

Evaluating design features and analysing the intra-cluster correlation coefficients for pupil health outcomes in school-based cluster randomised controlled trials

Submitted by Kitty Parker to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Medical Studies in May 2023

This thesis is available for Library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature

A handwritten signature in cursive script, appearing to read 'K Parker', is written over a horizontal dotted line. The signature is centered on the page.

Dedication

To my wonderfully supportive family and friends.

Acknowledgements

Firstly, to my supervisors. Professor Obioha C Ukoumunne, you have been the most incredible supervisor, mentor and role model. You have been extremely generous with your time and knowledge, and I am truly grateful for your support during my PhD. Dr Michael Nunns, thank you for your guidance, friendship and humour throughout my PhD; I always look forward to our meetings together. Dr ZhiMin Xiao, your belief in my ability and wise words were of great support to me, and my thesis is infinitely stronger as a result of your input. Professor Tamsin Ford, it has been a pleasure to work with and learn from you, thank you for all your time and encouragement.

Thank you to my wonderful colleagues and fellow PhD students. Kate Allen, Mary Fredlund, Emily Taylor and Sara Eddy, thank you for your friendship and for sharing the PhD journey with me. To Saskia Eddy and Professor Sandra Eldridge, thank you for all your help and advice regarding pilot and feasibility trials, and your hard work on my systematic review. Thank you to Professor Tamsin Ford, Professor Paul Stallard, Professor Willem Kuyken and Dr Nick Axford for granting me use of their datasets.

Finally, thank you to my dear friends who have helped me retain balance and perspective over the past three years. To my brother, Charlie, thank you for providing positive and welcome distraction from my PhD. Thank you to my parents who have supported me to follow my dream of completing a PhD, I could not have done it without you.

Page intentionally left blank

Abstract

Cluster randomised trials (CRTs) are used in schools to evaluate interventions for improving pupil health outcomes. Little is known about the methodological practices of these studies and plausible values of the intra-cluster correlation coefficient (ICC) of pupil outcomes to inform sample size calculation for CRTs.

Systematic reviews were undertaken to identify the practices of definitive and feasibility CRTs. ICC estimates for pupil health outcomes were collated from published reports of school-based CRTs worldwide, and the relationships between these and the design and contextual characteristics of the studies were examined. A secondary analysis of raw data from five UK school-based CRTs explored patterns in ICCs for pupil social emotional functioning outcomes.

The rate of publication of school-based CRTs is increasing. Estimates of the ICC are poorly reported in such studies. Better use could be made of feasibility CRTs to assess challenges that are specific to studies that allocate school-based clusters.

The median (interquartile range; range) ICC for pupil health outcomes worldwide was 0.031 (0.011 to 0.08; 0 to 0.47) at the school level and 0.063 (0.024 to 0.1; -0.009 to 0.262) at the class level. There were no clear associations between study characteristics and the ICC, other than estimates being larger in definitive trials than feasibility CRTs.

School-level and class-level ICCs for pupil social emotional functioning outcomes reported by the same teacher for all pupils in the same class were larger than ICCs for the parent- and pupil-reported versions of the same outcomes. School-level ICCs were larger in the study that sampled only one class from each school compared to the other studies that included pupils from multiple classes in each school.

When specifying an ICC for the sample size calculation for school-based CRTs, the potential impact of the different levels of clustering in the data and the outcome reporter need to be considered.

Page intentionally left blank

Contents

Dedication.....	2
Acknowledgements.....	3
Abstract.....	5
List of Tables	11
Author's Declaration	15
Abbreviations.....	17
Chapter 1: Introduction	20
1.1 Cluster Randomised Trials: Overview	20
1.2 Design features and methodological considerations of cluster randomised trials.....	22
1.3 Trials evaluating interventions for improving health outcomes on children and adolescents	32
1.4 Cluster randomised trials in school settings.....	34
1.5 Justification and aim of the thesis.....	39
1.6 Research objectives	40
1.7 Overview of thesis	40
Chapter 2: Thesis overview	43
2.1 Chapter 3 overview.....	43
2.2 Chapter 4 overview.....	45
2.3 Chapter 5 overview.....	46
2.4 Chapter 6 overview.....	47
2.8 Chapter summary	49
Chapter 3: The characteristics and practices of school-based cluster randomised controlled trials for improving health outcomes on pupils in the United Kingdom: A systematic review of definitive trials.....	51
3.1 Summary	51
3.2 Background	51
3.3 Aims and Objectives.....	52
3.4 Methods.....	52
3.5 Results.....	70
3.6 Discussion	99
3.7 Strengths and limitations	102
3.8 Implications.....	104
3.9 Conclusions.....	105

3.10 Chapter summary	105
Chapter 4: Systematic review of the characteristics of school-based feasibility cluster randomised trials of interventions for improving the health outcomes of pupils in the United Kingdom.....	108
4.1 Summary	108
4.2 Background	108
4.3 Aims and objectives.....	110
4.4 Methods.....	111
4.5 Results.....	120
4.5.10 Estimated intra-cluster correlation coefficients	135
4.6 Discussion	137
4.7 Strengths and limitations	140
4.8 Implications.....	141
4.9 Conclusions	142
4.10 Chapter summary	142
Chapter 5: Intra-cluster correlation coefficients from school-based cluster randomised trials of interventions for improving health outcomes on pupils	145
5.1 Summary	145
5.2 Background	145
5.3 Aims and objectives.....	148
5.4 Methods.....	148
5.5 Results.....	159
5.6 Discussion	179
5.7 Strengths and limitations	181
5.8 Implications.....	182
5.9 Conclusions	182
5.10 Chapter summary	183
Chapter 6: Estimating intra-cluster correlation coefficients and components of variance for social emotional functioning outcomes of pupils in school-based cluster randomised trials	185
6.1 Summary	185
6.2 Background	185
6.3 Aims and objectives.....	188
6.4 Methods.....	189
6.5 Results.....	206

6.6 Discussion	232
6.7 Strengths and limitations	236
6.8 Implications.....	237
6.9 Conclusions.....	238
6.10 Chapter summary	238
Chapter 7: Discussion.....	240
7.1 Chapter summary and contribution to knowledge.....	240
7.2 Strengths and limitations	249
7.3 Implications.....	252
7.4 Future research	256
7.5 Closing remarks.....	257
Appendices	258
References	364

Page intentionally left blank

List of Tables

Table 3.1. MEDLINE (via Ovid) search strategy

Table 3.2. Reasons for exclusion at full text screening

Table 3.3. Data extracted from included studies

Table 3.4. Setting and participant characteristics of included studies (N=64)

Table 3.5. Intervention type characteristics of included studies (N=64)

Table 3.6. Primary outcome characteristics of included studies (N=64)

Table 3.7. Study design characteristics of included studies (N=64)

Table 3.8. Cluster-level characteristics used to balance randomisation (N=51)

Table 3.9. Sample size calculation characteristics of included studies (N=64)

Table 3.10. Ethics and consent characteristics of included studies (N=64)

Table 3.11. Analysis methods characteristics of included studies (N=64)

Table 3.12. Other areas of methodological interest from included studies (N=64)

Table 3.13. Estimated intra-cluster correlation coefficients (ICCs) for primary outcomes (N=29)

Table 4.1. Reasons for exclusion at full text screening

Table 4.2. Data extracted from school-based feasibility CRTs

Table 4.3. Summary of methodological characteristics of included studies (N=24)

Table 4.4. Reported intra-cluster correlation coefficients (ICCs) for primary outcomes (N=8)

Table 5.1. Reasons for exclusion at full text screening

Table 5.2. Data extracted from included articles

Table 5.3. Criteria used to select which intra-cluster correlation coefficient (ICC) or between-cluster coefficient of variation of the outcome (CV) to extract

Table 5.4. Summary of study features and design characteristics (N=246).

Table 5.5. Median (IQR; range) school-level intra-cluster correlation coefficient (ICC) by world region, outcome area and education stage (N=210)

Table 5.6. Median (IQR; range) school-level intra-cluster correlation coefficient (ICC) by region, health outcome area and education stage summarised separately for continuous and binary outcomes

Table 6.1. Characteristics of the school-based cluster randomised trials at randomisation

Table 6.2. Description of outcomes, outcome measures and outcome scoring

Table 6.3. Demographic characteristics of participants (N indicates sample size)

Table 6.4. STARS study intra-cluster correlation coefficients (ICCs) for the Strengths and Difficulties Questionnaire (SDQ) outcomes at different time points

Table 6.5. STARS study intra-cluster correlation coefficients (ICCs) for the teacher-reported Pupil Behaviour Questionnaire and the pupil-reported 'How I Feel About My School' measure and at different time points

Table 6.6. KiVa study intra-cluster correlation coefficients (ICCs) for the teacher-reported Strengths and Difficulties Questionnaire (SDQ) outcomes at different time points

Table 6.7. KiVa study intra-cluster correlation coefficients (ICCs) for pupil-reported Olweus Bully/Victim Questionnaire (OBVQ) and bullying outcomes (KiVa questionnaire) at different time points

Table 6.8. PACES study intra-cluster correlation coefficients (ICCs) for the Revised Child Anxiety and Depression Scale (RCADS-30) at different time points

Table 6.9. PACES study intra-cluster correlation coefficients (ICCs) for the parent-reported Strengths and Difficulties Questionnaire (SDQ) outcomes at different time points

Table 6.10. PACES study intra-cluster correlation coefficients (ICCs) for pupil-reported outcomes at different time points

Table 6.11. PROMISE study intra-cluster correlation coefficients (ICCs) for the pupil-reported Revised Child Anxiety and Depression Scale (RCADS-30) at different time points

Table 6.12. PROMISE study intra-cluster correlation coefficients (ICCs) for pupil-reported outcomes at different time points

Table 6.13. MYRIAD study intra-cluster correlation coefficients (ICCs) for the pupil and teacher-reported Strengths and Difficulties Questionnaire outcomes at different time points

Table 6.14. MYRIAD study intra-cluster correlation coefficients (ICCs) for the pupil and teacher-reported Behaviour Rating Inventory of Executive Function, Second Edition (BRIEF-2) outcomes at different time points

Table 6.15. MYRIAD study intra-cluster correlation coefficients (ICCs) for the pupil-reported Revised Child Anxiety and Depression Scale (RCADS-30) outcomes at different time points

Table 6.16. MYRIAD study intra-cluster correlation coefficients (ICCs) for the pupil-reported Centre for Epidemiologic Studies for Depression Scale, Warwick-Edinburgh Mental Well-being Scale, Child and Adolescent Mindfulness Measure, suicide ideation and self-harm outcomes at different time points

Table 6.17. MYRIAD study intra-cluster correlation coefficients (ICCs) for the pupil-reported School Climate and Connectedness Survey (SCCS) outcomes at different time points

Table 7.1. Key findings from this thesis

Table 7.2. Key recommendations for future research based on the findings from this thesis

List of Figures

Figure 1.1. A diagrammatic representation of an individually randomised controlled trial and a cluster randomised controlled trial

Figure 1.2. A diagrammatic representation of the structure of this thesis

Figure 3.1. PRISMA flow diagram summarising the results of the literature search and screening for eligibility

Figure 3.2. Published cluster randomised trials indexed in MEDLINE from inception to 30th June 2020 (N=64)

Figure 3.3. Estimated intra-cluster correlation coefficient (ICC) for primary outcomes versus ICC assumed in sample size calculation (N=15)

Figure 4.1. PRISMA flow diagram summarising the results of the literature search and screening for eligibility

Figure 4.2. Published feasibility cluster randomised trials indexed on MEDLINE from inception to 31st December 2020 (N=24)

Figure 5.1. PRISMA flow diagram summarising the results of the literature search and screening for eligibility

Figure 5.2. The distribution of school-level intra-cluster correlation coefficients (ICCs) in school-based CRTs (N=210)

Figure 5.3. The frequency of class-level intra-cluster correlation coefficients (ICCs) in school-based CRTs (N=46)

Figure 5.4. Dot plots of school-level intra-cluster correlation coefficients (ICCs) by region, outcome area and education stage

Author's Declaration

The thesis comprises chapters presenting studies in their original, unpublished forms. Chapters 3, 4 and 5 have also been written as papers published in peer reviewed academic journals and are presented in Appendices 1-4. The contribution that I personally made to each of these papers, and the contribution made by each co-author, is described below.

Characteristics and practices of school-based cluster randomised controlled trials for improving health outcomes in pupils in the United Kingdom: a methodological systematic review – presented in Chapter 3

KP, MN, ZMX, TF and OU conceived the study. ZMX and TF advised on the design of the study and contributed to the protocol. KP, MN and OU contributed to the design of the study, wrote the protocol and designed the data extraction form. KP and OU undertook screening and data extraction. KP conducted the analyses of the data. All authors had full access to all the data. KP took primary responsibility for writing the manuscript. All authors provided feedback on all versions of the review. All authors read and approved the final manuscript.

Systematic review of the characteristics of school-based feasibility cluster randomised trials of interventions for improving the health of pupils in the UK – presented in Chapter 4

KP, SEd, MN, ZMX, TF, SE and OU conceived the study. MN, ZMX, TF and SE advised on the design of the study and contributed to the protocol. KP, SEd and OU contributed to the design of the study, wrote the protocol and designed the data extraction form. KP, SEd and OU undertook screening and data extraction. KP conducted the analyses of the data. All authors had full access to all the data. KP took primary responsibility for writing the manuscript. All authors provided feedback on all versions of the paper. All authors read and approved the final manuscript.

Intra-cluster correlation coefficients from school-based cluster randomised trials of interventions for improving health outcomes in pupils – presented in Chapter 5

KP, MN, ZMX, TF and OU conceived the study. ZMX and TF advised on the design of the study and contributed to the protocol. KP, MN and OU contributed to the design of the study,

wrote the protocol and designed the data extraction form. KP and OU undertook screening and data extraction. KP conducted the analyses of the data. All authors had full access to all the data. KP took primary responsibility for writing the manuscript. All authors provided feedback on all versions of the paper. All authors read and approved the final manuscript.

Abbreviations

ADHD	Attention Deficit Hyperactivity Disorder
BMI	Body Mass Index
BRIEF	Behaviour Rating Inventory of Executive Function
CAMM	Child and Adolescent Mindfulness Measure
CATS	Children's Automatic Thoughts Scale
CES-D	Centre for Epidemiologic Studies Depression Scale
CI	Confidence Interval
CRT	Cluster Randomised Trial
CV	Between-cluster Coefficient of Variation of the outcome
DARE	Database of Abstracts of Reviews of Effects
DE	Design Effect
EMBASE	Excerpta Medica Database
ERIC	Education Resources Information Centre
GAD	Generalised Anxiety Disorder
GEEs	Generalised Estimating Equations
HIFAMS	'How I Feel About My School' measure
ICC	Intra-cluster Correlation Coefficient
IDACI	Income Deprivation Affecting Children Index
IMD	Index of Multiple Deprivation
IQR	Interquartile range
ISCED	The International Standard Classification of Education
ISRCTN	International Standard Randomised Controlled Trial Number
MEDLINE	Medical Literature Analysis and Retrieval System Online
MeSH	Medical Subject Headings
MVPA	Moderate-to-Vigorous Physical Activity
NHS	National Health Service
NIHR	National Institution for Health and Care Research
OBVQ	Olweus Bully/Victim Questionnaire
OCD	Obsessive Compulsive Disorder
PBQ	Pupil Behavioural Questionnaire
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PSHE	Personal, Social, Health and Economic education
PsycINFO	Psychological Information Database
RCADS	Revised Child Anxiety and Depression Scale
RCT	Randomised Controlled Trial
SAD	Separation Anxiety Disorder
SCCS	School Climate and Collectedness Survey
SD	Standard Deviation
SDQ	Strengths and Difficulties Questionnaire
SE	Standard Error
SES	Socio-Economic Status
SMFQ	Short Moods and Feelings Questionnaire
UK	United Kingdom
USA	United States of America

UNESCO	The United Nations Educational, Scientific and Cultural Organisation
WEMWBS	Warwick-Edinburgh Mental Well-being Scales

Page intentionally left blank

Chapter 1: Introduction

This chapter presents an overview of cluster randomised controlled trials (CRTs), describes the contexts in which they may be used, and outlines their characteristic features and methodological challenges. The chapter then summarises the current methodological literature describing the use of CRTs to evaluate interventions for improving health outcomes on children and adolescents. The chapter describes the methodological considerations of using CRTs in school settings. The chapter then concludes by outlining the objectives and scope of the thesis.

1.1 Cluster Randomised Trials: Overview

Randomised controlled trials (RCTs) are considered the gold standard design for evaluating new interventions or treatments [1](p1-7). In the traditional RCT design, participating individuals are randomly allocated to either an intervention (experimental) arm or a control arm (either an alternative intervention or no intervention). The effect of the intervention is then quantified by comparing outcomes on the participants between the trial arms. If the number of individuals is sufficiently large, researchers can be confident that differences in the outcomes observed between trial arms are a result of the intervention, rather than a result of differences on other known or unknown factors. Random allocation prevents selection bias by ensuring that participants with different characteristics have the same chance of being allocated to the intervention arm [1](p1-7). It enhances the internal validity of the comparison, that is the extent to which the observed results are estimating the true intervention effect in the study population.

Cluster randomised controlled trials (CRTs), also known as group randomised trials, place randomised trials or community randomised trials, are studies in which groups (clusters) of participating individuals are allocated to trial arms rather than individuals themselves [2-6]. CRTs differ from traditional RCTs in that rather than randomising individual participants, entire clusters are the units of randomisation with outcomes measured on participants within the clusters. Clusters may be health organisations (e.g., hospitals, general practices), non-health organisations (e.g., workplaces, schools) or geographical areas (e.g., towns, villages). A diagrammatic representation of the difference between an individually RCT and a CRT is shown in Figure 1.1.

