} = R^2_{pup}$) but allows R^2_{class} to vary. MDES estimates (Y axis) are again plotted against ICC_{class} values (X axis).

Table 4 provides a reference point for Figure 1b. Table 4 presents the MDES estimates for a 3-level CRT that assumes a class level ICC of 0.00. A total of nine MDES estimates are shown; for each of the three sample scenarios an MDES estimate for three levels of covariate explanatory power at school and pupil levels is shown. For example, with school & pupil level explanatory power set at R^2=0.49, with 100 schools an MDES estimate of 0.20 sds is found for scenario 1, 0.19 sds for scenario 2 and 0.18 for scenario 3.

Table 4: MDES estimates assuming zero class-level clustering
K=100 schools, P=0.05, School ICC=0.20; α=0.05; 1-β=0.80

R^2 (School & Pupil levels)	**Scenario 1** 20 pupils per school	**Scenario 2** 20 pupils per school	**Scenario 3** 150 pupils per school
R^2=0.00	0.28	0.26	0.26
R^2=0.49	0.20	0.19	0.18
R^2=0.81	0.12	0.11	0.11

Figure 1b takes each of the nine MDES estimates in Table 4 and illustrates how they increase when the strength of class level clustering is allowed to increase from (ICC_{Class}=) 0.00 to 0.80. In each of the nine charts in Figure 1b, three lines are used show the relationship between class level clustering and MDES estimates for three levels of class-level explanatory power; R^2_{Class}=0.00; 0.49 and 0.81.

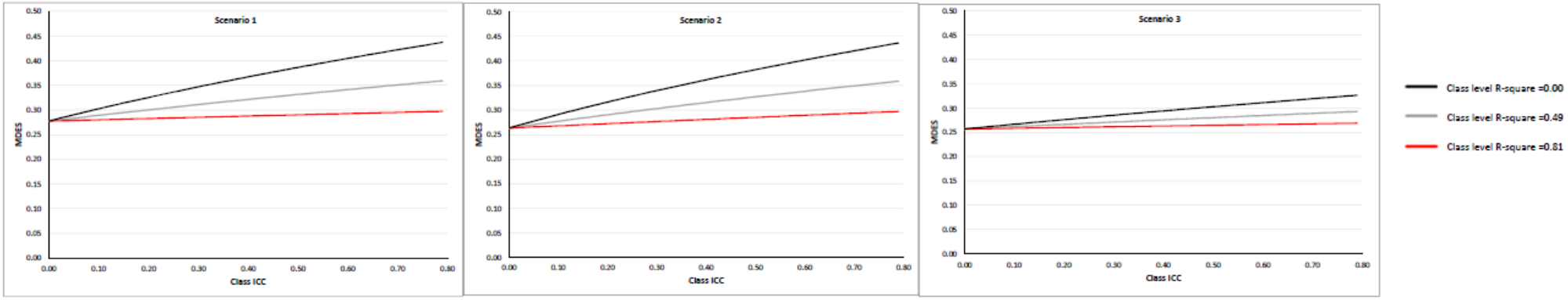
For example**, when school and pupil level explanatory power =0.00,** the MDES estimates for scenario 2 (2 classes of 25 pupils per school) begin at 0.26 sds (when ICC_{Class}=0.00, see

Table 4) but are seen to increase with ICC_{Class} in Figure 1b. However, the size of increased MDES is seen to reduce with increasing strength of class-level explanatory power. With zero class-level explanatory power, the MDES estimate is seen to increase to 0.28 sds when ICC_{Class}=0.05; to 0.29 sds when ICC_{Class}=0.10; 0.32 sds when ICC_{Class}=0.20; 0.38 sds when ICC_{Class}=0.50. With high class-level explanatory power (0.81), the MDES estimate remains at 0.20 sds until ICC_{Class}=0.05; remains at 0.27 sds until ICC_{Class}=0.26 increasing to 0.29 sds when ICC_{Class}=0.50.

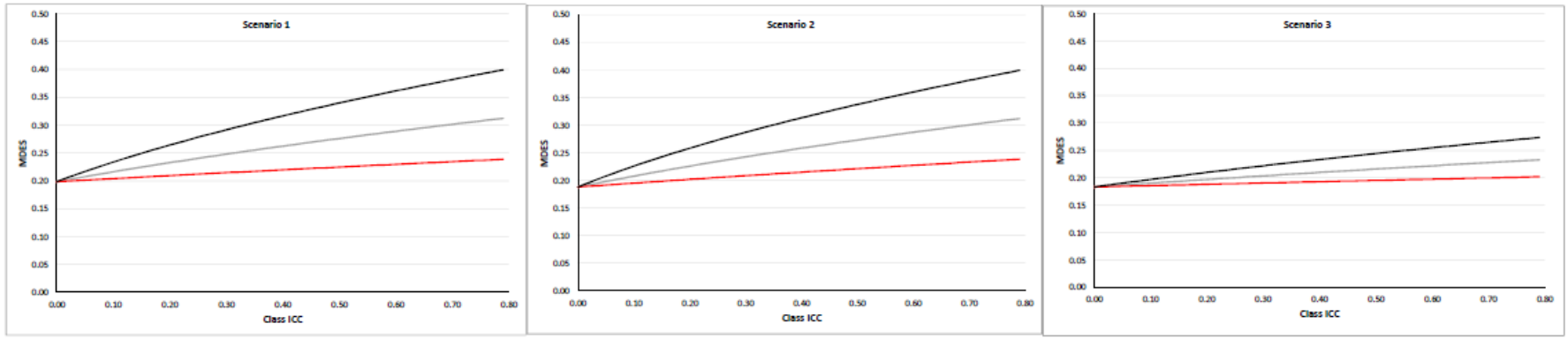
When school and pupil level explanatory power =0.81, the MDES estimates for scenario 2 (2 classes of 25 pupils per school) begin at 0.11 sds (when ICC_{Class}=0.00, see Table 4) but are seen to increase with ICC_{Class} in Figure 1b. With zero class-level explanatory power, the MDES estimate is seen to increase to 0.15 sds when ICC_{Class}=0.05; to 0.17 sds when ICC_{Class}=0.10; 0.21 sds when ICC_{Class}=0.20; 0.20 sds when ICC_{Class}=0.50. With high class-level explanatory power (0.81), the MDES estimate increases to 0.12 sds when ICC_{Class}=0.01 to 0.13 sds when ICC_{Class}=0.10; 0.14 sds when ICC_{Class}=0.20; 0.17 sds when ICC_{Class}=0.50.

Figure 1b: Class ICC (X) v MDES Estimate (Y) for three levels of class-level explanatory power (0.00; 0.49; 0.81)

School ICC=0.20 / School & Pupil R-square=0.00; 0.49; 0.81
number of schools=100 / school ICC=0.2 / alpha=0.05 / 1-Beta=0.8
...school & individual R-square=0



...school & individual R-square=0.49



...school & individual R-square=0.81

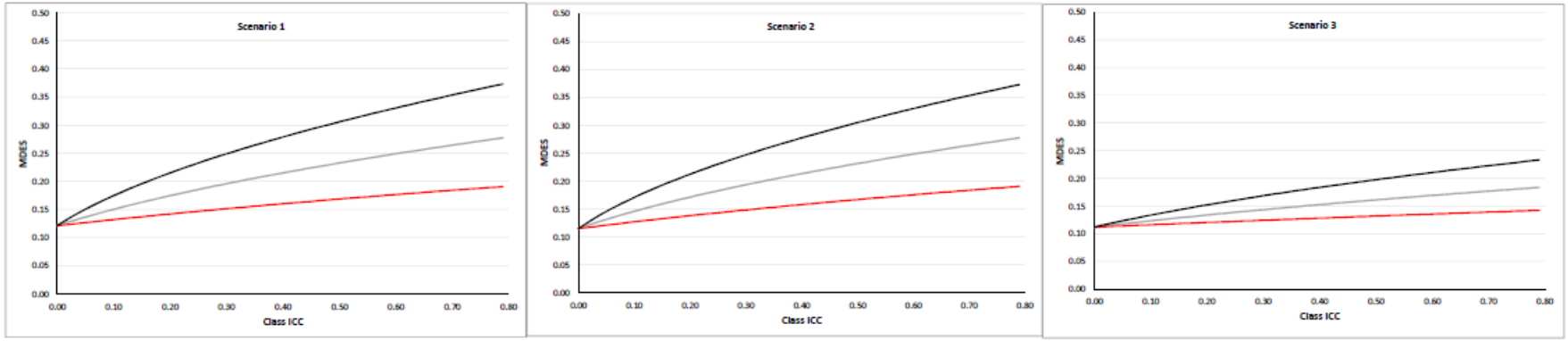


Figure 1b illustrates the following:

- Class-level explanatory power mitigates the impact of class level clustering
- The number of classes-per-school has a stronger influence on sensitivity compared with the number of pupils per class (i.e. it is methodologically preferable to have smaller samples across many classes compared with a few complete classes).
- Class-level explanatory power is particularly important when the number of classes per school is relatively small - the extent to which class-level explanatory power mitigates the increase in MDES with increasing class-level clustering diminishes with increasing numbers of classes.

It is worth noting that testing a small group of pupils within a class can bring practical and ethical problems that might be avoided by testing the whole class. So, whilst it is methodologically preferable to have a small samples of pupils spread across many classes, practicality suggests that it would be preferable to test all pupils in many classes - which has testing cost implications

Perspective 2: Detecting an MDES of 0.20 sds

The relationship between class level clustering and statistical power

Assuming a 2-level two-armed balanced CRT design (or a 3-level design where $ICC_{class} = 0.00$), the sample sizes needed to detect an MDES of 0.20 standard deviations have been calculated using equation 2.0 (or 3.0) for different strengths of covariate explanatory power and school-level clustering.

Table 5a below focuses on a 3-level two-armed balanced CRT design with 2 classes of 25 pupils per school (i.e. scenario 2, 50 pupils per school) and how class level clustering might impact on statistical sensitivity and power. For simplicity, in Table 5a, the same covariate explanatory power is assumed at all levels (i.e. $R^2_{ALL} = R^2_{sch} = R^2_{class} = R^2_{pup}$)

- At the top of Table 5a, the number of schools (**K**) required for an MDES of 0.20 sds when $ICC_{class} = 0.00$ is shown for two strengths of school-level clustering ($ICC_{Sch} = 0.05$ and $ICC_{Sch} = 0.20$).
- When school level clustering is relatively weak ($ICC_{Sch} = 0.05$), if class-level clustering is ignored, the number of schools needed to detect an MDES of 0.20 varies between 14 (when R^2_{ALL}=0.81) and 54 (when R^2_{ALL}=0.00).

- When school level clustering is stronger ($ICC_{Sch} = 0.20$), if class-level clustering is ignored, the number of schools needed to detect an MDES of 0.20 varies between 34 (when R^2_{ALL}=0.81) and 164 (when R^2_{ALL}=0.00).

Table 5a shows the statistical power for detecting an effect size of 0.20 or higher as significant (p<0.05) with a 3-level CRT design when class ICC varies between 0.00 and 0.50 . Alongside this are the actual MDES estimates for a 3-level CRT design. Finally, at the bottom of Table 5a, the number of schools that would be needed to detect an effect size of 0.20 sds or higher as statistically significant with a power of 80% is shown (and how many more schools this is to when class clustering is ignored or not present).

Table 5a: Number of schools required for MDES of 0.20 sds or higher.
Assuming 2 classes of 25 pupils per school.& same explanatory power at school, class and pupil levels

K: *number of schools required to detect MDES of 0.20 sds ignoring class level*

School level ICC of 0.05.

K=	*54*	*30*	*14*

R^2 *Explanatory Power (school, class & pupil levels)*

R^2=	*0.00*	*0.49*	*0.81*

Class level ICC	***Statistical power [actual MDES with 80% power]***		
0.00	80%	82%	86%
0.05	68% [0.24]	69% [0.23]	71% [0.22]
0.10	58% [0.27]	59% [0.26]	59% [0.25]
0.20	44% [0.32]	44% [0.31]	42% [0.29]
0.50	26% [0.43]	25% [0.42]	22% [0.40]

Class level ICC	***Total number of schools needed [Number additional to K]***		
0.05	72 [+18]	38 [+8]	16 [+2]
0.10	90 [+36]	48 [+18]	20 [+6]
0.20	126 [+72]	66 [+36]	26 [+12]
0.50	234 [+180]	120 [+90]	46 [+32]

School level ICC of 0.20.

K=	*164*	*84*	*34*

R^2 *Explanatory Power (school, class & pupil levels)*

R^2=	*0.00*	*0.49*	*0.81*

Class level ICC	***Statistical power [actual MDES with 80% power]***		
0.00	80%	80%	82%
0.05	76% [0.22]	74% [0.22]	71% [0.21]
0.10	72% [0.23]	69% [0.23]	62% [0.22]
0.20	65% [0.25]	60% [0.25]	49% [0.24]
0.50	49% [0.30]	42% [0.30]	30% [0.29]

Class level ICC	***Total number of schools needed [Number additional to K]***		
0.05	182 [+18]	94 [+10]	38 [+4]
0.10	200 [+36]	104 [+18]	40 [+6]
0.20	236 [+72]	122 [+38]	48 (+14]
0.50	344 [+180]	176 [+92]	68 [+34]

When school level clustering is relatively weak (ICC_{sch}=0.05)

A 2-level design that ignored class level clustering (ICC_{class}=0.00) and zero explanatory power (R^2=0.00) would need 54 schools to detect an effect size of 0.20 sds or higher as statistically significant (p<0.05) with a statistical power of 80% or higher. Even when class level clustering is weak (ICC_{class}=0.05), the statistical power for detecting an effect size of 0.20 is drops to 68%. 54 schools would be able to detect an MDES of 0.24 rather than 0.20 sds; it would take an additional 18 schools (K=72) for this design to be able to detect an MDES of 0.20 sds with 80% power. If class level clustering was stronger (ICC_{class}=0.20), statistical power for detecting an effect size of 0.20 drops further to 44%; 54 schools would be able to detect an MDES of 0.31 rather than 0.20 sds and it would take an additional 72 schools (K=126) to detect an MDES of 0.20 sds with 80% power.