Individually Randomised Controlled Trial

Cluster Randomised Controlled Trial

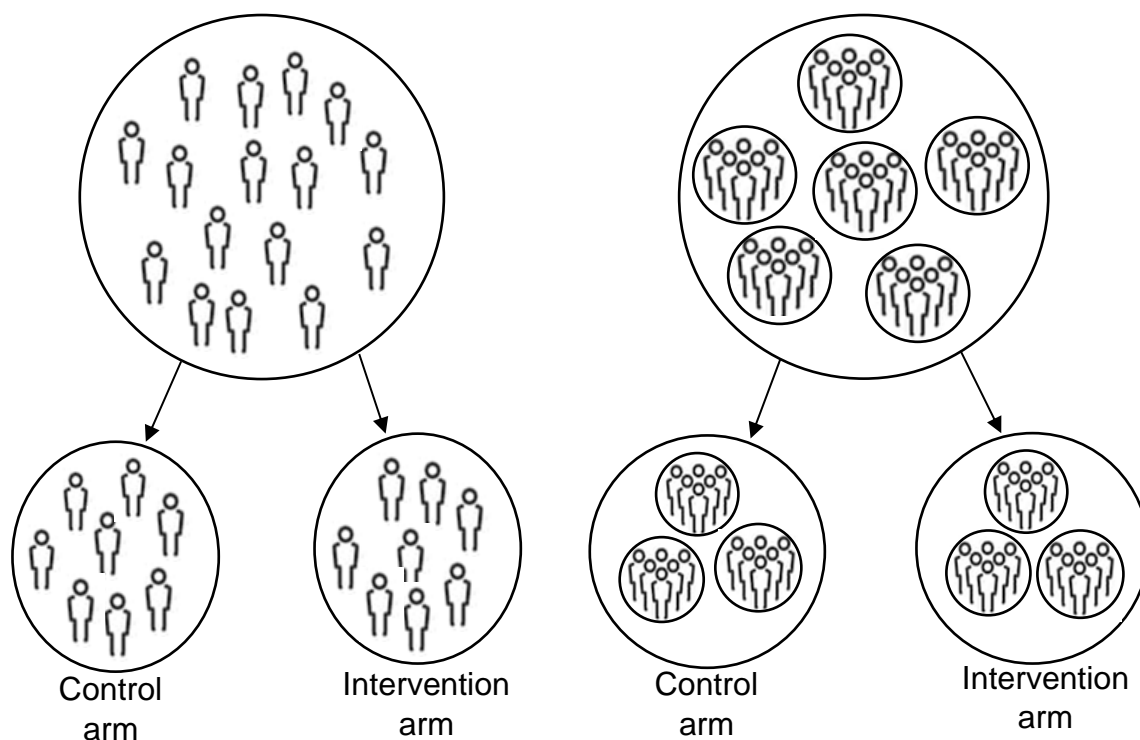


Fig. 1.1. A diagrammatic representation of an individually randomised controlled trial and a cluster randomised controlled trial.

Randomisation of clusters rather than individuals may be preferred for a number of reasons [2](p13). Clusters may be randomised because the intervention has been designed to be delivered at the cluster level and not the level of the individual [3](p10). An example of this would be the implementation of water fluoridation of towns. It would not be feasible to provide fluoridated water to specific individuals, so the entire towns (clusters) would be randomised when evaluating the intervention. Cluster randomisation is also more pragmatic here as it reflects real life delivery. Another reason for randomising clusters is to reduce the risk of 'contamination' that may otherwise occur between trial arms if individuals were randomised [3](p11-12). In other words, the CRT design minimises the possibility of individuals from different trial arms interacting and diluting the effect of the intervention. For example, in a trial evaluating a change in diet, individuals in the control arm might learn about the experimental diet and implement it themselves [7]. Contamination between trial arms can also occur at the level of the person delivering the intervention as they may find it hard to not deliver it to participants in the control arm if individuals are randomised [3](p11-12). By

randomising clusters, contamination is avoided as all members of a given cluster are allocated to the same trial arm [3](p11-12), provided participants in a intervention cluster do not interact with participants in a control cluster. The CRT design has also been used for logistical reasons, cost and administrative convenience [3](p12).

Despite there being a number of reasons why researchers may prefer to randomise clusters rather than individuals, cluster randomisation should only be used where there is a strong methodological justification [3](p10). Outcomes of participants in the same cluster tend to be more similar to each other than with outcomes of participants from different clusters. As a result of this within-cluster similarity, a larger number of participants are required in CRTs than if individual randomisation were used [8]. Because CRTs typically randomise few units (clusters), they are more susceptible than individually RCTs to imbalance in baseline characteristics between the trial arms [8]. Researchers should provide a clear rationale for choosing cluster randomisation over individual randomisation in order to justify the larger sample size inherent in the CRT design [4].

1.2 Design features and methodological considerations of cluster randomised trials

1.2.1 Intra-cluster correlation

A characteristic feature of CRTs is that outcome observations on participants who belong to the same cluster are usually more similar to each other than observations on participants from different clusters [4](p6-7). For example, patients registered with the same general practice (cluster) are more likely to have similar health outcomes to each other than those registered with different practices [9]. This similarity, or correlation, between participants in the same cluster can occur for three main reasons. First, people may choose the cluster they belong to, for example, individuals may choose the town they live in. Second, the cluster may exert a common influence on all members of the same cluster, for example, school policy may impact on all pupils in a similar way. Finally, participants may interact within their cluster and this may lead to more similar outcomes, for example, individuals interact within workplaces on a daily basis and this may lead to similarity with some outcomes.

The intra-cluster correlation coefficient (ICC), denoted ρ , quantifies the similarity of observations for a specific outcome on individuals within the same cluster. A common definition used for continuous outcomes is that the ICC is the proportion of the total variation in the outcome that is between clusters (σ_b^2) as opposed to between individuals within clusters (σ_w^2) [10, 11]:

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

Under this definition, the ICC can take values between 0 and 1. The larger the ICC is, the greater the similarity between individuals within clusters, or, equivalently, the greater the proportion of variability in the outcome that is between clusters [10]. If the outcomes of individuals from the same cluster are no more similar to each than the outcomes of individuals from different clusters then the ICC is zero.

The between-cluster component of variance (σ_b^2) [12] quantifies the variation in the outcome between clusters; it is the square of the standard deviation (SD) of mean outcomes across clusters. A variance quantifies how far a set of numbers are spread from the mean value, and in the present context σ_b^2 measures how far the mean outcomes from different clusters vary around the overall mean [12]. The within-cluster component of variance (σ_w^2) quantifies the variation in the outcome across participants within the same cluster; that is, how far the participants' outcomes in a given cluster vary around the mean outcome in the cluster.

The proportion of variance definition of the ICC is expressed differently for binary outcomes [11], for which the overall variance of the outcome, $\pi(1 - \pi)$, depends on the outcome prevalence, π [13]. The definition for the ICC, $\rho_{b(\text{linear})}$, for a binary outcome is:

$$\rho_{b(\text{linear})} = \frac{\text{var}(\pi_i)}{\pi(1 - \pi)}$$

where π_i is the proportion with the binary trait in the i th cluster and $\text{var}(\pi_i)$ is the variance of the cluster proportions (between-cluster variation). Under this definition the total outcome variance is expressed on the linear (proportions) scale.

There is different definition of the ICC for binary outcomes where the between-cluster variation is expressed on the logit, or log odds, transformation of π_i :

$$\text{logit}(\pi_i) = \ln(\pi_i/(1 - \pi_i))$$

This definition of the ICC assumes that the binary outcome is the dichotomised version of an underlying latent continuous variable that represents the tendency of an individual level cluster member to have the binary trait [10]. Individuals for whom the value of this latent variable is over a certain threshold, have the binary trait (coded 1) while the remaining individuals do not have the trait (coded 0). The underlying continuous variable is assumed to follow a logistic distribution. The definition for the ICC, $\rho_{b(\text{logit})}$, for a binary outcome is then:

$$\rho_{b(\text{logit})} = \frac{\text{var}(\text{logit}(\pi_i))}{\text{var}(\text{logit}(\pi_i)) + (\pi^2/3)}$$

where $\text{var}(\text{logit}(\pi_i))$ is the between-cluster variance on the logit scale, π is the mathematical constant (~ 3.141592654), and $\pi^2/3$ is the within-cluster variance on the logit scale [10].

Another way of quantifying the correlation of outcomes of participants from the same cluster is using the between-cluster coefficient of variation of the outcome (CV) [14]:

$$CV = \frac{\sigma_b}{\mu}$$

where σ_b is the between-cluster standard deviation (the square root of the between-cluster variance component), and μ is the mean outcome across the clusters. The higher the CV , the greater the level of variation of the outcome across clusters, and the greater the correlation of the outcome within clusters [14].

The similarity, or lack of statistical independence, between observations on individuals from the same cluster means that the usual methods for calculating sample size and analysing data in individually RCTs should not be used in CRTs [4]. The use of standard sample size methods to calculate the number of participating individuals needed in CRTs will result in studies that contain too few individuals and lack power to detect the pre-specified intervention effect of interest (i.e., the smallest effect that is worth detecting) [4]. The use of standard analysis methods to estimate the intervention effect from the resulting trial data will produce confidence intervals (CIs) that are narrower and p-values that are smaller than they should be, thus exaggerating evidence for the benefit of the intervention [4]. Therefore, sample size and analysis methods that take account of clustering should be used in CRTs.

Information about clustering (i.e., the ICC or CV of the outcome) for study outcomes is invaluable when calculating the sample size for a planned CRT and can be obtained from previous studies with the same type of cluster and outcomes to the planned trial [11]. Such studies may be previous CRTs, multi-centre individually randomised trials where the "centre" is the cluster of interest, multistage (cluster) surveys, or routine datasets [11]. In order to aid the design of future similar studies, authors reporting results from CRTs should provide estimates of the ICCs (or CVs) from their study. These should be reported with CIs because CRTs typically have few clusters, resulting in imprecise ICC estimates [10, 11].

1.2.2 Sample size calculation

When calculating the sample size required for a CRT, both the total number of clusters that need to be recruited and the number of individuals that need to be recruited from within each cluster must be determined. In order to detect a specified intervention effect, CRTs require more participants than traditional RCTs where individual participants are randomised [3](p137). Correlation of outcome observations within clusters means that each participant in a CRT provides less information than each participant in a trial that randomises individuals.

A consideration that researchers have to make when determining the sample size for a planned CRT is the trade-off between having large numbers of clusters and having large numbers of individuals within clusters. The total number of clusters is the key driver for increasing the power to detect the effect of the intervention in a CRT especially when the assumed ICC is large [15]. The total required number of participants decreases as the number of clusters increases, but it is often impractical and more expensive to recruit many clusters. Many studies have an upper limit on the number of clusters that is feasible to recruit. This may result in the study not being feasible as the maximum achievable power is heavily limited by the number of clusters, regardless of how many individual participants are included within each cluster [16].

1.2.2.1 Using the ICC to estimate the sample size

When the number of individuals in each cluster is fixed and known in advance, the total number of individuals required in a CRT is calculated by inflating the number of individuals that would be required in an individually RCT by the *design effect (DE)* [17]:

$$DE = 1 + (\bar{m} - 1)\rho$$

where \bar{m} is the mean number of participants providing outcome data in each cluster, often referred to as the *cluster size*, and ρ is the ICC of the outcome [3](p142). When calculating the sample size for binary outcomes must be specified on the linear (proportions) scale of the outcome, using $\rho_{b(\text{linear})}$. The *DE* is often referred to as the *variance inflation factor* in the literature because it is the amount by which the variance of the intervention effect estimate is increased as a result of using the CRT design [18]. Having calculated the total number of participants required, this is divided by the cluster size to obtain the total number of clusters that is required in the CRT. If the cluster size is large, then even a small ICC may drastically increase the total number of participants required [19]. For example, with a ICC of 0.05 (considered to be a conservatively high estimate for patient outcomes in general practice clusters [20]) and a cluster size of 100, the total number of individuals required for a parallel arm CRT is six times that under individual randomisation.

For scenarios where the total number of clusters available for a CRT is fixed and known, an alternative calculation based on the same DE approach is used to calculate the number of participants that need to be recruited from each cluster [16].

In addition to incorporating the DE, when calculating the sample size in CRTs, a *degrees of freedom correction* should be incorporated to take account of the anticipated uncertainty with which variability in the outcome across clusters is estimated in the analysis of the intervention effect. A further inflation of the sample size should be considered to allow for loss of efficiency that results from recruiting unequal numbers of participants from the clusters [12].

1.2.2.2 Using the between-cluster coefficient of variation of the outcome (CV) to estimate the sample size

The between-cluster CV can be used as an alternative measure of outcome clustering when adjusting the sample size calculation in CRTs with binary and count/rate (i.e., incidence rates of a disease) outcome data [3](p145-146); it is incorporated in to a modified design effect formula. Like the ICC, the larger the CV is the greater the between-cluster variation (or equivalently the within-cluster correlation) and the greater the inflation that is required for the sample size calculation for the CRT [14].

1.2.3 Analysis methods

1.2.3.1 Estimating intervention effect

When estimating the intervention effect in CRTs, analytical methods should take account of clustering, otherwise confidence intervals for the intervention effect will be narrower and corresponding p-values will be smaller than they should be. Furthermore, the degrees of freedom used for calculating the confidence interval and p-value for the intervention effect should take account of the number of clusters in the study [21]. The use of standard analytical methods that incorrectly assume there is no within-cluster correlation results in an exaggeration of the amount of evidence for a true intervention effect and the precision with which the effect is estimated [4].

In CRTs, statistical analyses of the intervention effect may be undertaken at the cluster-level or at the individual-level. For cluster-level analyses, the outcome is summarised for each cluster, for example, by calculating the means for continuous outcomes or percentages for binary outcomes across individuals in the cluster. Standard analytical methods are then used to compare the outcome between the trial arms using the cluster-level summary statistics as the observations [22]. Important covariates that need to be adjusted for in the analysis can be incorporated through the use of regression modelling at the cluster level [3](p107-109). This method of analysis is valid because the cluster is both the unit of randomisation and the unit of analysis.

Individual-level analysis of data from CRTs involves the application of regression-based methods that allow for the within-cluster correlation. This approach is exemplified by methods such as mixed effects (“multilevel”) models [23] and marginal models estimated using Generalised Estimating Equations (GEEs), usually assuming an exchangeable correlation structure with information sandwich (“robust”) estimates of standard error [24]. These methods readily facilitate adjustment for individual- and cluster-level factors that are potentially predictive of the outcome.

1.2.3.2 Estimating ICCs from outcome data

When publishing the results of CRTs, the ICCs for outcome variables should be reported to help inform the design of future similar trials. Different methods can be used to estimate

ICCs [10, 25], such as random effects analysis of variance and the regression-based methods discussed above [10]. Regression-based methods are commonly used to analyse data from CRTs and calculate ICCs since the ICC is estimated as part of the model fitting process [26] or to weight the analysis [24].

As described earlier, there are different definitions of the ICC for binary data depending on whether the binary outcome is analysed on the linear (proportions) scale or the logistic scale [10, 27]. The ICC for binary variables can be estimated using methods that allow for clustering. Random effects analysis of variance, marginal regression models using GEEs, and random effects (“multilevel”) linear regression can be used to estimate the ICC for a binary outcome on the linear scale, $\rho_{b(\text{linear})}$ [27]. Random effects (“multilevel”) logistic regression can be used to estimate the ICC for a binary outcome on the logistic scale, $\rho_{b(\text{logit})}$ [27].

1.2.4 Recruitment and consent processes

The clusters and individuals recruited to take part in CRTs should, ideally, be representative of the wider study population. In order to improve generalisability of the findings, a diverse range of clusters should be recruited, or investigators should ensure that settings are representative of the wider population. Investigators should also limit inclusion/exclusion criteria, encourage high uptake from eligible clusters and participants through means such as incentivisation, and provide them with adequate information about the trial [3](p23-24).

Ideally, participants should be identified and recruited before the clusters are randomised [28]. However, in some CRTs, this is not possible, in which case, the person recruiting the participants should be blind to the trial arm status of the cluster, otherwise the number and characteristics of recruited individuals may then differ between the trial arms, resulting in selection bias and compromising study validity [29].

1.2.5 Consent processes

In CRTs, consent should be obtained at the level of the cluster and the individual participant [30]; this makes the consent process more complicated than for individually RCTs. Consent can be sought for different components of a trial including randomisation, participation in the intervention and data collection [30, 31]. Cluster guardians/gatekeepers are appointed to

make decisions regarding participation in the trial on behalf of the individuals within those clusters (e.g., headteacher of a school, community leader of a village) [32]. In CRTs, it is rarely possible for individual participants to consent to randomisation and the intervention as these decisions are taken at the cluster level [30]. Usually, individuals are asked only for consent for data collection. The challenges with consent in CRTs have been documented extensively in the literature [30-35].

1.2.6 Restricted randomisation

As there are often few clusters included in CRTs due to logistical and financial constraints, simple randomisation may not evenly balance key cluster characteristics that are predictive of the outcomes across the trial arms. Completely randomised designs, where the interventions are assigned at random to clusters without reference to the characteristics of the clusters, are the simplest. In recent years, however, it has become more common to use some form of *restricted* allocation in CRTs [3](p75-76).

Restricted randomisation, or restricted allocation, involves modifying the randomisation process to reduce the chance of poor allocations and ensure trial arms are similar (or balanced) with respect to specific cluster characteristics while retaining the benefits of randomisation [36]. Restricted allocation enhances the face validity of the subsequent comparisons of the outcomes between the trial arms. Adjusting for characteristics used to balance the allocation when analysing the data from the CRT provides greater precision for estimating the intervention effect if those characteristics are predictive (i.e., prognostic) of the trial outcomes [3](p76-78). Restricted allocation can be used in combination with *blocking* (i.e., allocation of equal numbers of clusters to each trial arm within blocks based on order of recruitment) to ensure balance between trial arms with respect to the number of clusters [3](p82-83).

Besides balancing on cluster-level prognostic factors, restricted allocation may be used for other more practical reasons. For example, one might balance the randomisation based on cluster size to ensure that the total number of participants recruited is similar in each trial arm [37]. If the trial is undertaken in different geographical areas then randomisation may be balanced on location to ensure that different areas have an equal chance of being allocated the intervention [37]. Finally, if there is the need to investigate a cluster-level

characteristic as a potential moderator of the effect of the intervention then restricted allocation may be used to ensure similar numbers of intervention and control clusters in each level of the characteristic, for example, type of school (e.g. state funded versus independent) [37]. Balancing on such characteristics increases the power of tests of moderation.

A commonly used form of restricted randomisation is *stratification*. In stratified designs, clusters are grouped into strata based on having the same characteristics on factors that the investigators want to balance on between trial arms, such as geographic location or socio-economic status (SES). Within these strata, clusters are then randomly assigned to the trial arms [2](p45). This ensures balance with regard those cluster-level factors. A special case of the stratified design where each stratum contains only two clusters (i.e., a pair of clusters) is called a *matched pair* design. Under this design, one cluster from each pair is randomly allocated to each trial arm. A disadvantage of the matched pair method is that if one cluster from the pair drops out of the study, the other cannot be included in the analysis. Although there can be gains in statistical efficiency from matching, clusters within each pair need to be sufficiently similar with respect to the outcome or the method will result in a decrease in statistical power if matching status is incorporated in the analysis [3](p84).