When weak school level clustering (ICC_{sch}=0.05) but high covariate explanatory power (R^2=0.81), when ICC_{class}=0.00, 14 schools would be needed to detect an MDES of 0.20sds. When (ICC_{class}=0.05), the statistical power for detecting an effect size of 0.20 drops to 74%; 20 schools would be able to detect an MDES of 0.22 rather than 0.20 sds and it would take an additional 2 schools (K=16) to detect an MDES of 0.20sds with a power of 80%. If class level clustering was stronger (ICC_{class}=0.20), statistical power to detect an effect size of 0.20 drops further to 48% and it would take an additional 12 schools (K=26) to detect an MDES of 0.20 sds with 80% power.

When school level clustering is relatively strong (ICC_{sch}=0.20)

A 2-level design that ignored class level clustering (ICC_{class}=0.00), with zero covariate explanatory power, 164 schools would be needed to detect an MDES of 0.20 sds. When class level clustering is weak (ICC_{class}=0.05), the statistical power to detect an effect size of 0.20 drops to 76%; 164 schools would be able to detect an MDES of 0.22 and it would take an additional 18 schools (K=182) to detect an MDES of 0.20sds with 80% power. If class level clustering was stronger (ICC_{class}=0.20), statistical power to detect an effect size of 0.20 drops further to 65% and it would take an additional 72 schools (K=236) to detect an MDES of 0.20 sds with 80% power.

With strong school clustering (ICC_{sch}=0.20) and high covariate explanatory power (R^2=0.81), if class level clustering is ignored 34 schools would be needed to detect an MDES of 0.20 sds. When class level clustering is weak (ICC_{class}=0.05), the statistical power to detect an

effect size of 0.20 drops to 77%.; 34 schools would be able to detect an MDES of 0.21 rather than 0.20 standard deviations and it would take an additional 4 schools (K=38) for this design to be able to detect an MDES of 0.20sds with 80% power. If class level clustering was stronger (ICC_{class}=0.20), statistical power for the MDES of 0.20 drops further to 66% and it would take an additional 14 schools (K=48) to detect an MDES of 0.20 sds with 80% power.

As noted earlier, with increasing school level clustering, the possible proportion of clustering at the class and/or individual level decreases. For example, when school ICC=0.05, ICC values at class and/or individual levels could not be greater than 0.95; when school ICC=0.20, ICC values at class and/or individual levels could not be greater than 0.80 etc. Additionally, the statistical sensitivity for both a 2-level and 3-level design is influenced more by school-level compared with class-level clustering. Therefore, if a 2-level trial is powered to account for school clustering, this will provide some protection from the impact of class level clustering on statistical sensitivity; the greater this clustering at the school level is, the greater the protection it provides.

The estimates shown in Table 5a include a sizable assumption regarding explanatory power. In reality, it is highly unlikely that the same strength of covariate explanatory power would be found at all levels. However, at this point of 'equivalence' (i.e. $R^2_{ALL} = R^2_{sch} = R^2_{class} = R^2_{pup}$), a 2-level trial would be powered for this at both school and pupil levels and the inclusion of an additional class level with the same explanatory power does not provide additional protection against the impact of class clustering on statistical sensitivity. However, if $R^2_{class} > R^2_{Sch}$ some additional sensitivity would be gained and if $R^2_{class} < R^2_{Sch}$ sensitivity would be lost.

Table 5b looks more directly at how covariate explanatory power at the class-level mitigates against the impact of class-level clustering on statistical power, sensitivity and sample size. Assuming a school level ICC of 0.20 and 2 classes of 25 pupils per school, and fixing the school and pupil level explanatory power at three levels (0.00, 0.49 and 0.81), Table 5b shows the impact of class level clustering and covariate explanatory power on statistical power, sensitivity and required sample size.

Table 5b: Number of schools required for MDES of 0.20 sds or higher.
Assuming 2 classes of 25 pupils per school, school ICC=0.20, same explanatory power at school and pupil levels .
Class-level explanatory power allowed to vary between 0.00 and 0.81

School & Pupil level explanatory power = 0.00; K=164 schools.

R^2 Explanatory Power (class level only)

R^2=	*0.00*	*0.49*	*0.81*

Class level ICC	*Statistical power [actual MDES with 80% power]*		
0.00	80%	80%	80%
0.05	76% [0.22]	78% [0.21]	79% [0.21]
0.10	72% [0.23]	76% [0.22]	79% [0.21]
0.20	65% [0.25]	72% [0.23]	77% [0.21]
0.50	49% [0.30]	62% [0.25]	74% [0.22]

Class level ICC	*Total number of schools needed [Number additional to K]*		
0.05	182 [+18]	174 [+10]	168 [+4]
0.10	200 [+36]	182 [+18]	170 [+6]
0.20	236 [+72]	200 [+36]	176 [+12]
0.50	344 [+180]	252 [+88]	192 [+28]

School & Pupil level explanatory power = 0.49; K=84 schools.

R^2 Explanatory Power (class level only)

R^2=	*0.00*	*0.49*	*0.81*

Class level ICC	*Statistical power [actual MDES with 80% power]*		
0.00	80%	80%	80%
0.05	70% [0.23]	74% [0.22]	77% [0.21]
0.10	62% [0.25]	69% [0.23]	74% [0.21]
0.20	50% [0.28]	60% [0.25]	68% [0.22]
0.50	31% [0.37]	42% [0.30]	55% [0.24]

Class level ICC	*Total number of schools needed [Number additional to K]*		
0.05	104 [+20]	94 [+10]	88 [+4]
0.10	122 [+38]	104 [+18]	92 [+8]
0.20	158 [+74]	122 [+38]	98 [+14]
0.50	268 [+184]	176 [+92]	116 [+32]

School & Pupil level explanatory power = 0.81; K=34 schools.

R^2 *Explanatory Power (class level only)*

R^2=	*0.00*	*0.49*	*0.81*

Class level ICC	***Statistical power [actual MDES with 80% power]***		
0.00	82%	82%	82%
0.05	57% [0.25]	65% [0.23]	71% [0.21]
0.10	42% [0.30]	53% [0.26]	62% [0.22]
0.20	28% [0.37]	38% [0.30]	49% [0.24]
0.50	15% [0.53]	21% [0.41]	30% [0.29]

Class level ICC	***Total number of schools needed [Number additional to K]***		
0.05	52 [+18]	44 [+10]	38 [+4]
0.10	70 [+26]	52 [+18]	40 [+6]
0.20	108 [+74]	72 [+28]	48 (+14]
0.50	218 [+184]	128 [+94]	68 [+34]

With zero school and pupil level covariate explanatory power, 164 schools would be sufficient to detect an MDES of 0.20 sds if no class-level clustering were present. With zero class-level covariate explanatory power, the relationship between class level clustering and statistical power is identical to that shown in Table 5a (falling from 76% when class ICC=0.05 to 49% when class ICC=0.50). With high class-level covariate explanatory power ($R^2_{class} = 0.81$) this loss of power is reduced (falling from 79% when class ICC=0.05 to 72% when class ICC=0.50).

With high school and pupil level covariate explanatory power($R^2_{sch} = R^2_{pup} = 0.81$) 34 schools would be sufficient to detect an MDES of 0.20 sds if no class-level clustering were present. With high class-level covariate explanatory power ($R^2_{class} = 0.81$), the relationship between class level clustering and statistical power is identical to that shown in Table 5a (falling from 71% when class ICC=0.05 to 30% when class ICC=0.50). With zero class-level covariate explanatory power ($R^2_{class} = 0.81$) this loss of power is increased (falling from 57% when class ICC=0.05 to 15% when class ICC=0.50).

Figure 2a presents three charts plotting the relationship between class-level covariate explanatory power (X axis) by statistical power (Y axis). A chart is shown for three strengths

of school and pupil level explanatory power ($R^2_{sch} = R^2_{pup} = 0.00; 0.49 \& 0.81$). For each chart, a line is shown for five strengths of class-level clustering (ICC_{class}=0.00 to 0.50). For example, when zero class-level clustering is present, 164 schools are needed to detect an effect size of 0.20 sds or higher with a statistical power of 80% or higher (shown by the horizontal line in Figure 2a). If zero class-level explanatory power is assumed, the negative impact of class-level clustering on statistical power can be seen where each of the six lines cross the Y-axis. As class-level covariate explanatory power increases, statistical power is seen to tend upwards towards the 80% limit.

It seems apparent that if a CRT is designed and powered ignoring clustering at the class-level, gains brought by covariate explanatory power (at school / pupil levels) can be lost by an increasing negative influence of class level clustering on statistical power.

Figure 2a: Class-level R-square (X) v Statistical Power (Y) to detect an MDES of 0.20 sds

School ICC=0.20 / a=0.05 (two tailed, single outcome) / 2 classes of 25 pupils per school

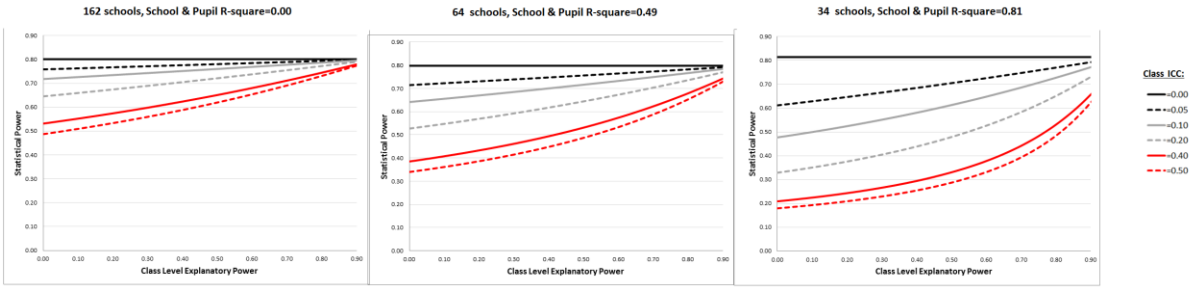
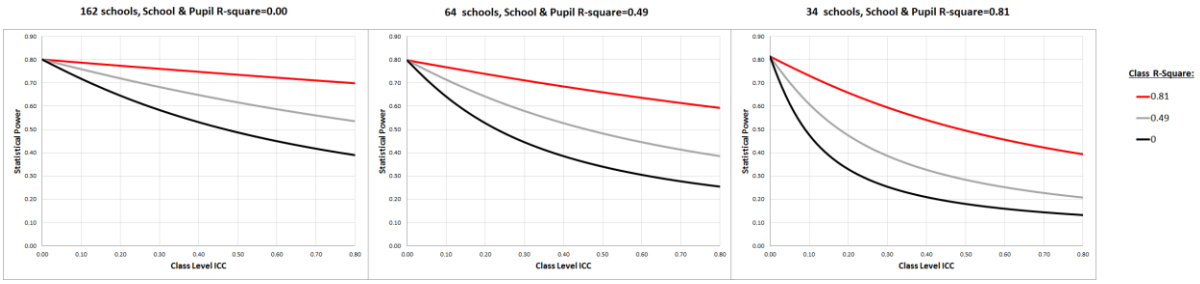


Figure 2b presents three charts plotting the relationship between class-level clustering (X axis) by statistical power (Y axis). A chart is shown for three strengths of school and pupil level explanatory power ($R^2_{sch} = R^2_{pup} = 0.00; 0.49 \& 0.81$). For each chart, a line is shown for three strengths of class-level explanatory power (R^2_{class}=0.00; 0.49 & 0.81). For example, when zero class-level clustering is present, 164 schools are needed to detect an effect size of 0.20 sds or higher with a statistical power of 80% or higher (shown by the horizontal line in Figure 2a). If zero class-level clustering is assumed, the positive impact of class-level explanatory power on statistical power can be seen where the three lines cross the Y-axis. As class-level clustering increases, statistical power is seen to tend downwards away from

80% - but this drop in power is less when there is high class-level covariate explanatory power.

Figure 2b: Class-level ICC (X) v Statistical Power (Y) to detect an MDES of 0.20 sds

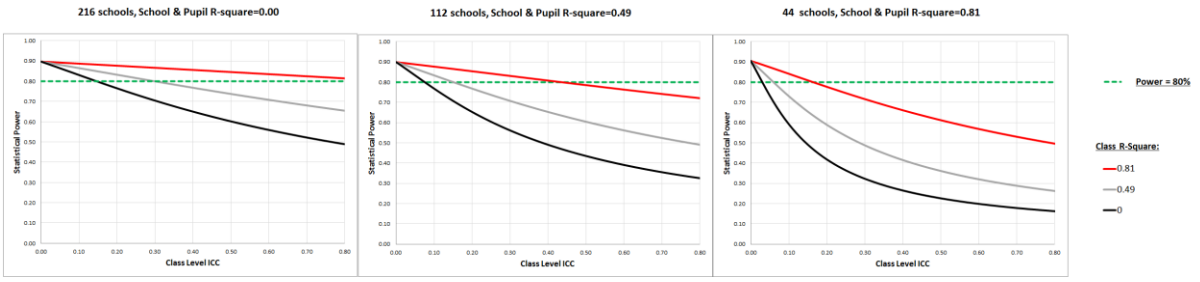
School ICC=0.20 / a=0.05 (two tailed, single outcome) / 2 classes of 25 pupils per school



One practical method that would provide a buffer against the impact of (hidden) class level clustering in a CRT design would be to use a higher statistical power for the 2-level design. Figure 2c replicates Figure 2b but assumes a statistical power of 90% or higher. This inevitably results in an increase in the required number of schools than would be needed compared with 80% statistical power. Figure 2c shows how doing this would ensure a statistical power above 80% up to a particular strength of class-level clustering which is determined by how the strength covariate explanatory power. For example, if zero school/pupil level explanatory power and class level clustering is assumed, 216 schools would be needed to detect an effect size of 0.20 or higher as significant with a power of 90% or higher. With zero class-level explanatory power, it is not until a class-level ICC of 0.14 or higher before power drops below 80%. With moderate class-level covariate explanatory power (R^2_{class}=0.49) it is not until a class-level ICC of 0.30 or higher before power drops below 80%.

Figure 2c: Class-level ICC (X) v Statistical Power (Y) to detect an MDES of 0.20 sds assuming a statistical power of 90%

School ICC=0.20 / a=0.05 (two tailed, single outcome) / 2 classes of 25 pupils per school



Figures 2a to 2c highlights X things:

- The impact of class level clustering on statistical power weakens with increasing school level clustering.
- If the sample size of a 2-level CRT sample size is selected with reference to covariate explanatory power and, within a 3-level CRT design, the same explanatory power is assumed for all levels, the relationship between class-level clustering and statistical power is not mitigated further by class level explanatory power.