When there are few clusters in a CRT and a large number of cluster-level characteristics to balance on, *constrained* randomisation is more practical than stratification and matching. Constrained randomisation involves undertaking a large number of randomisations for the recruited clusters and randomly selecting a randomisation sequence for which there is a reasonable balance, based on pre-specified criteria, in the cluster characteristics [38, 39].

In situations where clusters are recruited and randomised sequentially over time, rather than in a single batch, a method called *minimisation* can be used to allocate clusters whilst maintaining balance on the cluster characteristics across trial arms. Under minimisation, the assignment of each new cluster partially depends on the current balance (or imbalance) in the cluster characteristics across trial arms and, therefore, the method is only a pseudo-random process. Newly recruited clusters are allocated in a manner that is weighted towards maintaining the balance in cluster-level characteristics between trial arms [40].

1.2.7 Follow-up designs in CRTs

There are two main designs that are used in CRTs to undertake follow-up assessments on the participants: the *cohort* design and the *repeated cross-sectional* design. In CRTs, measurements are taken on individuals within clusters. In cohort designs the same participants provide data from each cluster at each measurement occasion. In repeated cross-sectional designs, a different set of participants provide data in each cluster at each measurement occasion. The cohort design is more useful for determining how an intervention affects individual-level outcomes as the same individuals provide outcome across all time points [41]. The repeated cross-sectional design is more useful when the aim is to measure the effect of an intervention at the cluster level [3](p86). In some circumstances it is only possible to use a repeated cross-section design, for example, when evaluating childbirth outcomes as the mothers cannot give birth at each measurement occasion [42].

1.2.8 Feasibility studies

Feasibility studies are often used ahead of the main definitive trials to explore any potential challenges in delivering the trial, establish if the trial is something that can be done, if it should be done and how it should be done [43]. Feasibility studies are smaller scale studies that focus on uncertainties in the main trial. For example, challenges in the randomisation and recruitment processes, the delivery and acceptability of the intervention, and estimating parameters such as recruitment and follow-up rates to inform the design of the future study.

Feasibility studies undertaken prior to a definitive CRT differ from feasibility studies performed prior to an individually RCT as they may be used to address concerns specific to the CRT design. These can include some of the challenges discussed earlier in this chapter, such as the possibility of recruitment bias if clusters are randomised before individual participants are recruited [29], and estimating the ICC of the primary outcome for use in the sample size calculation for the definitive trial, although authors warn of imprecise ICC estimates resulting from the small number of clusters usually included in feasibility studies [44].

1.3 Trials evaluating interventions for improving health outcomes on children and adolescents

In 2020, it was estimated that a third of the global population was under 20 years old [45]. In recent years, there has been an increased focus on improving childhood and adolescent health and intervening early in life in order to prevent adult disease, particularly for public health concerns such as obesity, physical inactivity and mental illness [46-48]. Developmental and physiological processes in children and adolescents differ from those in adults, and some conditions may only present in childhood/adolescence or present differently compared with adults [49], making it essential to investigate such health challenges in childhood. However, using children and adolescents as participants in health research presents unique challenges to the design and conduct of trials.

Children are a heterogeneous group with respect to their physiology, behaviour, physical and mental development [50]. This makes it challenging when planning a research trial as there is generally less information available regarding the rationale for choice of comparators for the control arm, validity of outcome measures, and long-term adverse events compared with trials where adults are the participants [51]. Additionally, it can be harder to recruit children than adults to health research studies as the burden of participating in a trial may be more apparent in children (e.g., needing a parent/carer to help them travel to the research venue) [52]. Furthermore, consent must be obtained from parents/carers in order for their children to participate, which may be more difficult for potentially controversial interventions such as vaccinations against diseases [33].

Another challenge when conducting health research in children and adolescents regards the choice of outcomes measure. Child outcome measures differ from those used for adults in that the former have child-specific elements [53]. Children and young people grow through developmental stages meaning it is not always appropriate to use the same outcome measures when studying children that span different ages, or where there is a need to follow-up children over a long period of time [53]. Some outcomes are more challenging to measure and report in younger children and require age-appropriate tools. The language used by researchers is important to ensure that children understand what is required of them and what is being done [54]. There are also issues around the reliability of children's responses

to self-reported outcome measures, particularly for younger children [55]. One way of overcoming this may be for outcomes to be reported by-proxy by a responsible adult (i.e., parents/carers, teacher) as well as the children in order to assess the outcome from different perspectives [54].

There are additional ethical concerns when enrolling children and young people into trials. It is important to ensure that children are protected as research subjects, and that age appropriate information is provided to inform children of their role in the trial [50]. Assent, an agreement given by a child/young person who is not legally empowered to give consent, may be obtained from the child/young person for participation. Children and young people also require proxy decision-makers (e.g., parents/carers) to provide consent on their behalf, which may result in some children not taking part in the trial when they otherwise would have if they were able to consent themselves [56].

1.3.1 Cluster randomised trials in child and adolescent populations

CRTs may be particularly appropriate for child health research as such trials often focus on non-pharmaceutical interventions, such as for behaviour change, aimed at improving public health [53]. This is because many of these interventions are delivered at the level of the cluster. Despite this, there is still a relative lack of methodological literature examining the use of the CRT design to investigate interventions for improving health outcomes on infants, children and adolescents. A 2011 systematic review of CRTs in children found that the rate of publication of such studies has increased since 2004, with studies most commonly undertaken in health areas such as infectious diseases (21%), diet/physical activity interventions (19%), health-risk behaviours (15%), and undernutrition (13%) [53]. The review found that the greatest proportions of CRTs were undertaken in Europe (29%), Asia (23%), and North America (21%). Of the studies included in the review, 72% randomised schools as the cluster unit. The review also highlighted poor reporting, with only 34% of CRTs adequately reporting on more than half of the CONSORT-CRT [57] criteria. Information was often missing regarding how clustering was accounted for in the sample size calculation (41% of CRTs) and analyses (35%), and the ICC for the primary outcomes was only reported in 37% of trials.

Another systematic review examining CRTs in maternal and infant health also found poor quality of reporting, with 10 of the 35 CRTs included not reporting accounting for clustering in the sample size calculation, and 7 not accounting for clustering in their analysis [58]. The review also comments that the shortcomings in the reporting of the sample size calculations made it difficult to evaluate whether an appropriate sample size was used and suggests better reporting and sharing of ICC values are needed [58].

1.4 Cluster randomised trials in school settings

1.4.1 Within-cluster correlation of pupil outcomes in school settings

The CRT design is often used in the school setting to evaluate the effect of interventions on pupil outcomes [53, 59]. The design respects the natural hierarchical structure in schools (i.e., pupils nested within classes (or class-teachers), nested within year groups, nested within schools). Often the interventions assessed using school-based CRTs are designed to be administered at the school/class (cluster) level (e.g., change to school meal policy). Additionally, the CRT design may be used to avoid contamination between the trial arms that might otherwise occur if pupils are randomised, given that they interact within schools (clusters) [50] (e.g., in a CRT evaluating an intervention for improving nutrition intake, pupils in the control arm might learn about the recommendations and adopt them themselves). Therefore, cluster randomisation is generally more appropriate than individual randomisation in the school setting.

As for CRTs undertaken in other settings, for a number of reasons, the outcomes of pupils (individuals) within the same school (cluster) will be more correlated than outcomes of pupils from different schools. First, pupils and their parents/carers may select the school they attend, and any given school is likely to attract pupils with similar characteristics, who are more likely to share similar behaviours and outcomes [50, 60]. Selection into schools in this manner results in pupils having more similar characteristics or behaviours than expected if selection into schools was random [60].

Second, the school itself can influence the behaviour of pupils through its culture and physical environment, ethos and policies [61]; these are termed “contextual effects”. Some researchers place a prominence on contextual effects to explain the association between

schools and the behaviours of pupils [62]. Pupils can choose to accept the instructional (gaining knowledge) and regulatory orders (appropriate behaviours) from a school. This in turn, may impact on whether pupils accept the school's values or engage with student groups within the school that are more conducive to negative behaviours [63].

Third, the characteristics of pupils in the school can have a common influence on the pupils within schools [50]. Such influences can be termed "compositional effects", where the impact of the collective properties of the members within the cluster can influence an individual's behaviours [63]. For example, some schools will have a higher proportion of pupils from socio-economically advantaged families, who may influence other pupils within the school in particular ways. Compositional effects can have both positive and negative influences on all pupils within the school [64] (i.e., increase positive and/or negative behaviours). Compositional effects are generally a consequence of both the selection into schools and the socialisation (interaction) of pupils within the school environment [65]. The peer contagion effect [66], where the behaviours and feelings of pupils can be transmitted between them, and social mimicry [67], where pupils adopt similar behaviours to increase social acceptance and boost self-esteem, are examples of compositional effects within the school social environment that may explain behavioural similarities amongst pupils.

1.4.2 Methodological challenges of school-based CRTs

School-based CRTs share many of the same methodological challenges as CRTs conducted in other settings, but some challenges are more salient [50]. As explained in Section 1.4.1, pupils who attend the same school are more likely to have similar outcomes than pupils attending different schools. This correlation between pupils within schools must be accounted for when designing and analysing school-based CRTs. Authors have previously reported that ICCs for pupil health outcomes are usually smaller than for educational outcomes in school settings [68-70]. This might be expected given that the main purpose of schools is to provide education [71]. ICCs for health outcomes in health care settings are well established, especially in primary health care where empirical data indicate that ICCs at general practice level are generally less than 0.05 [3, 20, 26, 72, 73]. Less is known about the size of ICCs for pupil health outcomes for school-based clusters.

Recruitment of clusters in CRTs has long been reported as a challenge in the literature [4]. It is important that an adequate number of representative clusters are recruited to achieve internal and external validity [3]. It is widely known that the additional demands placed on schools in terms of time and costs is a barrier to the recruitment of schools [33, 74-76]. This is similar to the recruitment of clusters in other settings such as primary care [77-79]. Researchers need to obtain agreement from school management, and potentially consult with parents/carers and pupils before recruitment [80]. Additionally, recruiting schools and participants to research that involves sensitive topics can lead to further barriers to successful recruitment if the intervention does not fit the school ethos or the schools regard the topic as too sensitive [33, 81]. This can lead to the exclusion of certain types of school and compromise the representativeness of the sample.

There are challenges regarding informed consent in school-based CRTs. Multiple stakeholders (i.e., researchers, parents/carers, pupils, school leaders and headteachers) are involved in the informed consent process, which adds complexity [82]. Consent for the school to be randomised and allocated the intervention are usually provided by the senior leadership team of the school. There may be interventions delivered to entire schools/classes, however, that some parents do not want their children to receive (e.g., sex education) [32, 56, 82].

Cluster-level attrition is an issue in CRTs in general, but can be particularly salient in school-based CRTs due to the demands of trials on schools [83]. Other methodological challenges include validity of data that are self-reported by pupils (particularly younger children), lack of long-term follow-up, high pupil drop-out rates, and how best to handle the analysis of data from pupils that change schools (clusters) during the course of the study [55, 81].

1.4.3 School-based CRTs evaluating educational interventions

Some of the first RCTs were undertaken in the field of education early in the 20th century [84]. Since the 1990s, there has been a greater focus on the use of evidence from RCTs to inform educational decision-making [85]. Since 2010, there has been an increase in the use of CRTs to test the efficacy of educational interventions in school settings [59, 86]. As highlighted previously, the CRT design respects the natural clustering in the education setting, and many educational interventions are delivered at the cluster-level. In the United

Kingdom (UK), as of 2020, it is estimated that over a third of English schools are now involved in RCTs [85]. Despite the increasing number of school-based RCTs, often using a CRT design, very few authors have investigated the unique methodological challenges associated with the CRT design for evaluating interventions for improving educational outcomes on pupils.

The literature in the field focuses heavily on the educational system in the United States (US) [87-90], which may not be applicable to other educational systems. Several studies have summarised ICC estimates for educational outcomes from school-based CRTs for use in sample size calculations [88-92]. Prior research has reported school-level ICC estimates between 0.10 and 0.25 for educational attainment outcomes [92]. Additionally, some authors explored patterns between grade level and the size of the ICC finding there may be a negative correlation between them [88, 91]. Although the size and pattern of ICC estimates have been described for educational outcomes in the school setting, there is still a relative lack of methodological literature on school-based CRTs with educational outcomes.

1.4.4 School-based CRTs evaluating health interventions

As for interventions to improve educational outcomes on pupils, schools provide a convenient environment to evaluate interventions for improving the health outcomes of pupils. Schools are an ideal setting in which to deliver health interventions as a large proportion of the world's child and adolescent population attend them. Worldwide, almost 90% of children aged 6 to 11 years are enrolled in primary education and 66% of adolescents aged 12 to 17 years are enrolled in secondary education [93]. Due to the amount of time children spend in school, schools provide a natural setting in which to recruit children and adolescents for participation in research studies, deliver interventions for improving health, and measure health outcomes [50, 74]. At a policy level, there is increasing awareness of the potential for using the school setting to deliver, non-pharmacological, complex public health interventions, and promote health from an early age [53, 94, 95].

The CRT design is well suited to the school setting when evaluating interventions for improving health outcomes on pupils as it reflects the hierarchical structure found within schools. The 2011 methodological systematic review examining the characteristics and quality of reporting of CRTs measuring health outcomes on children reported that 72% of

included studies randomised schools as clusters [53].

Even with the increasing use of the CRT design for school-based health research, few studies have investigated the methodological challenges specific to school-based CRTs measuring health outcomes on pupils. A recent paper highlighted that a lack of suitable ICC estimates for use in sample size calculations is a key issue in school-based CRTs [50]. Researchers have reported ICCs for health outcomes to be generally smaller than those for educational outcomes in school settings [61, 68, 70, 96]. For example, one study reported that the majority of the ICCs for health outcomes such as tobacco use, alcohol use, illicit drug use and risky sexual behaviour were lower than 0.10, compared to the ICCs for academic achievement, which were between 0.19 and 0.25 in the same samples [70]. A number of studies have provided estimates of ICCs from school-based CRTs or surveys but tend to focus on specific health areas such as substance use [61, 71, 97-104], nutrition [105-107], physical activity [61, 107-109], and mental health and behaviour [61, 69, 96], and summarise studies from the US. Additionally, compared with ICCs for health outcomes from CRTs in health care settings, the ICCs for health outcomes in school settings are less established [26, 61, 72, 110-114]. For example, the University of Aberdeen (<https://www.abdn.ac.uk/hsru/what-we-do/tools/>) has a database of ICC estimates for use in sample size calculations for CRTs in healthcare settings but there are no ICCs for pupil health outcomes in the school setting in this database [73].

Given there is currently limited literature on the methodological challenges of school-based CRTs measuring health outcomes on pupils, they may be less well known to researchers in this setting compared with other settings. Many issues are expected to be common to studies that randomise school units, for example, issues with high pupil drop-out rates, and how best to handle the analysis of data from pupils that change schools (clusters) during the course of the study [55, 81]. However, other issues may be more specific for CRTs with health outcomes. Some pupil health outcomes may be more difficult for teachers to report than educational outcomes, given their primary role is to provide education. Teachers may find it hard to understand how to rate particular health outcomes, or understand what the measure means. There may also be more of a research burden on schools for health interventions than educational ones as schools are more used to implementing education, whereas health interventions may result in more disruptive changes (e.g., to timetabling) [50]. Education policy focuses predominantly on maximising academic attainment and less

so on pupil's broader well-being, personal development, and health [115], resulting in less incentive for schools to promote health. Furthermore, some challenges associated with informed consent may be more specific to health interventions, particularly if the interventions are polarising amongst school leaders and parents/carers, for example, interventions surrounding sensitive topics such as vaccination and sexual health [33, 81].

Although school-based CRTs share many common methodological features with CRTs in other settings, some features may be more relevant in schools. Despite the increase in the number of school-based CRTs evaluating the effect of health interventions on pupils [50, 53], there is still limited knowledge regarding the current methodological practices of such studies and more research is needed to understand the specific challenges.

1.5 Justification and aim of the thesis

The aim of this thesis is to describe the methodological characteristics and challenges common to school-based CRTs for evaluating the effect of interventions on pupil health outcomes and collate and explore patterns in ICCs to aid future sample size calculations. In so doing, the thesis will provide knowledge for researchers planning CRTs of interventions for improving health outcomes of school children. The thesis will provide estimates of the ICC of pupil health outcomes based on school-related clusters, and examine whether design features, such as educational level (i.e., pre-school, primary and secondary school), are predictive of the size of the ICC. This knowledge will inform the sample size calculation and wider design of future school-based CRTs and provide plausible parameter values for simulation-based studies that use synthetic data to evaluate the statistical properties of methods used to design and analyse data from such studies. Simulation involves the random generation of synthetic data to evaluate the properties of statistical methods. Amongst other things, simulation studies can be used to estimate: the bias of estimates provided by statistical methods (i.e., the extent to which the estimates systematically deviate from the truth); the coverage of confidence intervals (the probability that the confidence interval includes the true value of the parameter that is being estimated); the power of statistical tests (the probability that the test will provide a statistically significant result when the intervention is effective at a specified level). Taken together, the findings from the thesis will help inform researchers on the design, conduct and analysis of school-based CRTs with health outcomes measured on pupils.

1.6 Research objectives

- 1) Undertake a methodological systematic review to summarise the characteristics and common challenges of school-based definitive CRTs used to evaluate interventions for improving health outcomes on pupils in the UK.
- 2) Undertake a methodological systematic review to summarise the characteristics and objectives of feasibility studies that are undertaken to aid the planning of definitive school-based CRTs used to evaluate interventions for improving health outcomes on pupils in the UK.
- 3) Collate and summarise estimates of the ICC for pupil health outcomes reported in previously published school-based CRTs worldwide and describe the relationships between the ICC and study characteristics.
- 4) Describe patterns in the size of ICC estimates using secondary analysis of raw data from school-based CRTs used to assess interventions for improving social emotional functioning outcomes on pupils in the UK.

1.7 Overview of thesis

The thesis describes and addresses methodological challenges when undertaking school-based CRTs that measure health outcomes on pupils. Original research was conducted as reported in Chapters 3 to 6. The results of these studies are brought together for discussion in Chapter 7 (Discussion), with a consideration of the implications, strengths and limitations of the body of work as a whole, and potential areas for further research. The original research consists of four chapters written in their unpublished forms. Published versions of the research undertaken in Chapters 3, 4, and 5 are in Appendices 1-4.