5. Discussion

From a theoretical perspective, class-level clustering has been shown to 'matter' for the design of educational trials in terms of statistical sensitivity, sample size and/or statistical power. If an intervention being evaluated seeks to cause change in pupil level attainment through 'something' (e.g. the introduction of teaching materials, new methods and/or pedagogue) introduced into classrooms (or in other within-school pupil clusters), failing to account for class-level clustering will result in underestimated MDES statistics and/or a loss of statistical power for a design to detect a specified MDES. The loss of sensitivity and statistical power grows with increasing strength of class-level clustering can be mitigated through the use of covariate explanatory power and/or increasing the size of the trial

Seven 3-level CRTs with school level randomisation and a class/teacher/TA level that have been completed or ongoing in England were identified. To date, only three of these trials have reported but two more are due to report in 2019. This means that the empirical evidence base to draw on for the design of 3-level CRTs is currently rather limited. Research has shown how variance structure is dependent on the number of levels included in a multilevel design (e.g. Hutchison & Healy, 2001) and that variance from ignored levels of clustering is redistributed to other levels. This means that drawing on guidance/past trials for 2-level CRT designs for providing estimates for school-level clustering for 3-level designs will be inaccurate. To improve accuracy, 3-level designs should draw on guidance and past trials for 3-level CRTs. From the limited number of 3-level CRTs conducted, there is a suggestion that class-level clustering is stronger in secondary schools compared with primary schools and this seems to reflect pervading patterns in the use of setting and/or steaming in English schools, and therefore subjects in which settings is most common.

The most striking strength of class-level clustering was observed in the evaluation of the Multiplicative Reasoning Project (MRP) amongst Y7, Y8 and Y9 pupils (Boylan et al., 2015). The MRP evaluation can be used to illustrate differences in the variance decomposition of 2-level and 3-level CRT designs that have been noted by Hutchison & Healy (2001) and others. Table 6 shows this for the three MRP pupil cohorts and shows variance decomposition and ICC values for a 3-level design (school, class & pupil) and when class level is dropped for a 2-

level design (school & pupil). The table echoes what has been reported in the literature to show how variance in the ignored level (class) is redistributed to both school and pupil levels.

Table 6: Variance decomposition & ICC estimates
3-level and 2-level CRT designs compared
MRP evaluation, outcome = GL PTiM

	Y7		Y8		Y9	
Original Sample						
Variance decomposition	3-lev	2-lev	3-lev	2-lev	3-lev	2-lev
School Level	51.6	86.5	67.5	130.6	4.6	89.2
Class Level	101.3	/	117.7	/	141.1	/
Pupil Level	89.4	143.0	65.0	119.4	56.2	117.4
Total	242.3	229.6	250.2	250.0	201.9	206.6
ICC_{sch}	0.21	0.38	0.27	0.52	0.02	0.43
ICC_{class}	0.42	/	0.47	/	0.70	/
ICC_{pup}	0.37	0.62	0.26	0.48	0.28	0.57

	Y7		Y8		Y9	
Restricted Sample *(Comprehensive schools with 2+ classes of 10+ pupils)*						
Variance decomposition	3-lev	2-lev	3-lev	2-lev	3-lev	2-lev
School Level	1.3	51.4	4.6	86.3	0.0	48.6
Class Level	110.6	/	154.9	/	120.3	/
Pupil Level	92.8	150.0	61.1	134.7	51.0	120.8
Total	204.7	201.4	220.6	221.0	171.3	169.4
ICC_{sch}	0.01	0.26	0.02	0.39	0.00	0.29
ICC_{class}	0.54	/	0.70	/	0.70	/
ICC_{pup}	0.45	0.74	0.28	0.61	0.30	0.71

It seems that whilst the 3-level models clearly show the highest concentration of variance at the class-level, when this is ignored the variance is redistributed to both school and pupil levels. Therefore, estimates for ICC values for 3-level designs at both school and class level should draw on past 3-level trials rather than assuming the same school-level ICC that might be used for a 2-level design. In order to do this, a larger bank of 3-level CRT trials will be needed. In the meantime, some proposed class level ICC values are given in the next section

Suggested class-level ICC values

A range of suggested class-level ICC values are given below - it is important that these are udpdated as the evidence base grows. Class level ICC values are given for school phases and key stages:

- Within the Primary schooling phase,
    - KS1 - class ICC of 0.05
    - KS2 - class ICC = 0.10

These can only relate to trials that recruit Primary schools with 2+ classes per year. It should be noted that is is relatively common for Primary schools to have one class per year (single-form entry). For these schools it would mean that, within each year group, the class and school level would be identical and so a 3-level CRT design is not feasible unless the trial involves multiple year groups.

- Within the Secondary schooling phase,
    - Y7-Y11 Maths - class ICC=0.50
    - Y7-Y11 English/Science - class ICC=0.40
    - Y7-Y11 other subjects - class ICC=0.20

Class level clustering of outcome data will be a result of setting/streaming and other factors such as the teacher(s), classroom environment etc. It does seem that class level clustering is particularly strong in mathematics but this draws on a single CRT with limitions (Boylan et al., 2015). However, given that the cluster pattern seems to reflect what is known about the widespread use of setting/streaming in maths in Secondary schools in England, it seems wise to assume a class level ICC of 0.50 until evidence suggests otherwise.

We have less detail on class level clustering in other subjects from educational trials. However, Table 7 illustrates how the use of setting/streaming is also widespread in Science and English using OFSTED inspection data from 2010. It seems unlikely that there has been a move away from segregating pupils by perceived or measured ability in England in the decade since the data shown in Table 6 were collected. Until we have evidence to suggest otherwise, it seems advisable to assume a class level ICC of around 0.40 for these subjects.

Table 7: Percentage of lessons observed by OFSTED in 2010 that grouped pupils by measured or perceived ability.

	All Subjects	Maths	Science	English
Y1	5.9	5.8	1.6	7.8
Y2	8.2	10.7	1.0	9.4
Y3	10.4	17.1	1.6	9.9
Y4	11.0	19.0	1.4	10.0
Y5	15.2	26.2	1.5	12.3
Y6	21.7	33.7	3.3	19.4
Y7	35.1	61.8	45.3	49.4
Y8	44.2	72.6	63.4	56.7
Y9	46.6	73.8	64.7	59.0
Y10	42.5	72.0	65.3	60.7
Y11	45.0	73.8	64.8	64.8
Primary Phase (Y1-Y6)	11.8	18.6	1.7	11.3
KS1 (Y1-Y2)	6.9	8.1	1.3	8.5
KS2 (Y3-Y6)	14.5	24.0	1.9	12.9
Secondary Phase (Y7-Y11)	42.7	70.8	61.5	58.3
KS3 (Y7-Y9)	42.2	69.2	58.6	55.2
KS4 (Y10-Y11)	43.6	72.8	65.1	62.6

Source: https://giftedphoenix.wordpress.com/2014/11/12/the-politics-of-setting/ & https://www.theyworkforyou.com/wrans/?id=2011-07-20a.340.4&s=%28ability+grouping+schools%29+section%3Awrans+section%3Awms#g340.6

Estimates for school-level ICC values might initially draw on the current evidence from 2-level trials but given what is shown in Table 6, a greater understanding of how attainment (and other) data is clustered at school and class levels will be needed.

Practicalities

At the start of this review, having struggled to identify 3-level CRTs, I contacted EEF to request details on EEF trials that included a class and school level. The EEF similarly struggled to identify 3-level CRTs and sought help/advice from the EEF advisory group. The advisory group responded with some reflections on the problems of obtaining class level ID detail:

> "There are limitations with the submitted data for teaching groupings – we thought about asking for class IDs, but it is not straightforward. Consider the following scenarios
>
> 1. Class/ registration group remains constant for all lessons (e.g. primary school which does not set)
> 2. Registration group differs from subject groupings (e.g. tutor group registration in secondary schools)
> 3. Class groupings differ for literacy and numeracy lessons (e.g. setting in primary schools – in some cases a primary pupil would be in 3 different "classes")
> 4. Class groupings differ for different subjects (e.g. secondary timetable)
>
> Then you add the complexity of intervention focus (e.g. literacy only, literacy and numeracy, whole school) and you can see that collecting the data could be a significant burden on schools and evaluators… We decided to abandon the attempt."

These are sizable practical problems and three further complexities can be added here.

5. When schools introduce setting/streaming during a trial period
6. Pupil and/or teacher movement between classes
7. Multiple teachers for one class

Having had experience in collecting class level details for four trials, I have some familiarity with the 'significant burden' for evaluators noted by the EEF advisory group. Collecting and checking details of classes, teachers and pupil class lists is very time consuming (and hence costly). However, collecting class level detail and in developing good-practice processes will bring methodological benefits to trial design in terms of greater accuracy in estimated MDES statistics and statistical power. Further, whilst perhaps pragmatic, ignoring class level clustering seems out-of-step within a 'gold standard' evaluation methodology; particularly given the extent of policies that Dracup (2014) calls 'within-school selection' and Francis (2017) calls 'pupil segregation' practiced in English schools (i.e. setting and streaming).

Setting/streaming

OECD identified that around a fifth of pupils in the UK are in selective schools but nearly all (99%) schools had policies of grouping pupils by measured/perceived 'ability' (OECD, 2012, Table 2.2 p57). Focusing solely on England, Dracup (2014) presented data from 2009/10 based on English, maths and science lessons observed by OFSTED inspectors and reported a greater proportion of lessons taught in 'ability' groups in Secondary schools (43%) compared with Primary Schools (12%) and a greater proportion of maths lessons using 'ability' grouping through Primary (19% overall; 26% in Y5; 34% in Y6) and Secondary (71% overall; 74% in Y11) compared with English (11% Primary; 58% Secondary) and Science (2% Primary; 62% Secondary). More recently, Francis et al. (2017) reported on the rarity of all attainment (or mixed ability) teaching (and entrenched use of 'ability' grouping) in the English education system and their difficulties in recruiting Secondary schools that either currently practiced mixed ability or were willing to introduce it for a recent EEF-funded study into grouping practices (Roy et al., 2018b). Finally, from data we collected in 2018 from 119 schools for a 3-level CRT evaluating the impact of a secondary (KS3) maths intervention, 106 schools (89%) reported to use setting or streaming within Y7 maths.

OECD highlight that segregation of pupils between schools (fee paying schools & selective state schools) and within schools (setting, streaming & banding) are key barriers to equity and quality in an education system.

"The evidence is conclusive: equity in education pays off. The highest performing

> education systems across OECD countries are those that combine high quality and equity. In such education systems, the vast majority of students can attain high level skills and knowledge that depend on their ability and drive, more than on their socio-economic background." OECD, 2012 p14.

It is relatively straight forward for designers of educational trials to avoid methodological issues that are brought by between-school pupil segregation. For example, trials might decide to only recruit state schools that operate a non-selective (comprehensive) admissions policy. This would help to minimise the strength of clustering in an outcome variable at the school level (ICC_{sch}). Given that the vast majority of state schools are non-selective, this does not bring too many problems. However, the entrenched use of within-school pupil segregation within 'comprehensive' schools means that the methodological issues brought by class-level clustering are less easy to side step (seemingly impossible in England). Simply ignoring clustering will not make the problems disappear but will bring hidden bias into the design of educational trials and therefore increase the risk of making incorrect conclusions about the impact of educational interventions.

Further, ignoring the issue of class-level clustering seems to jar with EEFs mission of "breaking the link between family income and educational achievement"(EEF, 2018). Pupil segregation (between and within schools) is not socially neutral and tends to result in the sorting (clustering) of pupils along socio-economic, ethnic, gender and SEND lines (Gillborn & Mirza, 2000; Kutnick et al., 2005). Further, it is established that setting and streaming does not result in better attainment outcomes for most pupils (Slavin, 1990; Francis, 2017). Therefore, pupil segregation seems to be a monolithic obstacle for "breaking the link" between income and educational success in England and deserves close attention. EEF has funded two evaluations looking directly at best practice in mixed ability and setting in Y7 and Y8 maths and English (Roy et al., 2018a & 2018b); both using 2-level CRT designs. Unfortunately, these evaluations suffered from either difficulties in recruiting schools (mixed ability) or drop-outs (setting) and so more research will be needed. It seems important that evidence used to inform educational policy at a national or school level is robust and therefore trial designs should reflect the structure of the system/schools in which they take place. The lack of a class (or teacher) level in the context of such

widespread within-school pupil segregation currently serves to undermine the robust nature of evidence from educational trials.

Class-level; Teacher-level

Through this paper, within-school clustering has been referred to as either class-level or teacher/TA level clustering. This was done to reflect how within-school clustering is described in evaluation reports and wider literature. However, it should be noted that these two descriptions are not necessarily the same. For example, one class may be taught by multiple teachers and one teacher might teach multiple classes. Some observational studies have looked at both class and teacher level clustering (Opdenakker & Van Damme, 2000; Noortgate et al., 2004) but no experimental studies with this (4-level) design were found. Given that these are still relatively early time for educational trials in England, future designs might explore such a design.

Many of the educational interventions that have been evaluated to date are directly targeted at teachers or TAs through professional development (PD) training with the hope/theory that this leads to change in the classroom and eventual gains in pupil level attainment. 2-level CRTs with just school and pupil levels do not adequately reflect the structural realities of teacher/TA PD interventions they used to evaluate. This paper illustrated what this means in terms of statistical sensitivity but 2-level CRTs are also problematic when it comes to being able to comprehensively scrutinise findings of impact evaluations.

RCT and CRT designs are widely identified as a 'gold standard' for evaluating impact and are methodologically unique in being able capture educational effects that have been 'caused' by an educational intervention. However, to unpack the mechanism(s) that led to (cause) the observed 'effects', a theory of change is drawn on which commonly will include components at teacher and/or classroom levels.

Alongside impact evaluations, Implementation and Process (IPE) evaluations usually gather data at these levels in order to help shine some light onto such causal mechanism(s). However, with a 2-level CRT design, follow-on impact analyses (e.g. on treatment or CACE)

are unable to draw on this data directly and are forced to aggregate class/teacher level data to the school level. This means that for these follow-on analyses, there is no acknowledgement of within-school variation (for example, differences in how teachers within one school engaged with an intervention and how their classes of pupils experienced it). This structural disconnect seems remarkably imprecise/fuzzy in the robust world of RCTs. 3 (possibly 4) level CRT designs would enable data at teacher and/or class levels to be included at the appropriate level for follow-on analyses and therefore help to pinpoint causal mechanism(s) behind any observed effects.