The studies that comprise this thesis are as follows (see Figure 1.2):

- a methodological systematic review of definitive school-based CRTs used to evaluate interventions for improving pupil health outcomes in the UK (Chapter 3)
- a methodological systematic review of school-based feasibility CRTs undertaken in advance of planned definitive trials for evaluating interventions for improving pupil health outcomes in the UK (Chapter 4)

- a summary of ICC estimates from school-based CRTs worldwide of interventions for improving health outcomes on pupils (Chapter 5)
- a secondary analysis of raw data from five UK school-based CRTs to estimate the intra-cluster correlation coefficients of social emotional functioning outcomes on pupils (Chapter 6).

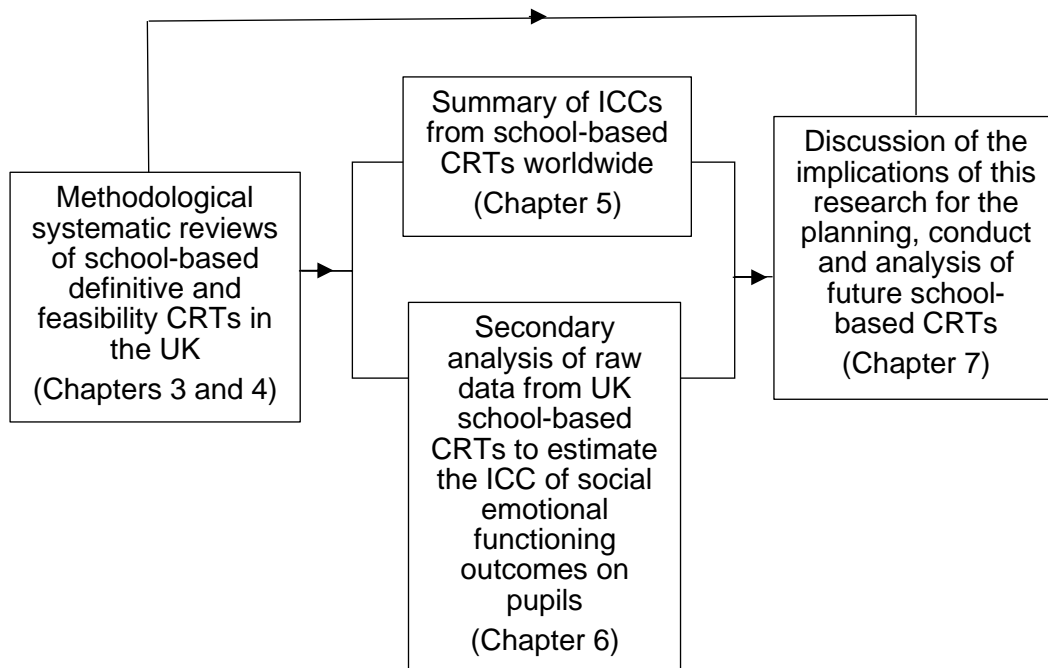


Fig. 1.2. Diagrammatic representation of the structure of this thesis

Page intentionally left blank

Chapter 2: Thesis overview

This chapter provides an overview of the studies presented in the thesis, details the rationale and objectives for each study, and justifies the choice of methodology used to answer the research questions.

2.1 Chapter 3 overview

The first study in this thesis (Chapter 3) is a systematic review of articles reporting the findings of definitive school-based CRTs that evaluated interventions for improving pupil health outcomes in the UK. Two peer reviewed journal articles have been published based on the study: a protocol paper [116] (Appendix 1), and the findings of the systematic review [117] (Appendix 2). The systematic review is reported in detail in Chapter 3. The roles of the researchers involved in the study are specified in the 'Author's Declaration' section of the thesis.

2.1.1 Aims and rationale

The aim of the systematic review was to describe the characteristics and practices of definitive school-based CRTs, including the following aspects: participant and setting characteristics, study design, sample size assumptions, intervention and outcome details, analysis methods, ethics and consent procedures, number of clusters and pupils recruited and followed-up, and the intra-cluster correlation coefficient (ICC) of the primary outcome.

To date, no systematic reviews have been published describing the characteristics and practices of school-based CRTs of interventions for improving health outcomes of pupils. This was, therefore, a logical and essential first step in order to establish common methodological practices in school-based CRTs and identify gaps that could be addressed in the thesis. Through summarising these methodological characteristics and practices, the systematic review also provides knowledge for researchers to help them better plan and conduct their future studies. Finally, the review provides parameter values to inform the design of simulation studies that use synthetic data to assess the statistical properties of methods used to analyse data from school-based CRTs.

2.1.2 Methodology

A systematic approach was used for this literature review as it is considered to be the best approach for finding and synthesising evidence from studies in relation to a specific research question [118, 119]. The purpose of a systematic review is to carry out a precise summary of available primary research evidence relating to a specific research question in order to provide informative and evidence-based answers. This type of review uses systematic searching to find relevant papers and involves the development of a detailed pre-specified plan and search strategy. Systematic reviews are reproducible, and the approach helps to minimise selection bias that would arise if authors were to identify articles themselves in an *ad hoc* manner. Systematic reviews are characterised by a systematic presentation and synthesis of the characteristics and findings of the included studies [120].

The features of a systematic review include:

- Defining a clear research question that the systematic review aims to address
- Outlining the aims, providing pre-defined eligibility criteria for studies to be included
- Clear and reproducible methods
- A rigorous search strategy to find eligible studies
- Critical appraisal of included studies
- A systematic presentation and synthesis of included studies

A systematic review starts by searching sources of evidence (e.g., databases and citation indexes) for relevant studies, using a pre-defined search strategy. Then, using pre-defined eligibility criteria, the titles and abstracts of studies are screened for eligibility. Potentially eligible studies undergo another round of screening using the same criteria but this time the full text of the article is used to assess eligibility. Each study is then assessed in terms of methodological quality using a critical appraisal tool. Lastly, the evidence from each study is extracted and synthesised. This process may or may not include a meta-analysis, a statistical method used to pool the results across the studies.

The systematic review undertaken in Chapter 3 did not use a critical appraisal tool as the aim was to conduct a review of the methodology and characteristics of the studies, rather than collate estimates of the intervention effects. Meta-analysis of the ICC was not undertaken as the studies were methodologically and clinically diverse (i.e., different

outcomes and health conditions). Summarising the variability in ICC estimates is more useful as this provides a range of plausible values within which to assess the sensitivity of the sample size calculation for a CRT [121].

2.2 Chapter 4 overview

The second study in the thesis (Chapter 4) is a systematic review describing the characteristics, methodological practices and objectives of school-based feasibility CRTs measuring health outcomes on pupils. A peer-reviewed journal article was published of this systematic review [122] (Appendix 3). The entire, unpublished version of the systematic review is presented in Chapter 4. The roles of the researchers involved in the study are specified in the 'Author's Declaration' section of the thesis.

2.2.1 Aims and rationale

The aim of this systematic review was to summarise the design features and report the feasibility-related objectives of school-based feasibility CRTs measuring health outcomes on pupils in the UK. Particularly, the review aimed to summarise design features and objectives that were related to using a clustered design, including: the percentage of clusters that are followed up; willingness for clusters to be randomised; estimation of the ICC to calculate the sample size for the definitive study; and the planned and achieved sample sizes at the cluster and individual levels.

No systematic review has summarised the characteristics of school-based feasibility CRTs for improving pupil health outcomes. Therefore, undertaking a systematic review enabled the identification of common practices and gaps in the existing methodological literature. The review helps to highlight areas where improvements could be made to the design and conduct of feasibility CRTs. Furthermore, reporting the feasibility objectives from school-based feasibility CRTs helps to identify areas in which better use of such studies could be made to address uncertainties that are specific to the CRT design.

2.2.2 Methodology

The systematic review was the best methodology to use to address the objectives in this study and used a repeatable approach for finding relevant papers. The key methodological

features of a systematic review and justifications for this choice of methodology have already been provided in Section 2.1.2.

2.3 Chapter 5 overview

The third study in the thesis (Chapter 5) summarises estimates of the ICC of pupil health outcomes from published school-based CRTs undertaken worldwide. A peer-reviewed journal article reporting the study findings has been published [123] (Appendix 4). The study is reported in detail in Chapter 5. The roles of the researchers involved in the study are specified in the 'Author's Declaration' section of the thesis.

2.3.1 Aims and rationale

The aims of this study were to collate and summarise estimates of the ICC of pupil health outcomes reported in school-based CRTs worldwide and examine the relationship between methodological characteristics of the CRTs and the ICC.

A summary of ICC estimates for pupil health outcomes from school-based CRTs in different settings will help researchers design future CRTs by providing plausible values that can be used in sample size calculations. Estimates from CRTs, rather than from surveys, may be more relevant as this information is more generalisable and reflective of the population of schools that participate in health-based trials [3](p175). Identifying relationships between the ICC and design features of CRTs, such as health outcome area, educational level, and region, will help to inform the specification of assumed ICC in sample size calculations in situations where no relevant previous estimates have been reported for the specific outcome in the planned study.

2.3.2 Methodology

A systematic searching strategy was used to identify relevant published school-based CRTs reporting health outcomes on pupils. The method was used because the search strategy for identifying school-based CRTs of interventions for improving health outcomes on pupils had already been developed (as used in Chapters 3 and 4) and because the resulting process used to find the papers was repeatable. A systematic search approach was the most

practical and rigorous way of finding relevant papers reporting estimates of the ICC from school-based CRTs with pupil health outcomes.

Mann-Whitney and Kruskal-Wallis tests were used to compare the ICC across subgroups of papers defined by study characteristics. The Mann-Whitney test was used to compare the ICC estimates between two subgroups and the Kruskal-Wallis test was used to compare the ICC estimates across three or more subgroups. These tests, the non-parametric alternatives to using the two-sample *t*-test and analysis of variance, respectively, were used because the ICC is not Normally distributed.

2.4 Chapter 6 overview

The fourth and final study in the thesis (Chapter 6) was a secondary analysis using data from five school-based CRTs to estimate ICCs and components of variance at different levels of clustering for pupil mental health and social emotional functioning [124] outcomes (e.g., mental health, mood, well-being, self-esteem, bullying, school climate). The roles of the researchers involved in the study are specified in the 'Author's Declaration' section of the thesis.

2.4.1 Aims and rationale

The aim of this study was to use raw data from five UK school-based CRTs to estimate ICCs and components of variance at different levels of clustering for pupil social emotional functioning outcomes and compare estimates of the ICC across studies, for different levels of clustering (e.g., school- versus class-level), for different reporters for the same outcome (i.e., pupils, parents, teachers), and over time.

The richness of the raw data provided the opportunity to examine ICC patterns in greater depth than was possible based on only using reported data in published papers. The analysis of raw data also facilitated the use of *within*-study information to identify the determinants of the ICC, thus avoiding the limitation of study-level confounding that results when comparing ICC estimates *across* studies as was undertaken in Chapter 5. Raw data also provided the opportunity to comprehensively report the ICC for all relevant outcomes in the studies. This in-depth exploration of the patterns of ICCs and components of variance will aid researchers when calculating the sample size in future CRTs.

2.4.2 Methodology

Mixed effects (“multilevel”) linear regression models were fitted to the pupil mental health and social emotional functioning outcomes to estimate ICCs and variance components. Mixed effects models are characterised by having both fixed effects for participant and cluster characteristics (e.g., trial arm status, pupil age, percentage of children in school that are eligible for free school meals) and random effects (residuals) that are used to explicitly model the variation in outcome across clusters and across individuals within the clusters [23]. Mixed effects models are fitted for data with a hierarchical or clustered structure, such as encountered in CRTs. For example, a simple two-level model mixed model would allow for clustering of pupils (level 1) within schools (level 2) by including random effects at each of those levels. The school-level random effect is the effect of unobserved school characteristics that is common for all pupils in a given school. The model explicitly recognises the correlation of outcomes between pupils from the same school.

A two-level mixed effects model can be fitted to estimate the ICC from data that have a single level of clustering:

$$Y_{ij} = \alpha + u_i + e_{ij}$$

- Y_{ij} is the outcome for the j^{th} individual in the i^{th} cluster
- α is the constant
- u_i is the random effect of the i^{th} cluster, assumed to be Normally distributed with 0 mean and constant variance σ_u^2
- e_{ij} is the residual effect of the j^{th} individual in the i^{th} cluster assumed to be Normally distributed with 0 mean and constant variance σ_e^2

Using the same formula previously described in Section 1.2.1, the ICC (ρ) is calculated from the between-cluster (σ_u^2) and within-cluster (σ_e^2) components of variances which are estimated by the model using:

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$$

2.8 Chapter summary

This chapter explained the overarching aims of the four studies in the thesis and the rationale for investigating the specific research questions. The chapter also described the methodological approach used in each study and justified the choice of methodology.

Page intentionally left blank

Chapter 3: The characteristics and practices of school-based cluster randomised controlled trials for improving health outcomes on pupils in the United Kingdom: A systematic review of definitive trials

3.1 Summary

This chapter reports a systematic review of the characteristics and practices of definitive school-based CRTs that assessed interventions for improving health outcomes on pupils in the UK. A peer-reviewed journal article of the protocol [116] (Appendix 1), and an article reporting the findings of this systematic review [117] (Appendix 2) have been published. The role of the authors in these publications has been previously specified (see Author's Declaration).

3.2 Background

As discussed previously in Chapter 1, the CRT design is increasingly used in school settings to evaluate interventions for improving child and adolescent health outcomes [50, 53]. A systematic review published in 2011 examining the characteristics and quality of reporting of CRTs in child health research found an increase in the rate of publication of such studies between 2004 and 2010, with 72% using school clusters as the units of randomisation [53]. Despite this, no systematic review has specifically focussed on the characteristics and practices of school-based CRTs assessing interventions for improving the health outcomes of pupils.

The systematic review aims to identify common methodological challenges associated with conducting CRTs in school settings and provides valuable information for researchers planning similar trials. The review collates estimates of parameters (e.g., estimates of the intra-cluster correlation coefficient (ICC), sample sizes, follow-up rates) from the included studies that are of use to researchers. Furthermore, the findings of the systematic review can be used to inform the design of simulation studies that use synthetic data to evaluate

the properties of statistical methods applied in the context of school-based CRTs with health outcomes.

3.3 Aims and Objectives

The aim of this systematic review was to summarise the characteristics and methodological practices of definitive school-based CRTs undertaken in the UK that evaluated interventions for improving pupil health outcomes.

The review examined several aspects of methodology and study design. These included: participant and setting characteristics; study design; intervention type; health area and outcome measures; recruitment and retention, sampling and allocation methods; sample size calculation; consent and ethical approval procedures; and analysis methods.

The objectives of the systematic review were to:

- Conduct a systematic review of definitive school-based CRTs used to assess interventions for improving health outcomes on pupils in the UK.
- Summarise the methodological characteristics of the included studies.

3.4 Methods

The systematic review was reported in accordance with the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) statement [125]. The review protocol was registered on the International Prospective Register of Systematic Reviews (PROSPERO) (Registration number: CRD42020201792) and was also published in a peer-reviewed journal article [116] (Appendix 1).

3.4.1 Search strategy

3.4.1.1 Developing the search strategy

The search strategy for the Medical Literature Analysis and Retrieval System Online (MEDLINE) database (via Ovid) is described in Table 3.1. Development of the search strategy is outlined below.

Table 3.1. MEDLINE (via Ovid) search strategy

Search strategy
Randomised controlled trial terms:
1. random:.mp.
2. trial.ab, kw, ti.
Study type terms:
3. "cluster*".ab, kw, ti.
4. "communit*".ab, kw, ti.
5. group*adj2 random*.ab, kw, ti.
6. 3 OR 4 OR 5
School terms:
7. exp Schools/
8. School*.ab, kw, ti.
9. 7 OR 8
Final search:
10.1 AND 2 AND 6 AND 9
11.10 limited to English language

The RCT concept

The *RCT* concept terms were identified using an RCT filter for MEDLINE [126]. '*random:.mp.*' and '*trial*ab,kw,ti.*' were used as *RCT* concept terms as the current review aimed to identify randomised controlled trials. Terms such as '*placebo*' and '*clinical trials*', which were also included in the RCT filter, were removed as the review did not seek to identify clinical trials. '*random:.mp.*' was used as *.mp.* represents a term that is found in any field (i.e., title, abstract, key words). This term also works similarly to a Medical Subject Heading (MeSH) term in that it also encompasses official words or phrases that represent or are similar to the word of interest (i.e., '*random*' encompassed words such as '*randomly*' or '*randomised*'). There was no similar MeSH term for *trial*, therefore, '*trial*ab,kw,ti.*' was used in its truncated form in order to identify these terms in the abstract, keywords and title.

The study type concept

The *study type* concept was developed based upon the sensitive MEDLINE search strategy for identification of cluster randomised trials developed by Taljaard et al [127]. Developing on the free text terms (cluster\$ adj2 randomi\$.tw., ((communit\$ adj2 intervention\$) OR (communit\$ adj2 randomi\$)).tw., group\$ randomi\$.tw.), the search strategy in this

systematic review focused on the terms '*cluster*', '*community*' and '*group*' in order to develop the *study type* concept terms. Therefore, the terms used in the final search strategy were "*cluster**".*ab,kw,ti*, "*communit**".*ab,kw,ti*. and *group*adj2 random*.ab,kw,ti.* '*Cluster*' and '*community*' were both truncated (removal of the end of the word) and searched for in the abstract, keywords and title. A decision was made to remove the '*adj2 randomi\$*' from these terms because the *study type* concept terms in this search were combined with the *RCT* concept terms using the Boolean search command '*AND*', thus it would be unlikely terms from both concepts would be unrelated. Despite this, the search strategy included the term *group* adj2 random*.ab,kw,ti.* instead of using the term '*group*' on its own as this would result in too many unrelated results, for example, identifying terms like '*control group*', '*intervention group*' or '*treatment group*'.

The *setting (school)* concept

The *setting (school)* concept used the exploded *schools* MeSH term, *exp schools/*, and the free text term, *School*.ab, kw, ti*. The MeSH term was chosen to increase precision and efficiency of the search strategy as it encompasses official words or phrases that represent the *school* concept, instead of listing school related terms such as '*classroom*' or '*year group*'. The free text term '*School*.ab, kw, ti*' was also used as a precaution as articles can sometimes be indexed incorrectly using MeSH terms.

Terms in the concepts *study type* and *setting* were combined with the Boolean search command '*OR*'; for example, for the '*setting (school)*' concept, *exp schools/OR School*.ab, kw, ti*. All three concepts (*RCT* concept, *study type (CRT)* concept, and *Setting (school)* concept) were combined with the Boolean search command '*AND*' to produce the final search strategy. The search was then limited to English language as available resources made it unfeasible to translate papers.

3.4.1.2 Database choice

The search was run in MEDLINE to identify peer reviewed articles published from inception to 30th June 2020. Once the search strategy had been developed in MEDLINE, a pragmatic decision was made not to translate the search strategy to other databases.

Scoping had identified a wealth of literature in this area, and, although other databases were considered, MEDLINE is focused on health-related journals of interest and was manageable

with the limited time and resources available for the review. There is substantial overlap in the studies indexed in Excerpta Medica Database (EMBASE) and MEDLINE, therefore it was not considered to be time-efficient to search EMBASE [128].

Subject specific databases, such as Psychological Information Database (PsycINFO), were considered but were not used as this may have biased the search in favour of certain health areas (i.e., in the case of PsycINFO there would be a bias towards psychological research). Additionally, as this systematic review focused on health outcomes, using the Education Resources Information Centre (ERIC) would have resulted in screening many studies with educational outcomes that would ultimately be ineligible. Other multidisciplinary databases were also considered, such as Web of Science and Google Scholar, but significant overlap with MEDLINE was also anticipated and would have resulted in a large number of studies to screen with little return.

In order to validate the use of MEDLINE, scoping searches were undertaken by translating the search strategy into the EMBASE, the Database of Abstracts of Reviews of Effects (DARE), PsycINFO and ERIC databases, and titles and abstracts of publications from 1st January 2018 to 30th June 2020 were screened for potential eligible studies that were not identified in MEDLINE.

3.4.1.3 Limiting the search to the UK

The review focused on the studies undertaken in UK educational settings, rather than internationally. This was partially due to the considerable number of school-based CRTs published worldwide identified during scoping and the limited available resources, as only two reviewers were involved in the screening and data extraction process. Additionally, by focusing on one education system the findings of this systematic review would be more applicable to a given setting.

3.4.2 Eligibility criteria

The PICOS (Population, Intervention, Comparison, Outcome, Study type) framework is used to develop health-related research questions and eligibility criteria for systematic reviews [129]. As the systematic review did not focus on one type of intervention or comparison group, the PICOS framework was used loosely to describe the eligibility criteria in the review.

Eligible articles were those reporting the results from UK school-based CRTs used to evaluate health-related interventions that measured at least one primary health outcome on school pupils.

3.4.2.1 Population

The population of interest was children/young people of school age in full-time education in the UK. Participants were pupils in pre-school (1-4 years), primary school (4-11 years), secondary school (11-16 years), or sixth form/college settings (16-18 years). 'Pre-school' was defined as an organisation offering early childhood education to children before they begin compulsory (primary) education [130]. This included nursery schools and kindergartens.

The types of eligible clusters included schools, year-groups, classes/classrooms, teachers or any other relevant school-related unit. Studies that randomised school-related units as well as other types of clusters (e.g., communities, households) were eligible for inclusion in the review as long as the study characteristics of interest were reported separately for the school clusters (i.e., the authors did not pool summaries of characteristics across different types of clusters). Any school type was eligible for inclusion, including fee-paying and special needs schools.

3.4.2.2 Intervention

Any health-related interventions were considered. Interventions that were administered to the teachers, parents/carers or other third party which influenced the pupil were considered (e.g., training teachers in mindfulness), as long as the primary health outcome was measured on the pupil themselves. Interventions that were designed to specifically improve educational outcomes (e.g., academic test scores) were excluded. Interventions could be targeted (i.e., intervention for a specific subset of individuals within the population) or universal (i.e., intervention for any individuals within the population) in their mode of delivery [131].

3.4.2.3 Comparison

Studies had to use a control/usual care comparison group(s). Any type of comparator was eligible including active control group(s). An active control was defined as, 'a *control group*

in which participants engage in some task during the intervention period that differs from normal practice' [3](p88-89).

3.4.2.4 Outcome

The primary outcome had to be health-related and measured on school pupils. Studies for which the primary outcome was not health-based (e.g., improved academic test scores) were excluded, as well as studies that did not measure the primary outcome on pupils (i.e., on teachers or parents/carers).

3.4.2.5 Study type

Eligible studies had to use a CRT design. All types of CRT design were eligible, including parallel group, crossover, factorial and stepped wedge. Non-randomised and single arm trials were excluded as randomisation was a key methodological characteristic that this systematic review wished to investigate. Quasi-experimental designs (i.e., no random assignment) were also excluded.

3.4.2.6 Other eligibility criteria

Articles not published in English were excluded due to time and resource constraints to translate the papers into English. However, as this review focused on UK-based studies, this was not anticipated to be problematic.

Only definitive CRTs were included in the systematic review. Definitive studies were defined as, 'trials in which a pre-specified hypothesis is evaluated using a pre-defined methodology in order to provide sufficient or unequivocal evidence about a treatment's benefit to the participant [132]'.

If more than one publication of the primary outcome result for an eligible CRT was identified (i.e., sibling paper), an index paper was designated and used for data extraction. The index paper was defined as the first paper to publish the primary outcome.

Articles that did not report the primary outcome were excluded, along with pilot/feasibility studies (a 'small study for helping to design a further confirmatory study' [43]), protocol/design articles (a detailed plan of a study), process evaluations (an examination of how an intervention improves outcomes or why it does not improve outcomes), economic

evaluations/cost-effectiveness studies (a simultaneous comparison of the costs and outcomes of health care interventions), statistical analysis plans (description of the methods to be used in analyses of the trial data), commentaries (an explanatory series of notes or comments), and papers reporting only findings from mediation/mechanism analyses (used to explore the underlying mechanism or process by which the intervention influences the outcome).

3.4.3 Screening and selection

Titles and abstracts of the identified studies were retrieved from the MEDLINE database and exported to Endnote (X9) [133]. Any duplicate citations were removed using the 'Find duplicates' function in Endnote. The remaining citations were independently screened by KP and one other reviewer (OU) for eligibility against the inclusion criteria described above. Articles were coded (1) if they were thought to be eligible, or (2) if not. Once both reviewers had finished coding, the two Endnote libraries were merged in order to identify the articles where the coding differed. Disagreements were resolved through discussion, and any studies where uncertainty of inclusion remained were included in the full text screening stage.

A new Endnote library was created with the potentially eligible articles and PDF versions of each article were obtained for full text screening. Endnote was used to code each paper with a reason for inclusion/exclusion, using a letter that indicated the justification. This method was first piloted on a random sample of 15 articles. The coding reasons are provided in Table 3.2.

Table 3.2. Reasons for exclusion at full text screening

Code	Reason
a	Include
b	Exclude – Not undertaken in the UK
c	Exclude – Not a CRT
d	Exclude – Not school-based/schools not randomised
e	Exclude – Primary outcome not reported on pupils
f	Exclude – Not main outcome paper
g	Exclude – Sibling paper of index study
h	Exclude – Pilot/Feasibility study
i	Exclude – Protocol/Design
j	Exclude – Baseline results
k	Exclude – Mediation/mechanism analysis
l	Exclude – Cost-effectiveness/economic evaluation
m	Exclude – Sub-group analysis
n	Exclude – Statistical analysis plan
o	Exclude – Process evaluation
p	Exclude – Commentary

Two independent reviewers (KP and OU) then assessed articles for inclusion based on the criteria using the coding method. Once all articles had been coded, the two Endnote libraries were merged to identify the articles where coding differed. Disagreements which could not be resolved through discussion were sent to a third reviewer (ZMX) for a decision.

3.4.4 Data extraction

Before data extraction, each article was assigned a study ID number. Data extraction variables were developed after examining similar methodological systematic reviews [53, 111]. A bespoke data extraction form was developed using Microsoft Excel and initially piloted on a random sample of 10 included papers.

Developing the data extraction form was an iterative process and changes were made following piloting and throughout data extraction. Additional variables were added to aid the refinement and classification of the extracted data (this is detailed in Section 3.4.5). The final list of categories and associated variables that were extracted are presented in Table 3.3.

Table 3.3. Data extracted from included studies

Characteristic	Variables
<i>Publication details</i>	Author Surname; Year of publication; Title; Journal name; Corresponding author; Affiliation of first author; Funding sources.
<i>Setting and participant characteristics</i>	Country (England, Scotland, Northern Ireland, Wales); Region (e.g., South West England); School level (pre-school, primary, secondary, sixth-form/college); School type (state, local authority, foundation and voluntary-aided, academy, grammar, special, faith, independent); Co-educational status (co-ed, female only, male only); Age(s) of pupils; Year group(s) of pupils; Gender (female, male, both); Inclusion and exclusion criteria of the study at both the cluster and individual level (e.g., only included schools in deprived areas).
<i>Intervention</i>	Health area (e.g., dental health); Was the intervention universal or targeted? (i.e., aimed at all pupils or a subset of pupils); Was the intervention for primary prevention or secondary prevention? (i.e., did the intervention aim to prevent or treat the health problem); Who trained the intervention administer? (e.g., researcher); Who administered the intervention? (e.g., teacher); How was the intervention delivered? (e.g., through marketing material); Was there an intermediate target of intervention? (e.g., parents/carer); Intervention typology classification ¹ (Eldridge typology [3](p25-29) – individual-cluster, professional-cluster, cluster-cluster, external-cluster, multifaceted); Type of control group (e.g., usual care, active control); Was a wait-list (delayed intervention) control group used? (yes, no, not stated).

Characteristic	Variables
<i>Primary outcome</i>	Health area of the primary outcome (e.g., dental health); Name of primary outcome (e.g., body mass index); Type of outcome (e.g., continuous); Reporter of the primary outcome (e.g., pupil); Method of data collection/reporting (e.g., questionnaire); Was the primary outcome reporter blind to allocation status? (yes, no, not stated); Was the outcome assessment objective? (yes, no, not stated).
<i>Study design and analysis methods</i>	Was justification provided for using CRT design? (yes, no, not stated); If 'Yes', what was the justification? (i.e., to prevent contamination between trial arms); Unit of randomisation (e.g., school, classroom); Was there an intermediate level of clustering? (yes, no, not stated); If there was intermediate level of clustering, what was it? (e.g., classes); Type of CRT design used (e.g., parallel group, cross-over, factorial); Method used to sample schools (e.g., convenience sampling); Was allocation concealment used for the randomisation process? (yes, no, not stated); Was there allocation concealment with respect to the pupils? (yes, no, not stated); Were pupils recruited before the clusters were randomised? (yes, no, not stated); Were baseline data collected before clusters were randomised? (yes, no, not stated); Number of trial arms; Allocation ratio (e.g., 1:1 ratio); Method used to balance the randomisation (e.g., completely randomised, matched pair, stratified); What factors were used to balance the randomisation? (e.g., deprivation); Design of follow-up (e.g., cohort, repeated cross-sectional); Total length of follow-up (in months); Total number of follow-ups; Was the outcome reporter blind to trial arm they were randomised to? (Yes, no, not stated); Method used to account for clustering in the analysis (e.g., random effects linear regression); Baseline factors that

Characteristic	Variables
	were adjusted for in the analysis (e.g., school size); Was an intention-to-treat analysis used? (yes, no, not stated); Was multiple imputation used to account for missing data in the main analysis? (yes, no, not stated); Was a subgroup analysis undertaken? (yes, no, not stated)
<i>Sample size calculation</i>	ICC assumed in the sample size calculation; Assumed between-cluster coefficient of variation (CV) of the outcome in the sample size calculation (if provided); Where was the ICC (or CV) used in the sample size calculation obtained? (e.g., pilot study, reference); Assumed design effect (if provided); Power; Type 1 error rate; Was drop-out at cluster and/or individual level anticipated in calculation? (yes, no, not stated); Were equal or unequal cluster sizes assumed? (equal, unequal, not stated); Assumed coefficient of variation of cluster size; Assumed standard deviation of cluster size; Was intermediate level of clustering explicitly allowed for in sample size (yes, no, not stated, not applicable); Target number of clusters to recruit; Target number of pupils to recruit; Target number of pupils to provide data at follow-up.
<i>Ethics and consent procedures</i>	Was ethical approval granted? (yes, no, not stated); Name of committee that provided ethical approval; Was consent/assent sought for randomisation, the intervention and data collection from the headteacher/administrator (cluster-level consent), the parent/guardian and the child (yes, no, not stated); Was passive “opt-out” consent used? (yes, no, not stated); Were harm/adverse events recorded during the study? (yes, no, not stated).

Characteristic	Variables
<i>Other areas of methodological importance</i>	Number of schools, clusters and pupils that were recruited; Number of schools, clusters and pupils that were followed-up; Percentage of female pupils at baseline; Percentage of pupils of white ethnicity at baseline; Deprivation level of school (e.g., the mean Income Deprivation Affecting Children Index (IDACI) score); Coefficient of variation of cluster size; Mean (standard deviation) cluster size; Median (interquartile range; range) cluster size; p-value for the primary analysis of intervention effect; ICC estimate of the primary outcome; Was the ICC from adjusted analysis? (yes, no, not stated); Did any participants change cluster membership? (e.g., move between school clusters); Harms/adverse events; Methodological challenges highlighted by authors.

¹ Added post-hoc to aid classification.

Once agreement on the understanding of the data extraction form had been reached, data were extracted in full by two independent reviewers (KP and OU). If there were disagreements regarding particular items, the data obtained were checked by a third reviewer (ZMX) and resolved by further discussion. Missing information that was not available in the index papers was obtained from corresponding protocol papers and other sibling publications of the studies. No attempt was made to contact corresponding authors for missing information due to time and practical constraints.

3.4.5 Data processing

Once the data were extracted for all included texts, some data were processed by coding or further classification for ease of analysis. Data were originally extracted exactly as provided in each article. If data were not provided for a variable or the information was unclear, this was recorded as 'not stated'. This section details specific variables where coding/classification was used and specifies assumptions that were made during data extraction.

3.4.5.1 Setting and participant characteristics

If a region was not stated but a local authority or city or county was provided, then this information was used to identify the region (e.g., Exeter was recorded as South West, England).

School type was recorded as stated in the article and then categorised as listed on the UK government website [134]. State schools (also called comprehensive, state-maintained, state-funded) receive funding through their local authority or directly from the government. The most common types of state school in the UK are local authority, foundation and voluntary-aided schools. Academies are schools run by government and not-for-profit trusts and are independent of local authority. Grammar schools are run by local authorities, but intake is based on assessment of the pupils' academic ability. Special schools cater for pupils with special educational needs. Faith schools follow the national curriculum but can decide what they teach in religious studies. Independent schools do not need to follow the national curriculum and charge fees for attending pupils.

Additionally, school level and year groups across the devolved nations in the UK were standardised in relation to their equivalent school level and year group in

the English school system (i.e., *pre-school* (1 to 4 years), *primary school* (4 to 11 years), *secondary school* (11 to 16 years) and *sixth form/college* (16 to 18 years)). A table comparing school year groups across nations in the UK can be found in Appendix 5 [135].

3.4.5.2 Intervention type

The '*health area of the intervention*' categories were decided on by examining previous systematic reviews [53, 111]. For example, health difficulties such as mental health, behaviour, neurodiversity (e.g., Attention deficit hyperactivity disorder (ADHD)), well-being, quality of life, bullying, social and emotional learning, and self-esteem were categorised under '*Social emotional functioning and its influences*'.

'*Intervention type*' was summarised using the typology described by Eldridge et al [3](p25-29). The categories were as follows: '*Individual-cluster*' interventions which include components that are directed at individual participants (e.g., pupils) on whom outcomes are measured; '*Professional-cluster*' interventions which include components for training professionals in the cluster (e.g., teachers in schools) to deliver the intervention; '*External-cluster*' interventions which involve using additional staff outside the cluster to deliver the intervention (e.g., researchers, trained facilitators); '*Cluster-cluster*' interventions which include components that necessarily have to be administered to entire clusters (e.g., school policy); '*Multifaceted*' interventions which include components across more than one of the '*individual-cluster*', '*professional-cluster*', '*external-cluster*' and '*cluster-cluster*' categories.

Form of delivery was described using the most common classifications for the methods of delivering the intervention (e.g., videos, worksheets for use in lessons were recorded as 'lesson materials').

Interventions had components that were recorded as being 'universal', 'targeted' or 'indicated'. A universal intervention was defined as '*an intervention that is aimed at the whole population*'. A targeted intervention (sometimes called a selective intervention) was defined as '*an intervention which targets a subgroup of the population deemed at risk of developing a particular health problem*'. An indicated intervention was defined as '*an intervention which targets a subgroup*'.

of the population already exhibiting particular health problems or behaviours' [136].

Primary prevention was defined as '*an intervention that aims to prevent a disease or injury before it ever occurs*' (e.g., legislation of health eating practices). Secondary prevention was defined as '*an intervention that aims to reduce the impact of a disease or injury that has already occurred*' [137].

Type of control group was recorded as 'usual care' or 'active'. An active control was defined as, '*a control group in which participants engage in some task during the intervention period that differs from normal practice*' [3](p88-89). If the study had more than one control group this was also recorded.

Intermediate target of the intervention is an individual that is targeted by the intervention (e.g., teachers) but not the primary target for whom outcomes are measured on (e.g., pupils). The intermediate target will have influence over the primary target of the intervention [137].

3.4.5.3 Study design

Justification for use of the CRT design was categorised into reasons commonly cited and established in the methodological literature [3](p10-13). If the study provided more than one reason, multiple justifications were recorded.

Factors used to balance the randomisation were categorised into common themes. For example, *Deprivation* included the factors: percentage of pupils in the school that were eligible for free school meals; Townsend Index of Deprivation [138]; Income Deprivation Affecting Children Index (IDACI) [139]; and Index of Multiple Deprivation (IMD) [139].

It was assumed that a completely randomised design was used unless otherwise stated.

An objective outcome assessment was defined as, '*a measurement that is impartial and is usually measured by a type of diagnostic instrument (e.g., accelerometer)*' [140].

Allocation concealment with respect to the pupils was defined as, an approach used to prevent selection bias by concealing the allocation sequence. In other words, the pupil and anybody involved in recruiting that pupil do not know what

trial arm the pupil will be assigned to if they agree to take part. Recruitment of pupils before randomising clusters is one way of ensuring this.

For the extraction variable, 'Were participants recruited before the clusters were randomised?', the CONSORT flow diagram was primarily used to determine this. If it was not clear from the CONSORT flow diagram then information was extracted and verified from the main text of the article.