Conclusions

So, does the classroom level 'matter' in the design of educational trials? From a theoretical perspective the answer is that if within-school clustering (class/teacher/TA) exists, failing to account for this in the design and analyses of educational trials will result in drawing inaccurate conclusions about the impact of educational interventions; so yes, the classroom level matters. Empirically, it is not appropriate to draw firm conclusions because only seven 3-level CRTs have been conducted in England. However, the patterns observed within these few 3-level trials do reflect deeply engrained policies of setting and/or streaming in English schools; manifesting in sizable class-level clustering particularly in secondary schools. In primary schools, class-level clustering seems weaker and this reflects more limited use of setting here (although Dracup, 2014 did report over a third of KS2 Y6 maths classes took place in 'ability' groups in 2010). Clustering might be weaker in primary, but the practical problems of collecting data at the class and teacher level are less acute. This is because primary schools are smaller and there is less variation and movement of teachers (and pupils) between classes. So, in terms of statistical sensitivity, a class/teacher/TA level seems less important for evaluations of interventions in primary schools compared with secondary schools but the feasibility of collecting class-level details in order to accurately reflect the nested structure of schools-teacher/class-pupil is greater in primary compared with secondary schools. Given that a 3-level design brings greater precision for follow-on analyses exploring possible mechanism(s) behind observed effects, this seems worthwhile. In secondary schools the empirical evidence base is currently limited to a single 3-level CRT evaluating a KS3 maths intervention (Boylan et al., 2015). A second 3-level CRT evaluating a KS4 secondary maths intervention is not due to report until 2020. From experience, I know

that collecting class-level detail in secondary schools is more complex than primary. This is because secondary schools are generally larger and there is much greater variation of teachers (and pupils) between classes. However, given that sizable class-level clustering is very likely to be a structural reality in secondary schools, it seems highly problematic to ignore it. This is perhaps more acute in some subject areas where setting/streaming is widely practiced (maths, English) compared with other subject areas (history, art).

Complexity in the collection of class/teacher/TA level data and the associated burden on schools and evaluators will be most intense now because this is new. Schools are relatively experienced in providing pupil and school level data for external scrutiny but experience in providing details at a class level will be more limited. It seems likely that, within schools, variations (in pupil attainment) at a class and/or teacher level will be examined (e.g. by heads of departments or schools or MATs) but the reality may be mixed practices across schools. Further, within the context of the aftermath of GDPR and an increasingly marketised education system with measured school performance, effectiveness, commercial sensitivity and OFSTED, schools may be reticent / unwilling to provide class-level detail. However, in the few 3-level CRTs undertaken, reticence to providing class-level detail is not reported. As more 3-level trials are conducted, evaluators and schools will gain experience and this will help to develop approaches which both minimise burden and accurately capture the nested structure of schools involved in educational trials.

At this point in time it seems that, theoretically, including a class and/or teacher level into trial designs (particularly trials evaluating teacher PD programmes) will result in improved accuracy in trial designs but the empirical realities are less clear. One final point is that none of the 3-level CRT trials that have reported so far have included explanatory power at all three levels in the final analyses. Hedges & Hedberg (2013) recommend that an orthogonal approach is taken such that the pre-test is centred differently for each level: i.e. School level means centred around the overall school-level mean; class level means centred around the school level mean and pupil level raw scores centred around the class level mean. This approach ensures that the variables can only account for variance in an outcome at the level at which they are included. This approach may lead to additional gains in explanatory

power at school, class or pupil levels (and hence improved statistical sensitivity in the design) but this is yet to be seen empirically.

A pragmatic way forward is to encourage and fund more 3-level CRTs in order to build the empirical evidence base and use these trials to develop good practice guidance for the design, analyses and data collection methods for 3-level clustered educational trials. This would lead to wider benefits for educational research that are tangential to evaluation of educational interventions but aligned to EEFs mission; by providing data which can be used to examine patterns and practices of grouping/segregating of pupils in schools in England. Experiences in collecting this detail might be drawn upon for a larger observational study to measure class-level clustering across the English education system.

References

Atkins, M. (2017) Statistical Analysis Plan for Catch-Up Numeracy. https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/catch-up-numeracy-2015/

Atkins, M. (2018) Statistical Analysis Plan for Mathematics in Context. https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/maths-in-context/

Bloom, H.S. (1995) Minimum Detectable Effects: a simple way to report the statistical power of experimental designs. Evaluation Review 19(5) pp547-556.

Bloom, H.S. (2006) The Core Analytics of Randomized Experiments of Social Research. MDRC Working Papers on Research Methodology, August 2006. Available at https://www.mdrc.org/publication/core-analytics-randomized-experiments-social-research

Bloom, H, S., Richburg-Hayes, L. & Black, A. R. (2007) Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions Educational Evaluation and Policy Analysis 29(1) pp 30–59

Boylan, M., Demack, S., Willis, B., Stevens, A., Adams, G. & Verrier, D. (2015) Multiplicative Reasoning professional development programme: evaluation. https://www.gov.uk/government/publications/multiplicative-reasoning-professional-development-programme

Boylan, M., Demack, S., Wolstenholme, C, Reidy, J. & Reaney-Wood, S. (2018) ScratchMaths Evaluation Report. https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/scratch-programming/

Cabinet Office. 2013. “What Works Network.” https://www.gov.uk/guidance/what-works-network.

Connolly, P., Keenan, C. & Urbanska, K. (2018) The trials of evidence-based practice in education: a systematic review of randomised controlled trials in education research 1980-2016. Educational Research, 60(3) pp276-291

Dong, N. & Maynard, R. (2013) PowerUp: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-Experimental Design Studies, Journal of Research on Educational Effectiveness, 6(1) pp24-67

Dracup, T. 2014. “The Politics of Setting.” Available at: https://giftedphoenix.wordpress.com/2014/11/12/the-politicsof-setting/ .

EEF (2018) Statistical analysis guidance for EEF evaluations available at https://educationendowmentfoundation.org.uk/public/files/Grantee_guide_and_EEF_policies/Evaluation/Writing_a_Protocol_or_SAP/EEF_statistical_analysis_guidance_2018.pdf

Francis, B., L. Archer, J. Hodgen, D. Pepper, B. Taylor, and M.-C. Travers. (2016). Exploring the Relative Lack of Impact of Research on ‘Ability Grouping’ in England: A Discourse Analytic Account. Cambridge Journal of Education: 1–17.

Gillborn, D. & Mirza, H. (2000) Educational Inequality: Mapping race, class and gender. London: Ofsted. Available at http://dera.ioe.ac.uk/4428/

Goldacre, B. 2013. “Building Evidence into Education.” London. https://www.gov.uk/government/news/building-evidence-into-education.

Hammersley, M. 2007. Educational Research and Evidence-Based Practice. London: Sage.

Hedges, L.V. & Hedberg, E.C. (2013) Intraclass correlations and covariate outcome correlations for planning two- and three-level cluster-randomised experiments in education. Evaluation Review 37(6) pp445-489.

Hedges, L.V. & Rhoads, C. (2010) Statistical Power Analysis in Education Research. NCSER 2010-3006. Available at https://ies.ed.gov/ncser/pubs/20103006/pdf/20103006.pdf

Hill, P.W. & Rowe, K.J. (1996) Multilevel modelling in school effectiveness research. Journal of School Effectiveness and School Improvement 7(1) pp1-34.