3.4.5.4 Sample size assumptions

Target mean cluster size was calculated by dividing the target number of pupils at follow-up by the targeted number of clusters. Some articles did not provide the design effect (DE), therefore, this was calculated as: $DE = 1 + ((\text{targeted mean cluster size at follow-up} - 1) \times \text{assumed ICC in sample size calculation})$. When calculating the DE, it was assumed that all clusters had the same number of participants.

3.4.5.5 Consent and ethical approval

Consent was defined as *an agreement given by parent/carer*. Assent was defined as *an agreement given by a child/young person who is not legally empowered to give consent*. Information on whether consent and/or assent was obtained for randomisation, partaking in the intervention and data collection were extracted from the articles. This was recorded at different levels: from the cluster gatekeeper (individuals or bodies that represent the interests of cluster members, clusters, or organisations)[32]; headteacher/administrator (cluster-level); parent/carer; and child (individual-level). Consent/assent information was coded as whether consent/assent was obtained from the child, the parent/carer, both or neither. Passive 'opt-out' consent/assent was defined as, *the act of participants being included in research unless they give their express decision to be excluded (i.e., opt-out of the research)* [141].

3.4.5.6 Type of primary outcome

The primary outcome was identified as the health outcome stated in the paper as being the primary outcome. If there were multiple primary outcomes or the primary outcome was unclear, then the outcome presented in the title, first outcome presented in the *Outcome Measures* section in the methods, or first

outcome presented in the *Results* section was taken as the primary outcome (in this order of priority).

Primary outcome health area was categorised into broad health areas defined after consulting previous systematic reviews [53, 111]. For example, primary outcomes in the health area of mental health, well-being or behaviour were categorised into '*Social emotional functioning and its influences*'.

3.4.5.7 Analysis of primary outcome

The method of analysis used in each study to compare the primary outcome between the trial arms was categorised into broader approaches of analysis. For example, Generalised Estimating Equation (GEEs) was categorised as '*Adjusted individual-level analysis*'. If there was uncertainty regarding which was the primary time point, the last data collection time point was chosen.

3.4.5.8 Methodological parameter estimates

Information on recruitment and drop-out of clusters and pupils was extracted from the CONSORT flow diagram the included papers.

The percentage of pupils who were female was extracted as this was most commonly reported in the included studies. If there was no overall percentage of female pupils provided across trial arms, the percentage of female pupils was reported for the control arm.

Ethnicity was recorded as the percentage of white students as other ethnicities were often not reported or the manner in which the information was presented differed across studies.

Measures of deprivation were not easy to summarise across studies due to the number of different measures of deprivation. Therefore, a record was made of whether socio-economic status (SES)/deprivation was reported at the cluster and/or individual level.

When recording the number of clusters followed up for the primary outcome analysis, an assumption was made that all clusters were followed up unless stated otherwise.

3.4.6 Assessment of study quality

A quality/risk of bias assessment was not appropriate for this methodological systematic review as the focus was not on the specific estimates of intervention effects in the included studies. The review only aimed to describe the characteristics of the studies. However, much of the information extracted in the systematic review is indicative of quality in CRTs [57]. Examples of such information included: *justification for the use of the CRT design; whether allowed for clustering in the sample size calculation and analysis; whether the study used matching, stratification, or an alternative means of reducing chance imbalances on cluster-level characteristics at randomisation.*

3.4.7 Data analysis

Results of the search process were reported using a PRISMA flow diagram [125]. The reasons for exclusion at full text screening were also reported in the PRISMA flow diagram.

Once the data had been checked and 'cleaned' following data extraction, Stata 16 software [142] was used to undertake statistical analysis. Study characteristics were summarised using means and standard deviations, or medians and interquartile ranges for continuous variables, and numbers and percentages for categorical variables. A histogram was used to summarise the year of publication. A scatterplot was used to summarise the relationship between the ICC assumed in the sample size calculation and the estimated ICC for the primary outcome from the study data. Challenges (e.g., recruitment and retention difficulties) reported by the authors of each article were summarised narratively. Meta-analysis of the intervention effect was not appropriate as the review focuses on summarising methodological characteristics not evaluating the interventions

3.5 Results

3.5.1 Study selection and PRISMA flow diagram

Figure 3.1 summarises the flow of studies through the review. The search of the MEDLINE database from inception to 30th June 2020 identified a total of 3138 articles. After deduplication using Endnote, the titles and abstracts of 3103 articles were screened. This resulted in 159 texts which were included for full text screening. After full text screening, 64 articles were eligible and were included in the systematic review [143-206]. Agreement between reviewers on which articles should be included was 100%. Ninety-five (95) articles were excluded at full text screening, of which 19 articles were excluded because they reported on the same study as the index paper.

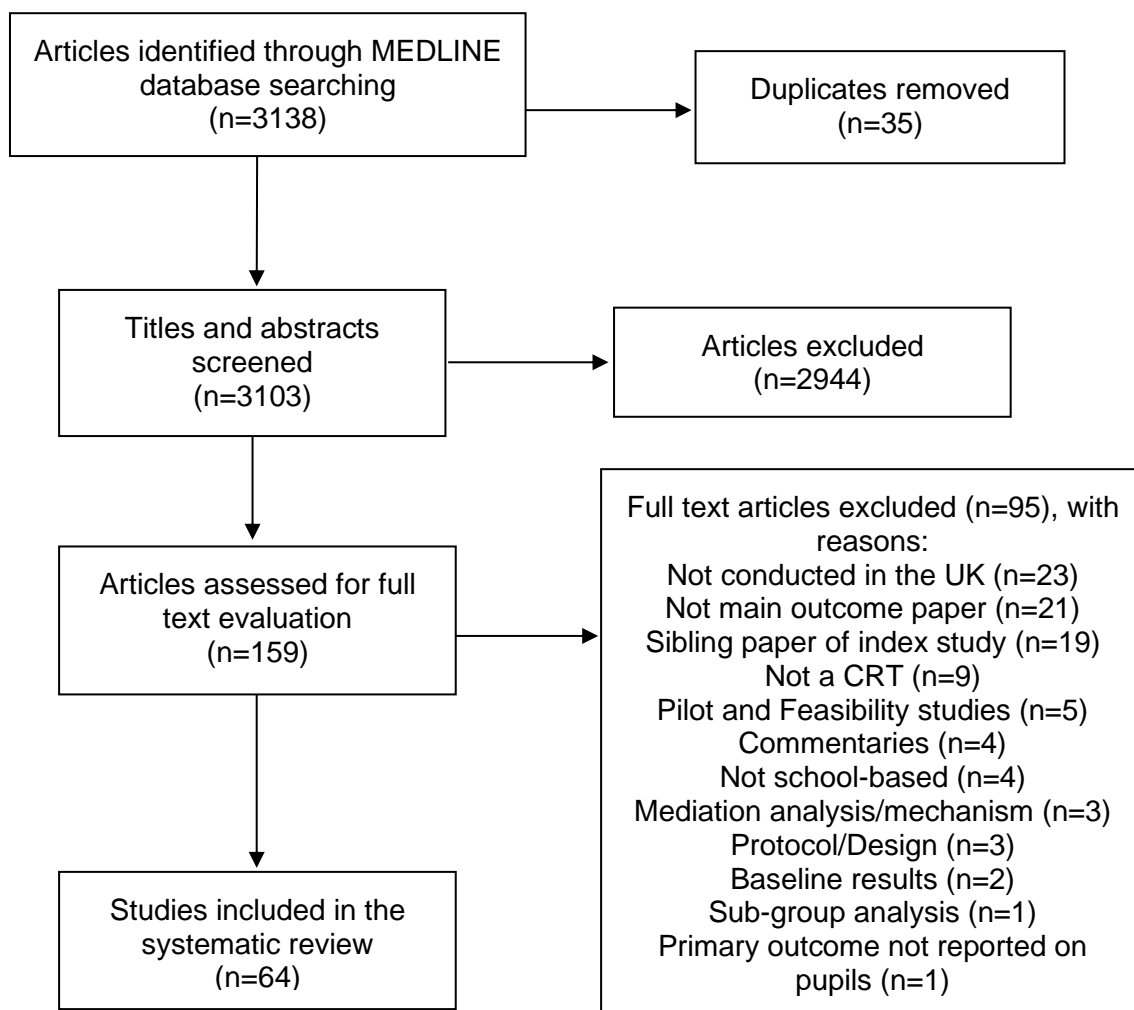


Fig 3.1. PRISMA flow diagram summarising the results of the literature search and screening for eligibility

3.5.2 Publication characteristics

The rate of publication of school-based CRTs evaluating interventions for improving health outcomes on pupils in the UK has increased since the earliest paper was published in 1993 (Figure 3.2). Twenty-three (23) articles were published between 2001 and 2010 compared with 37 articles in the ten years after (January 2011 to June 2020).

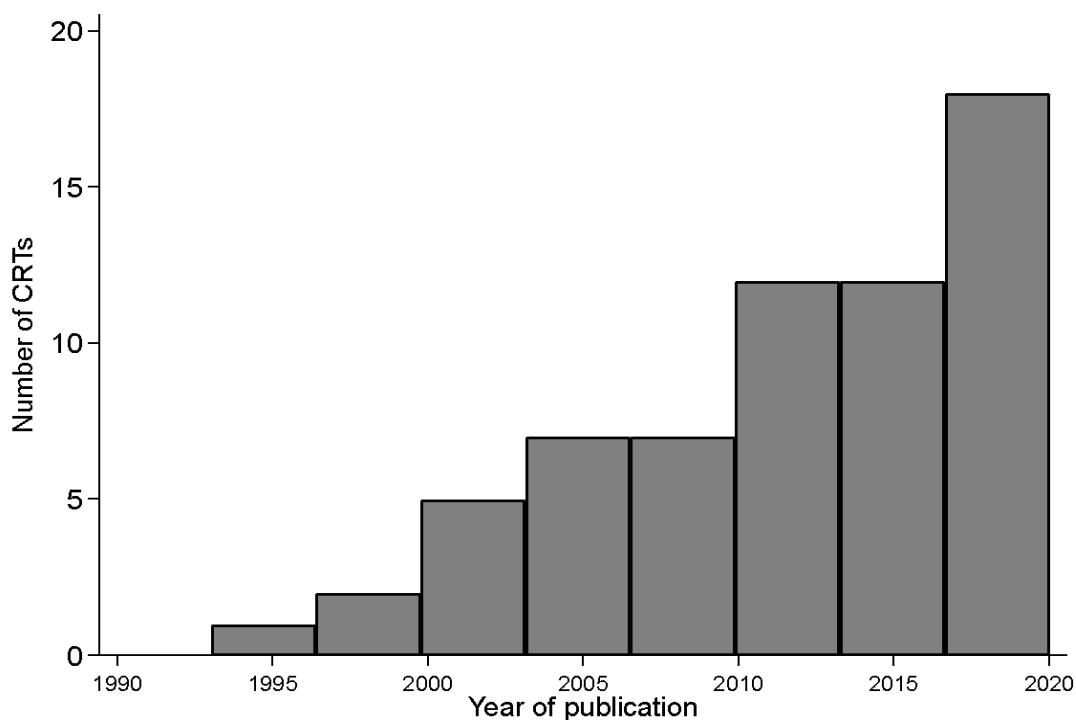


Fig. 3.2. Published CRTs indexed in MEDLINE from inception to 30th June 2020 (N=64)

The 64 articles were published in 36 different journals, most commonly the British Medical Journal (n=9; 14%). Sixty-one (61) studies stated where the funding for their research came from. Fifty-seven (57) different sources of funding were identified, of which 19% (n=11) were funded by the National Institute of Health Research (NIHR) Public Health Research programme. Further details on journals and funding sources are reported in Appendix 6.

3.5.3 Setting and participant characteristics

England was the most common country for school-based CRTs to be undertaken in the UK, with 73% (n=47) exclusively conducted there. Five (8%) studies were exclusively conducted in Scotland [156, 165, 168, 172, 193], 3 (5%) exclusively in Wales [145, 181, 188], 3 (5%) studies exclusively in Northern Ireland [153, 163,

192], and 6 (9%) studies were conducted in more than one country in the UK [148, 149, 158, 183, 186, 190]. For 63% (n=40) of studies, the schools were selected from only one geographic region (e.g. South West England).

More than half of the studies (n=36; 56%) took place exclusively in primary schools (4-11 years), 38% (n=24) exclusively in secondary schools (11-16 years), and 3% (n=2) took place exclusively in pre-schools (2-4 years) [162, 170]. Two (3%) studies took place in both primary and secondary school settings [200, 201]. No studies took place in sixth-form or college settings (16-18 years).

Forty-four (69%) studies reported information regarding types of schools recruited in the study. Of these studies, 93% (n=41) included state-funded schools among their eligible school types.

The majority of studies focused on recruiting children of middle childhood/ upper years of primary school age (8-11 years). Only one study each recruited pupils of 2 years old [170] and 3 years old [162], respectively. In 60 (94%) studies both boys and girls were eligible to participate. In 3 (5%) studies only girls were eligible [167, 174, 195], and in 1 (2%) study only boys were eligible to participate [169].

Of the 64 studies, 20 (31%) had cluster-level inclusion criteria when selecting clusters, and 11 (17%) had individual-level inclusion criteria when selecting pupils to participate. Thirteen (20%) studies had cluster-level exclusion criteria, and 8 (13%) had individual-level exclusion criteria.

Settings and participant characteristics are summarised in Table 3.4.

Table 3.4. Setting and participant characteristics of included studies (N=64)

Characteristic	N	Statistics (n (%))
Country	64	
England		47 (73)
Scotland		5 (8)
Wales		3 (5)
Northern Ireland		3 (5)
More than one country ¹		6 (9)
Number of regions from which schools were drawn ²	64	
One		40 (63)
Two		10 (16)
Three		1 (2)

Characteristic	N	Statistics (n (%))
Four		1 (2)
Unclear		12 (19)
School level	64	
Pre-school only		2 (3)
Primary only		36 (56)
Secondary only		24 (38)
Primary and Secondary		2 (3)
School types that were included [134] ³	44	
State		41 (93)
Independent		6 (14)
Academies		2 (5)
Grammar		2 (5)
Special		2 (5)
Voluntary-aided		2 (5)
Foundation		1 (2)
Faith		1 (2)
Age of pupils eligible to participate (years)	64	
2		1 (2)
3		1 (2)
4		5 (8)
5		8 (13)
6		10 (16)
7		19 (30)
8		24 (38)
9		27 (42)
10		20 (31)
11		23 (36)
12		19 (30)
13		15 (23)
14		12 (19)
15		6 (9)
16		3 (5)

¹ Studies that included schools from more than one country in the United Kingdom.

² England regions included: South West, South East (including Greater London), East of England, West Midlands, East Midlands, North West, North East, Yorkshire and The Humber, "Southern England", "Central England" and "West of England". Scotland regions included: Glasgow, Inverclyde, Tayside, Grampian, Lanarkshire, Lothian and Fife. Wales regions included: North Wales, South West Wales and South East Wales. Northern Ireland areas included: South Belfast, East Belfast, Ulster, Leinster, Connacht and Munster.

³ Some studies included more than one school type. This is the number of studies that included specific types of school.

3.5.4 Intervention type

Interventions in the included studies targeted 11 different health areas: *nutrition* (n=18; 28%) [143, 151, 155, 158-160, 165, 170, 175, 178-181, 186, 188, 195, 197, 201]; *physical activity* (n=15; 23%) [143, 147, 148, 155, 160, 162, 167, 170, 174, 178, 180, 189, 196, 197, 205]; *social emotional functioning and its influences* (n=15; 23%) [145, 146, 150, 153, 157, 161, 173, 191, 192, 198-200, 202, 203, 205]; *dental health* (n=7; 11%) [156, 166, 184, 185, 193, 194, 206]; *smoking* (n=5; 8%) [144, 149, 152, 182, 190]; *injury* (n=5; 8%) [169, 176, 177, 187, 196]; *sexual health* (n=3; 5%) [164, 168, 204]; *alcohol misuse* (n=2; 3%) [154, 183]; *cancer* (n=1; 2%) [172]; *communication skills (for children with Autism)* (n=1; 2%) [171]; and *health attitudes (breast feeding)* (n=1; 2%) [163].

The number of publications with interventions in the area of physical activity increased markedly (13 published in or after 2011 compared to 2 publications before 2011). Similarly, the number of studies evaluating interventions for improving social emotional functioning and its influences has also increased since 2011; of the 15 studies in this area, 13 were published since 2011. In contrast, of the 7 articles focusing on dental health interventions in schools, the most recent was published in 2011.

The interventions in 52 (81%) studies had components that were *universal* in their administration. Nine (14%) had intervention components that were *targeted* (e.g., for deprived schools) [148, 154, 162, 166, 170, 177, 186, 199, 202], and six (9%) had *indicated* components (e.g., for pupils with previous behavioural problems) [145, 171, 182, 191, 192, 198]. Sixty (94%) studies evaluated primary prevention interventions, and 4 (6%) evaluated secondary prevention interventions [171, 182, 191, 192].

The types of intervention components included: classrooms sessions (n=36; 56%); materials (e.g., lesson materials) (n=29; 45%); non-classroom based session (e.g., gardening) (n=6; 9%) [151, 159, 174, 180, 182, 192]; changing school environment (n=5;%) [181, 186, 188, 197, 201]; physical activity (n=3; 5%) [143, 147, 180]; application of dental varnish (n=2; 3%) [166, 184]; group sessions for parents (n=2; 3%) [198, 199]; one-to-one sessions for pupils (n=2; 3%) [180, 191]; peer support (n=2; 3%) [149, 205]; dental inspection (n=1; 2%) [156]; group meetings with teachers and students (n=1; 2%) [146]; group

sessions for pupils (n=1; 2%) [191]; group session for teachers (n=1; 2%) [198]; meetings (n=1; 2%) [196]; parents evenings (n=1; 2%) [183]; role play (n=1; 2%) [199]; screening (n=1; 2%) [185]; sports sessions (n=1; 2%) [169]; and supervised tooth brushing (n=1; 2%) [193].

An intervention trainer was used to train the intervention administrator (the person delivering the intervention) in 30 (47%) studies; this was most commonly a member of the research team (n=13). In just over half the studies (n=33; 52%), a member of school staff (i.e., teacher) delivered the intervention. There was an intermediate target of the intervention in 25% of studies (n=16).

Interventions were also classified using the typology based on the primary reason for adopting a clustered design [3] (p26). “*Cluster-cluster*” interventions were evaluated in 53 (83%) of studies. These interventions include components that necessarily have to be administered to entire clusters, such as educational lessons [163], gardening [151], breakfast clubs [201], provision of funding/resources [200], change in school policy [196] and advertisements [155]. In 11 (17%) studies, an “*individual-cluster*” intervention was evaluated. These interventions include components which are directed at the individual participant (pupil), such as use of fluoride varnish [166]. Just over half the studies (n=33, 52%) evaluated “*professional-cluster*” interventions. These interventions include components for training professionals in the cluster to deliver the intervention. In 30 (47%) studies the teacher was either trained in or provided with guidance to deliver components of the intervention, in 3 studies pupils themselves were trained to deliver peer-led intervention components [149, 204, 205], and in 1 study the school nurse was trained [192]. Thirty-two studies used “*external-cluster*” interventions. Such interventions involve using additional staff outside the cluster to deliver the intervention. External facilitators included researchers [163], trained facilitators [202], dental professionals [185], dance instructors [174] and student volunteers [148]. Fifty-two (81%) studies had multifaceted interventions that had more than one of the above types of intervention component.

Sixty (94%) studies described their type of control group. Thirty-three studies (55%) had a usual care control group, and 12 (17%) had an active control [150-152, 158, 159, 162, 169, 177, 182, 195, 202, 203]. Sixteen (25%) studies used a waitlist (delayed intervention) control arm where the control arm participants received the intervention after the final data collection point [145, 148, 161, 164,

165, 172, 176, 188, 189, 192, 194, 197, 198, 200, 205, 206]. Two studies [202, 203] had two control arms (one usual care and one active control). In the first study [202], the active control group received an 'attention control intervention' in which the class teacher delivered the usual curriculum for personal, social, and health education (PSHE), but two facilitators assisted with lesson delivery and engagement of the pupils. The usual care control group received no external input from the research team in PSHE lessons. In the second study [203], the active control arm received the 'school-led FRIENDS (10 sessions of cognitive behaviour therapy) programme' where sessions were led by a teacher trained in the programme and were supported by two facilitators. This differed from the intervention arm where the 'health-led FRIENDS programme' was led by two trained health facilitators working alongside the class teacher. The usual care control arm continued with usual PSHE sessions provided by the school and no external input from the research team.

Information on the type of interventions is summarised in Table 3.5.

Table 3.5. Intervention type characteristics of included studies (N=64)

Characteristic	N	Statistics (n (%))
Health area of intervention ¹	64	
Nutrition		18 (28)
Physical activity		15 (23)
Social emotional functioning and its influences ²		15 (23)
Dental health		7 (11)
Smoking		5 (8)
Injury		5 (8)
Sexual health		3 (5)
Alcohol misuse		2 (3)
Cancer		1 (2)
Communication skills (children with autism)		1 (2)
Health attitudes (breast feeding)		1 (2)
Delivery of intervention components	64	
Universal		52 (81)
Targeted		9 (14)
Indicated		6 (9)
Level of prevention	64	
Primary prevention		60 (94)
Secondary prevention		4 (6)
Type of intervention [3] ³	64	
Individual-cluster		11 (17)
Professional-cluster		33 (52)
External-cluster		32 (50)
Cluster-cluster		53 (83)
Multifaceted		52 (81)
Control group(s)	60	
Usual care ⁴		52 (87)
Active		12 (20)

¹ Some interventions targeted more than one health area.

² Includes mental health, behaviour, Attention deficit hyperactivity disorder (ADHD), well-being, quality of life, bullying, social and emotional learning, and self-esteem.

³ Many studies had 'Multifaceted' interventions that included components in more than one of the individual-cluster, professional-cluster, external-cluster and cluster-cluster categories.

⁴ Two studies used two control arms (one usual care and one active control group) [202, 203]

3.5.5 Primary outcome

The primary outcomes spanned 14 different health areas: *social emotional functioning and its influences* (n=15; 23%) [145, 146, 150, 153, 157, 161, 173, 191, 192, 198-203]; *nutrition* (n=10; 16%) [151, 158, 159, 165, 175, 179, 181, 186, 188, 195]; *adiposity* (n=7; 11%) [143, 147, 155, 160, 170, 180, 197]; *dental health* (n=7; 11%) [156, 166, 184, 185, 193, 194, 206]; *physical activity* (n=7; 11%) [148, 167, 174, 178, 189, 196, 205]; *smoking* (n=5; 8%) [144, 149, 152, 182, 190]; *injury* (n=3; 5%) [169, 177, 187]; *sexual health* (n=2; 3%) [164, 204]; *obstetrics* (n=2; 3%) [163, 168]; *alcohol misuse* (n=2; 3%) [154, 183]; *cancer* (n=1; 2%) [172]; *communication skills (for children with autism)* (n=1; 2%) [171]; *gross motor skills* (n=1; 2%) [162]; and *safety* (n=1; 2%) [176].

The most common primary outcomes were minutes per day of moderate-to-vigorous physical activity (MVPA) (n=5; 8%) [148, 167, 174, 178, 205], and body mass index (BMI) z-score (n=5; 8%) [143, 147, 170, 180, 197]. Forty-two (66%) primary outcomes were continuous variables, 18 (28%) were binary variables, 3 (5%) were count variables [168, 169, 184], and 1 (2%) was ordinal [171].

Questionnaires were the method of data collection in almost half the studies (n=31; 48%). Observations were used in 7 studies (11%) [158, 162, 169, 171, 181, 187, 199], accelerometer measurement in 6 studies (9%) [148, 167, 174, 178, 189, 205], anthropometric measurements in 6 studies (9%) [143, 147, 160, 170, 180, 197], dental assessment in 6 studies (9%) [166, 184, 185, 193, 194, 206]; diaries/recall were used in 6 studies (9%) [151, 159, 175, 186, 188, 195], and routine data in 2 studies (3%) [156, 168].

In just over half the studies (n=34; 53%) students self-reported the primary outcome. A member of the research team reported the primary outcome in 20% (n=13) of studies, dentists in 9% (n=6) [166, 184, 185, 193, 194, 206], parents in 8% (n=5) [151, 155, 159, 196, 198], teachers in 8% (n=5) [153, 161, 169, 173, 201], and routine data were used in two (3%) studies [156, 168]. The primary outcome reporter was blind to allocation status in 28% (n=18) of studies. The primary outcome was measured using an objective method in 22% (n=14) of studies included in this review.

Information about the primary outcomes is summarised in Table 3.6.

Table 3.6. Primary outcome characteristics of included studies (N=64)

Characteristic	N	Statistics (n (%))
Primary outcome health area	64	
Social emotional functioning and its influences ¹		15 (23)
Nutrition		10 (16)
Adiposity		7 (11)
Dental health		7 (11)
Physical activity		7 (11)
Smoking		5 (8)
Injury		3 (5)
Sexual health		2 (3)
Obstetrics		2 (3)
Alcohol misuse		2 (3)
Cancer		1 (2)
Communication skills (children with autism)		1 (2)
Gross motor skills		1 (2)
Safety		1 (2)
Type of primary outcome variable	64	
Continuous		42 (66)
Binary		18 (28)
Count		3 (5)
Ordinal		1 (2)
Method of data collection	64	
Questionnaire		31 (48)
Observation		7 (11)
Accelerometer measurement		6 (9)
Anthropometric measurement		6 (9)
Dental assessment		6 (9)
Diaries/recall		6 (9)
Routine data		2 (3)
Main reporter of primary outcome	64	
Pupil		34 (53)
Researcher		12 (19)
Dentist		6 (9)
Teacher		5 (8)
Parent		4 (6)
Routine data		2 (3)
Researcher and parent		1 (2)

¹ Includes mental health, behaviour, hyperactivity/inattention (ADHD), well-being, quality of life, bullying, social and emotional learning, and self-esteem (body image).

Characteristic	N	Statistics (n (%))
Primary outcome reporter blind to allocation status	64	
Yes		18 (28)
No		46 (72)
Primary outcome measurement was objective	64	
Yes		14 (22)
No		50 (78)

3.5.6 Study design

Only 17 (27%) studies provided explicit justification for the use of cluster randomisation. Of those that did, the most common reason was to avoid contamination between individuals in different trial arms (n=13; 76%) [160, 163, 166, 170, 172, 173, 191, 192, 194, 198, 202, 203, 206]. Other justifications provided were that the intervention was delivered at the cluster level (n=5; 30%) [145, 161, 163, 181, 197], logistical reasons (n=2; 12%) [166, 206], and to avoid selection bias (n=1; 6%) [172].

Eighty-eight percent (n=56) of studies randomised schools as the clusters, 6 (9%) studies randomised classes [150, 171, 175, 193, 195, 199], and 2 (3%) randomised year groups [166, 202]. Two reports noted that in order to optimise statistical power, classes were randomised instead of schools, but recognised that this may have led to contamination between trial arms within schools [150, 175].

There was an intermediate level of clustering in 10 (16%) studies. In 7 of these studies, one class was selected from each school cluster [153, 158, 161, 163, 187, 189, 205]. In 2 studies, one class from each of Year 5 and 6 was selected [186, 188], and in 1 study, school rugby teams were the intermediate level of clustering [169].

Sixty-one (95%) studies used a parallel arm design and 3 (5%) used a factorial design [190, 200, 205]. A factorial study is an experimental design that allows researchers to investigate the effects of two or more interventions. In one factorial study, one arm was given a 'wait list' control, one arm was given peer mentoring, one arm was given participative learning and one arm was given both peer mentoring and participative learning [205]. In another factorial study, one arm did

not receive any intervention (control arm), one arm received the Targeted Mental Health in Schools (TaMHS) programme, one arm received educational booklets, and one arm received TaMHS and booklets [200]. In the third factorial study, one arm was given no planned intervention (control group), one arm was given a family smoking education project only, one arm was given smoking and me project only, and one arm was given both projects in sequence [190].

Forty-six (72%) articles provided sufficient information to establish the approach used to sample schools. Of these, 33 studies initially invited all potentially eligible schools to participate, 4 used purposive sampling [150, 155, 180, 195], 3 used convenience sampling [161, 182, 189], 3 used simple random sampling [143, 144, 159], 2 used stratified random sampling (stratified on geographic area) [158, 160], and 1 study used a mixture of random sampling and convenience sampling [190].

The majority of studies had two trial arms (n= 55; 86%). Five (8%) studies had three trial arms [157, 171, 198, 202, 203], and 4 (6%) studies had four trial arms [156, 185, 190, 205]. All studies used a 1:1 allocation ratio except for one study which used a 2:3 ratio [163], which was chosen due to “time and financial constraints”.

Twenty-two (34%) studies specifically stated that there was allocation concealment with respect to the pupils (i.e., pupils did not know which trial arm their cluster was allocated to before they agreed to take part). In 15 (23%) studies, it was stated that there was no concealment of allocation with respect to pupils (i.e., pupils knew which trial arm their cluster was allocated to before recruitment).

A challenge of conducting CRTs is to avoid recruitment bias that might occur if participants are recruited after the clusters are randomised [29, 207]. One third (n=21; 33%) of studies recruited pupils before the clusters were randomised. Only one quarter (n=16; 25%) of studies reported collecting baseline data on pupils before clusters were randomised. This information, however, was unclear in many studies (n=26; 41%).

Most studies (97%) used a cohort design as their method of follow up, where the same pupils provide data at all study waves. One study used a repeated cross-sectional design where different pupils provided data at each wave [188], and one study used an *a priori* mixed design incorporating elements of the cohort and

repeated cross-sectional designs, with only a subset of participating pupils providing data at every wave [186].

Total length of follow-up ranged from 2 weeks [150] to 54 months [168]. The median (IQR) length of follow-up was 12 (6 to 22) months. Half (n=32) of the studies had one follow-up time point, 21 (33%) studies had two follow-ups, 6 (9%) studies had three follow-ups [148, 149, 161, 165, 170, 183], and 5 (8%) studies had four follow-ups [152-154, 189, 193].

In 18 (28%) studies the outcome reporter of the primary outcome was blind to trial arm allocation.

Information about study design characteristics are summarised in Table 3.7.

Table 3.7. Study design characteristics of included studies (N=64)

Characteristic	N	Statistics (n (%))
Justification provided for randomising clusters	64	
Yes		17 (27)
No		47 (73)
Reason for randomising clusters ¹	17	
Avoid contamination		13 (76)
Intervention delivered at the cluster level		5 (30)
Logistical reasons		2 (12)
Avoid selection bias		1 (6)
Unit of randomisation	64	
Schools		56 (88)
Classes		6 (9)
Year groups		2 (3)
Number of trial conditions	64	
Two		55 (86)
Three		5 (8)
Four		4 (6)
Study design	64	
Parallel group		61 (95)
Factorial		3 (5)
Method used to sample schools	46	
All potentially eligible schools invited		33 (72)
Purposive sample ²		4 (9)
Convenience sample ³		3 (7)
Random sample ⁴		3 (7)
Stratified random sample ⁵		2 (4)
Mixed random/convenience sample		1 (2)
Type of randomisation	64	
Completely randomised		13 (20)
Stratified		29 (45)
Matched		8 (13)
Minimisation		8 (13)
Constrained [190, 191]		6 (9)
Type of follow-up	64	
Cohort		62 (97)
Repeated cross-sectional		1 (2)
Mixed		1 (2)

Characteristic	N	Statistics (n (%))
Number of follow-ups	64	
1		32 (50)
2		21 (33)
3		6 (9)
4		5 (8)
Length of follow-up	64	
Up to 6 months		22 (34)
7 to 12 months		19 (30)
13 to 18 months		6 (9)
19 to 24 months		8 (13)
25 to 36 months		7 (11)
More than 36 months		2 (3)
Participating pupils recruited before clusters were randomised	64	
Yes		21 (33)
No		17 (27)
Unclear		26 (41)
Baseline data collected before clusters were randomised	64	
Yes		16 (25)
No		27 (42)
Unclear		21 (33)
Allocation concealment with respect to the pupils	64	
Yes		22 (34)
No		15 (23)
Unclear		27 (42)

¹ Four studies provided two reasons for randomising clusters [163, 166, 172, 206].

² Researchers rely on their own judgement when choosing clusters to participate and when making sure the sample represents certain characteristics of the wider population.

³ A sample taken from clusters easy to contact or to reach.

⁴ Each cluster has a known probability of being chosen (either equal or unequal probabilities).

⁵ The study population is divided into sub-groups (strata) where clusters share common characteristics and then a random selection of clusters is drawn from each strata.

Most studies (n=51; 80%) reported using a form of restricted randomisation to allocate the clusters to trial arms, balancing on selected cluster-level characteristics between trial arms. Of these, 29 (57%) used stratification, 8 (16%) used the matched-pair design [155, 166, 170, 179, 190, 192, 193, 197], 8 (16%) used minimisation [163-165, 173, 174, 186, 191, 196], and 6 (12%) used constrained randomisation [143, 147, 161, 168, 202, 203]. Randomisation was most commonly balanced on a measure of school-level deprivation (61% of the studies that used restricted randomisation). Other factors used to balance the randomisation are described in Table 3.8.

Of the 51 studies that used some form of restricted randomising, only 9 (18%) gave explicit justification for their choice of balancing factors [146, 160, 163, 179, 184, 190, 191, 204, 205]. For example, Bonell and colleagues stated that the factors they chose were key school-level determinants of violence (the primary outcome in their study) [146]. Other authors also chose factors that were strong predictors of the outcomes [163, 184, 191, 204, 205]. Another justification was that 'schools were matched for deprivation and size as it was felt that both these variables could have an impact on the effectiveness of the intervention on the primary outcome, nutrition knowledge' [179]. A further five (10%) studies only *implied* there was justification for their choice of balancing factors [147, 155, 164, 202, 203].

Table 3.8. Cluster-level characteristics used to balance randomisation (N=51)

Characteristic	Statistic (n (%))
Deprivation (school or area in which school is based)	31 (61)
Percentage of pupils eligible for free school meals	21 (41)
Townsend Index [138] ¹	2 (4)
Income Deprivation Affecting Children Index [139] ²	1 (2)
Index of Multiple Deprivation [139] ³	1 (2)
Unspecified ⁴	6 (12)
Cluster size ⁵	23 (45)
Area ⁶	14 (27)
Pupil ethnicity summary	5 (10)
Co-educational status of school	5 (10)
School performance ⁷	5 (10)
School	5 (10)
Baseline measures ⁸	3 (6)
Whether school has existing policy similar to the intervention ⁹	3 (6)
Local sexual health services ¹⁰	2 (4)
Number of classes per school	2 (4)
School type	2 (4)
Other ¹¹	21 (41)

¹ Townsend Index quantifies material deprivation within a population.

² Income Deprivation Affecting Children Index (IDACI) is the proportion of all children aged 0 to 15 living in income deprived families in different local areas across England.

³ Index of Multiple Deprivation (IMD) measures relative deprivation for small areas in England.

⁴ Did not state which measure of deprivation used.

⁵ Includes: Size of school; Size of year group; One versus more than one year-5 class.

⁶ Includes: Local authority area; Geographic area; Health and social care area; Urban vs rural/semi-rural area; Education and Library Board Area; Catchment area; Local health authority; Locality of the school.

⁷ Includes: Student attainment (GCSE); Proportion of pupils staying at school beyond the age of 16; Achievement at key stage 2; Level of educational attainment.

⁸ Includes: Cluster-average baseline moderate-to-vigorous physical activity; body mass index; Teacher reported baseline behaviour problems.

⁹ Includes: Planned road safety improvements during follow up period; School was already participating in “safe routes to school” or other related programmes; Whether the school already had a travel plan; Awarded “healthy schools” or “healthy schools plus” status; Existing policy on snacks at morning break.

¹⁰ Includes: Local sexual health services; Family planning.

¹¹Other balancing factors include: Percentage of students who actively commuted to school; Teaching of UK National Curriculum; Key Stage 1 versus Key Stage 2; Attitude of the school towards health promotion; English-speaking versus Welsh-speaking school; Number of students in year group; Number of year groups per school; Number of mixed-sex classes; Date of entry of school into study; Percentage of children speaking English as an additional language; Whether sex education was taught by a tutor or specialised team of teachers; Whether sex education was taught mainly in Year 9 or in Year 10; Quality and quantity of current school sex education; Percentage of pupils staying on after age 16 years; Special educational need status; School expressed preference for allocation (control versus intervention versus no preference); Health-promoting school status; Percentage of children in year group of interest with no dental decay; Frequency and timetabling of personal, social, and health education lessons; Preferred timetabling of the intervention; Facilitator of the intervention (Regional Project Manager).