Hussain, F., Wishart, R., Marshall, L., Frankenberg, S., Bussard, L., Chidley, S., Hudson, R., Votjkova, M. & Morris, S. (2017) Family Skills Evaluation Report. https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/family-skills/

Hutchison, D. & Healy, M. (2001). The effect of variance component estimates of ignoring a level in a multilevel model. Multilevel Modelling Newsletter, 13, 4-5. available at https://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/new13-2.pdf

Jay, T., Willis, B., Thomas, P., Taylor, R., Moore, N., Burnett, C., Merchant, G. & Stevens, A. (2017) Dialogic Teaching Evaluation Report.

https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/dialogic-teaching/

Kelcey, B., Spybrook, J., Phelps, G., Jones, N. & Zhang, J. (2017) Designing large scale multisite and cluster randomized studies of professional development. The Journal of Experimental Education, 85(3) pp389-410

Konstantopoulos, S. (2008) The power of the test for treatment effects in three-level cluster randomized designs. Journal of Research on Educational Effectiveness, 1(1) pp66-88

Konstantopoulos, S. (2007) A Comment on Variance Decomposition and Nesting Effects in Two- and Three-Level Designs. IZA Discussion Paper No 3178. Available at http://ftp.iza.org/dp3178.pdf

Kutnick, P., J. Sebba, P. Blatchford, M. Galton, J. Thorp, H. MacIntyre, and L. Berdondini. 2005. The Effects of Pupil Grouping: Literature Review. London: DCSF. http://dera.ioe.ac.uk/18143/1/RR688.pdf

Lauer, S.A., Kleinman, K.P. & Reich, N.G. (2015) The effect of cluster size variability on statistical power in cluster-randomised trials. PLoS One 10(4) doi: 10.1371/journal.pone.0119074

Leckie, G. (2013) Module 12(concepts): Cross-classified multilevel models available from Bristol University Centre for Multilevel Modelling at http://www.bristol.ac.uk/cmm/learning/online-course/

Lortie-Forgues, (2017) What can we learn from RCTs in education? Presented at the conference of Randomised Controlled in the Social Sciences, York University Sept 7th 2017.

Luo, W. & Azen, R. (2013) Determining Predictor Importance in Hierarchical Linear Models Using Dominance Analysis. Journal of Educational and Behavioral Statistics, 38(1) pp.3-31

Moerbeek, M. (2004) The Consequence of Ignoring a Level of Nesting in Multilevel Analysis. Multivariate Behavioral Research, 39(1), pp129-149,

OECD (2012) Equity and Quality in Education. Supporting disadvantaged students and schools. available at https://www.oecd.org/education/school/50293148.pdf

Opdenakker, M-C. & Van Damme, J.V. (2000) The Importance of Identifying Levels in Multilevel Analysis: An Illustration of the Effects of Ignoring the Top or Intermediate Levels

in School Effectiveness Research. Journal of School Effectiveness and School Improvement 11(1) pp103-130.

Raudenbush, S.W., & Bryk, A. S. (2002). Hierarchical linear models. Applications and data analysis methods. Thousand Oaks, CA: Sage.

Raudenbush, S. W., Spybrook, J., Bloom, H.S., Congdon, R., Hill, C. & Martinez, A. (2011). Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01) [Software]. Available from www.wtgrantfoundation.org.

Recchia, A. (2010) R-squared Measures for Two-Level Hierarchical Linear Models Using SAS. Journal of Statistical Software 32(2) pp1-9.

Rosnow, R.L. & Rosenthal, R. (2003) Effect sizes for experimenting psychologists. Canadian Journal of Experimental Psychology 57(3) pp221-237

Roy, P., Styles, B., Walker, M., Morrison, J., Nelson, J. & Kettlewell, K. (2018a) Best Practice in Grouping Students Intervention A: Best Practice in Setting Evaluation report. Available at https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/best-practice-in-grouping-students/

Roy, P., Styles, B., Walker, M., Bradshaw, S., Nelson, J. & Kettlewell, K. (2018b) Best Practice in Grouping Students Intervention B: Mixed Attainment Grouping Pilot report. Available at https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/best-practice-in-mixed-attainment-grouping/

Schagen, I. & Elliot, K. (eds) (2004) But what does it mean? The use of effect sizes in educational research. NFER/IoE Available at https://www.nfer.ac.uk/publications/SEF01/SEF01.pdf

Seymour, K. & Morris, S. (2018) Trial Evaluation Protocol: Evaluating the effectiveness of Eedi (previously Diagnostic Questions or DQ) formative assessment programme on raising attainment in mathematics at GCSE. Available at https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/diagnostic-questions/

Slavin, R. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. Review of Educational Research, 60, 471–499.

Snijders, T. A. B., & Bosker, R. J. (1999). Multilevel analysis: An introduction to basic and advanced multilevel modeling. London: Sage.

Spybrook, J. & Raudenbus, S.W. (2009) An examination of the precision and technical accuracy of the first wave of group randomized trials funded by the Institute of Education Sciences. Education Evaluation and Policy Analysis 31(3) pp298-318

Spybrook, J., Bloom, H.S., Congdon, R., Hill, C. Martinez, A. & Raudenbush, S. (2011). Documentation for the Optimal Design Software. Available from www.wtgrantfoundation.org .

Spybrook, J., Shi, R. & Kelcey, B. (2016) Progress in the past decade: an examination of the precision of cluster randomized trials funded by the U.S. Institute of Education Sciences, International Journal of Research & Method in Education, 39 (3), 255-267

Torgerson, C. J., and D. J. Torgerson. 2001. "The Need for Randomised Controlled Trials in Educational Research." British Journal of Educational Studies 49: 316–329. doi:10.1111/1467-8527.t01-1-00178

Torgerson, D. J., and C. J. Torgerson. 2008. Designing Randomised Trials in Health Education and the Social Sciences. Basingstoke: Palgrave McMillan.

Tymms, P., Merrell, C. & Wildy, H. (2015) The progress of pupils in their first school year across classes and education systems. British Educational Research Journal 41(3) pp. 365–380

Worth, J., Sizmur, J., Walker, M., Bradshaw, S. & Styles, B. (2017) Teacher Observation Evaluation Report. https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/teacher-observation/

Torgerson, D., Torgerson, C., Mitchell, N., Buckley, H., Ainsworth, H., Heaps, C. & Jefferson, L. (2014) Grammar for Writing Evaluation Report. https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/grammar-for-writing/

Trammer, M. & Steel, D.G. (2001) Ignoring a level in a multilevel model: evidence from UK census data. Environment and Planning A 33 pp941-948

Van den Noortgate, W., Opdenakker, M-C. & Onghena, P. (2005) The Effects of Ignoring a Level in Multilevel Analysis. Journal of School Effectiveness and School Improvement 16(3) pp281-303.

Van Landeghem, G., De Fraine, B. & Van Damme, J. (2005) The Consequence of Ignoring a Level of Nesting in Multilevel Analysis: A Comment. Multivariate Behavioral Research, 40(4), pp423-434,

Worth, J., Sizmur, J., Walker, M., Bradshaw, S. & Styles, B. (2017) Teacher Observation Evaluation Report. https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/teacher-observation/