3.5.7 Sample size calculation

Fifty (78%) studies accounted for clustering in their sample size calculation. The ICC (n=43; 67%) or between-cluster coefficient of variation (CV) of the outcome [14] (n=3; 5%) assumed in the sample size calculation was reported in 72% (n=46) of studies. Of the 43 studies that provided the ICC assumed in the sample size calculation, 37 randomised schools as the cluster unit, and 6 studies randomised classes [150, 166, 171, 175, 195, 202]. Of those that randomised schools, the median ICC (IQR; range) was 0.05 (0.02 to 0.1; 0.005 to 0.175). Of those that randomised classes, the median ICC (range) was 0.05 (0.025 to 0.1). Of the 3 studies that specified the CV that was used in the sample size calculation, 2 studies assumed it to be 0.2 [169, 204] and 1 assumed it to be 0.25 [179]. The median (range) assumed design effect was 2.21 (1.22 to 8.11) (n=36).

Based on the 46 studies that provided the information, the median (IQR; range) target number of clusters was 30 (20 to 40; 4 to 160). Based on 41 studies, the median (IQR; range) target number of schools was 30 (20 to 42; 4 to 160). Based on 45 studies, the median (IQR; range) target number of individuals was 964 (498 to 2000; 90 to 9000).

Of the studies that had an intermediate level of clustering (n=10), none explicitly stated allowing for this in their sample size calculation [153, 158, 161, 163, 169, 186-189, 205]. Ninety four percent (n=60) of studies did not state whether their sample size calculation allowed for loss of clusters at follow-up; 3 (5%) studies provided sufficient information to indicate that the sample size calculation allowed for loss to follow-up of clusters (i.e., they provided the assumed drop-out percentage at cluster-level) [143, 154, 169]; and it was unclear in 1 (2%) study [167]. Eighteen (28%) studies stated allowing for loss to follow-up of individuals

in the sample size calculation. The median drop-out percentage at individual-level assumed in the sample size calculation was 20%.

Information about the sample size calculations is summarised in Table 3.9.

Table 3.9. Sample size calculation characteristics of included studies (N=64)

Characteristic	N	Statistics
Accounted for clustering in sample size calculation	64	
Yes, n (%)		50 (78)
Assumed school-level ICC of outcome, median (IQR; range)	37	0.05 (0.02 to 0.1; 0.005 to 0.175)
Assumed design effect, median (IQR; range)	36	2.21 (1.98 to 3.53; 1.22 to 8.11)
Power ¹ specified in sample size calculation	64	
80% power, n (%)		30 (47)
90% power, n (%)		17 (27)
85% power, n (%)		3 (5)
81.6% power, n (%)		1 (2)
98% power, n (%)		1 (2)
“100% power” ² , n (%)		1 (2)
Not stated, n (%)		11 (17)
Type 1 error ³ rate specified in sample size calculation	64	
5% level, n (5)		53 (83)
Not stated, n (%)		11 (17)
Study allowed for drop-out at cluster level	64	
Yes ⁴ , n (%)		4 (6)
Not stated, n (%)		60 (94)
Study allowed for drop-out at individual level ⁵	62	
Yes, n (%)		18 (29)
Not stated, n (%)		44 (71)
Target number of clusters, median (IQR; range)	46	30 (20 to 40; 4 to 160)

Characteristic	N	Statistics
Target number of schools, median (IQR; range)	41	30 (20 to 42; 4 to 160)
Target number of individuals, median (IQR; range) ⁶	45	964 (498 to 2000; 90 to 9000)

¹ Power is the likelihood of a significance test detecting an effect when there actually is one.

² Although the study reported this it is not possible to have 100% power.

³ Type 1 error rate is the probability of rejecting the null hypothesis given that it is true. In other words, the probability of a false positive result.

⁴Unclear in one study but enough information to assumed that authors did allow for drop-out at cluster level.

⁵Summary excludes the two studies that did not use the cohort design.

⁶Summary excludes the two studies that did not use the cohort design.

3.5.8 Ethics and consent procedures

Ethical approval was granted for 57 (89%) studies. Six (9%) did not state whether ethical approval had been granted [144, 177, 186, 190, 194, 206]. One (2%) study [179] did not receive ethical committee approval because, “... *the study assessed a new curriculum and change in nutrition knowledge with no identifiable data or anthropometry measurements, ethical approval was not required. The head teachers of participating schools filled out a reply slip and gave consent for participation. No individual student or teacher consent was obtained.*”

Information regarding consent procedures and ethical approval was often not well reported. Consent for participation in the trial at the level of the cluster (e.g. headteacher/administrator/gatekeeper) was often implied rather than detailed. Sixty three percent (n=40) of studies explained that consent (permission for something to happen or agreement to do something) and/or assent (the expression of approval or agreement, often verbal) was sought from *both* parents/carers and their child for participation. Just under half (n=29; 45%) of studies reported that passive ‘opt-out’ consent [82] was used for participation in the study from either the parent/carer and/or pupil. Eight (13%) studies explicitly mentioned that harm/adverse events were recorded during the study.

Information about the ethics and consent procedures are summarised in Table 3.10.

Table 3.10. Ethics and consent characteristics of included studies (N = 64)

Characteristic	N	Statistics (n (%))
From whom was consent/assent sought for pupil participation?	64	
Parent/carer and pupil		40 (63)
Parent/carer only		15 (23)
Pupil only		2 (3)
No/Not stated		7 (11)
Opt-out consent/assent used for parent/carer and/or pupil	64	
Yes		29 (45)
No/Not stated		35 (55)

3.5.9 Analysis methods

Nearly three quarters of studies (n=46; 72%) analysed their data using individual-level analysis methods that allow for clustering (e.g., mixed effects models). Cluster-level analysis methods were used in 10 (16%) studies. Eight (12%) studies did not allow for clustering in their analysis.

Fifty-two (82%) studies adjusted for cluster-level and/or individual-level factors in their analysis. Twenty-seven (42%) studies adjusted for cluster-level factors in their analysis, most commonly a measure of deprivation (n=17). Forty-five (70%) studies adjusted for individual-level factors in their analysis, the most commonly being baseline measure of the outcome (n=35). Other cluster-level and individual-level characteristics adjusted for in the analysis of included studies are described in Appendix 7.

Forty-three (67%) studies stated that they used an intention-to-treat analysis to test the intervention effect. Only four (6%) studies reported using multiple imputation to handle missing data in their main analysis. Just over half (n=35; 55%) the studies undertook a subgroup analysis.

Information about the analysis methods are summarised in Table 3.11.

Table 3.11. Analysis methods characteristics of included studies (N=64)

Characteristic	N	Statistics (n (%))
Method of analysis	64	
Individual-level analysis that allows for clustering		46 (72)
Cluster-level analysis		10 (16)
Did not allow for clustering		8 (12)
Adjusted analysis ¹	64	
Yes, for cluster-level characteristics		27 (42)
Yes, for individual-level characteristics		45 (70)
No		12 (19)
Intention-to-treat analysis	64	
Yes		43 (67)
Subgroup analysis	64	
Yes		35 (55)

¹ Some studies adjusted for *both* cluster-level and individual-level characteristics.

3.5.10 Other areas of methodological interest

A median (IQR; range) of 31.5 (21 to 50; 4 to 486) clusters, 29 (15 to 50; 4 to 486) schools and 1308 (604 to 3201; 17 to 27435) pupils were recruited to the studies included in this review. One study recruited only 17 pupils [182]. The median (IQR; range) number of pupils per cluster was 38.9 (20.9 to 99.8; 4.9 to 327.9) and the median (IQR; range) number of pupils per school was 45.4 (25.9 to 116.0; 5.9 to 327.9). The CRTs with a cohort design that reported targeted and achieved recruitment figures at the cluster (n=45) and pupil (n=43) levels achieved those targets in 89% and 77% of studies, respectively.

Some authors noted challenges with recruitment of the clusters [148, 157, 196]. For example, Rowland and colleagues noted that, '*only half of the schools invited to participate took part. Most declined because they were too busy and were reluctant to take on the extra responsibility of school travel*' [196]. Breslin and colleagues similarly discussed that '*logistics and finite resources*' made the recruitment of schools challenging [148]. Diedrichs and colleagues commented

that a '*crowded timetable*' was a barrier to school's participation in their study [157].

Challenges with the recruitment of pupils were also reported [182, 183]. Markham and colleagues attributed the reason for failure to recruit pupils to, '*young people's lack of interest*' [182]. This study had such significant issues with recruitment that it was halted prematurely. Another study by McKay also highlighted the need for better understanding of barriers and facilitators to recruitment and stated, '*Research is needed to assess the relative efficacy of recruitment strategies such as incentives, mass media campaigns, the removal of barriers to attendance (e.g., providing transport and childcare) and the use of key community recruiters (influential individuals and organisations)*' [183]. Several authors commented on the challenges of blinding trial arm status [160, 169, 170, 198]).

Thirty out of 62 (48%) studies that provided information reported that at least one cluster was lost to follow-up. Of the studies that lost clusters, the median (IQR; range) percentage of clusters lost to follow-up was 6.5% (3.7% to 11.7%; 0.5% to 39.5%). Missing data resulting from entire schools dropping out was highlighted as a problem in some reports (for example, [162, 199, 204]). Stephenson and colleagues stated that: "*The withdrawal of one school had the biggest effect on missing data*" [204].

The median follow-up at pupil level was 79.9%. Of the 55 (86%) studies that reported information on loss to follow-up for pupils, the median (IQR; range) percentage of pupils missing was 21.9% (14.2% to 36.6%; 0.5% to 92.3%).

All but one study [156] reported baseline demographic information. Based on the 33 studies that provided data, the median (IQR) percentage of pupils categorised as "White" was 76.8% (51.5% to 86.2%). Fifty-five studies reported the percentage of female pupils at cluster level; the median (IQR) percentage was 49% (47.5% to 52.5%).

Thirty eight percent (n=24) of studies reported a p-value less than 0.05 for the primary analysis. Seven (11%) studies provided information suggesting that some pupils changed cluster membership during the course of the study [144, 149, 152, 153, 159, 165, 180]. Three (5%) studies reported harms/adverse events during their study [146, 180, 184].

In total, 29 (45%) studies reported the intra-cluster correlation coefficient (ICC) from the analysis of the primary outcome. Eighteen (49%) of the 37 studies published after 2010 reported the ICC. Of the 29 studies that reported the ICC, 5 (17%) also reported 95% confidence intervals. The median (range) ICC for studies in which schools were the cluster was 0.039 (0.0005 to 0.21).

Information regarding the other areas of methodological interest, including the ICC estimates, are summarised in Table 3.12. The ICCs reported are summarised in Table 3.13.

Table 3.12. Other areas of methodological interest from included studies (N=64)

Characteristic	N	Statistics Median (IQR; range)
Ethnicity: percentage of White pupils ¹	33	76.8 (51.5 to 86.2; 24 to 95.3)
Gender: percentage of female pupils ²	55	49 (47.5 to 52.5; 0 to 100)
Total number of clusters recruited	62	31.5 (21 to 50; 4 to 486)
Total number of schools recruited	63	29 (15 to 50; 4 to 486)
Total number of pupils recruited ³	60	1308 (604 to 3201; 17 to 27435)
Number of pupils per cluster	60	38.9 (20.9 to 99.8; 4.9 to 327.9)
Number of pupils per school	60	45.4 (25.9 to 116.0; 5.9 to 327.9)
Percentage of clusters followed-up for primary outcome	62	100 (92.5 to 100; 60.5 to 100)
Percentage of pupils followed-up for primary outcome ³	58	79.9 (64.1 to 87.5; 7.7 to 100)
Observed school-level ICC of primary outcome ³	26	0.039 (0.017 to 0.12; 0.0005 to 0.21)

¹ Hodgkinson and colleagues [170] - ethnicity was based on adults. Sharpe and colleagues [200] - ethnicity was based on sub-sample of pupils from primary schools. Diedrichs and colleagues [157] - ethnicity was calculated as an average across genders in the control arm.

²Sharpe et al [200] - the percentage of female pupils was based on sub-sample of pupils from primary schools.

³ Summary excludes the two studies that did not use the cohort design.

Table 3.13. Estimated intra-cluster correlation coefficients (ICCs) for primary outcomes (N=29)

Author	Year	Cluster unit	Outcome	Health area	Outcome type	ICC estimate (95% CI)
Stallard [202]	2012	Year groups	Symptoms of low mood (depression)	Social emotional functioning and its influences	Continuous	0.012 (<0.001 to 0.039)
Chisholm [150]	2016	Classes	Stigma of mental illness	Social emotional functioning and its influences	Continuous	0.1 (0.04 to 0.26)
Obsuth [191]	2017	Schools	School exclusion	Social emotional functioning and its influences	Binary	0.028
Connolly [153]	2018	Schools	Prosocial behaviour (Strengths and Difficulties Questionnaire (SDQ))	Social emotional functioning and its influences	Continuous	0.116
Ford [161]	2019	Schools	Total difficulties (Strengths and Difficulties Questionnaire (SDQ))	Social emotional functioning and its influences	Continuous	0.121
Axford [145]	2020	Schools	Victimisation (being bullied) (occurring at least twice a month in the last 2 months)	Social emotional functioning and its influences	Binary	0.019
Campbell ¹ [149]	2008	Schools	Smoking in the past week	Smoking	Binary	0.017 (0.004 to 0.029)
Conner [152]	2019	Schools	Ever smoking	Smoking	Binary	0.017

¹ ICC for control arm only

Author	Year	Cluster unit	Outcome	Health area	Outcome type	ICC estimate (95% CI)
McKay [183]	2018	Schools	Heavy episodic drinking in the previous 30-days (≥ 6 units for males and ≥ 4.5 units for females)	Alcohol misuse	Binary	0.121
Croker [155]	2012	Schools	Child's eating habits	Adiposity	Continuous	0.07
Fairclough [160]	2013	Schools	Waist circumference (cm)	Adiposity	Continuous	0.06
Hodgkinson ² [170]	2019	Schools	BMI z score	Adiposity	Continuous	0.0396
Lloyd [180]	2018	Schools	BMI z score	Adiposity	Continuous	0.014 (0.003 to 0.069)
Breheny [147]	2020	Schools	BMI z-score	Adiposity	Continuous	0.001
Jago [174]	2015	Schools	Moderate-to-Vigorous Physical Activity (mins/weekday)	Physical activity	Continuous	0.0005
Harrington [167]	2018	Schools	Moderate-to-Vigorous Physical Activity (mins/day)	Physical activity	Continuous	0.02
Tymms [205]	2016	Schools	Moderate-to-Vigorous Physical Activity (mins/day)	Physical activity	Continuous	0.19
Norris [189]	2018	Schools	Sedentary behaviour during the school day (mins)	Physical activity	Continuous	0.080

² ICC for control arm only

Author	Year	Cluster unit	Outcome	Health area	Outcome type	ICC estimate (95% CI)
James ³ [175]	2004	Classes	Consumption of carbonated drinks over 3-days (in glasses)	Nutrition	Continuous	-0.009 (- 0.03 to 0.05)
Christian [151]	2014	Schools	Combined daily fruit and vegetable intake (grams/day)	Nutrition	Continuous	0.003
Redmond [194]	1999	Schools	Proportion of teeth sites with caries at 6 months	Dental health	Continuous	0.16
Worthington [206]	2001	Schools	Plaque score	Dental health	Continuous	0.023
Milsom [185]	2006	Schools	Active caries in first permanent molars	Dental health	Binary	0.027
Kendrick [177]	2004	Schools	Ownership of cycling helmet	Injury	Binary	0.09
Mulvaney [187]	2006	Schools	Use of visibility aid (reflective and fluorescent slap wrap) while cycling	Injury	Binary	0.21
Kendrick [176]	2007	Schools	Knowledge score for fire and burn prevention	Safety	Continuous	0.187
Hubbard [172]	2016	Schools	Number of recognised cancer warning signs	Cancer	Continuous	0.038

³ The ICC in James (2004) was negative. True negative values are generally considered implausible in the context of cluster randomised trials.

Author	Year	Cluster unit	Outcome	Health area	Outcome type	ICC estimate (95% CI)
Henderson [168]	2007	Schools	Terminations of pregnancy by age 20	Obstetrics	Count	0.005
Giles [163]	2014	Schools	Intention to breastfeed	Obstetrics	Continuous	0.12

Three studies that randomised classes as clusters reported the ICC estimate for the primary outcome. The ICC values were -0.009 [175], 0.012 [202], and 0.1 [205]. For many studies that reported both values there was a marked difference between the assumed value of the school-level ICC in the sample size calculation and the ICC estimated for the primary outcome from the study data (Figure 3.3).

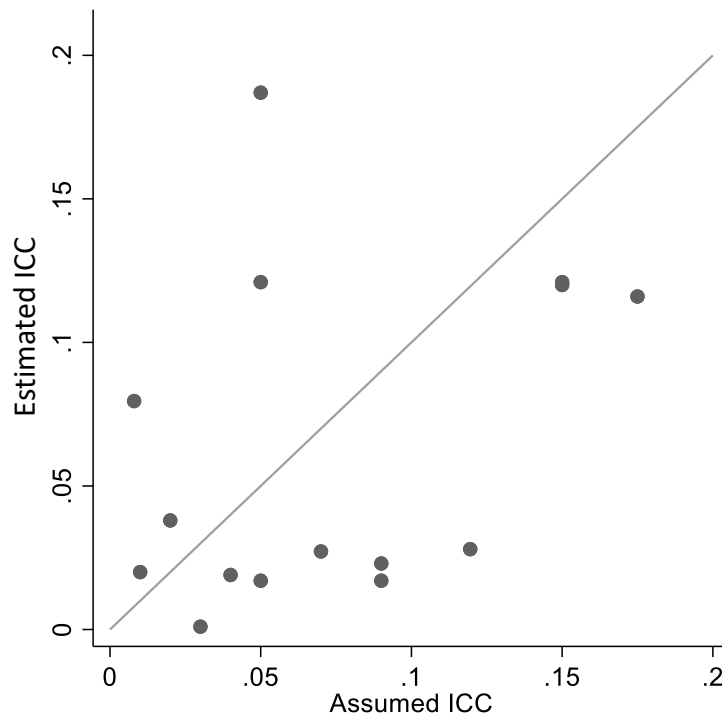


Fig. 3.3. Estimated intra-cluster correlation coefficient (ICC) for primary outcome versus ICC assumed in sample size calculation (N=15)

The median (range) of the differences between the estimated ICC and the ICC assumed in the sample size calculation, calculated as estimated minus assumed, was -0.029 (-0.092 to 0.137). This indicates that, on average, the estimated ICC was smaller than the assumed ICC. The most extreme example of this was one study which had an estimated ICC 0.092 smaller than the assumed value [151]. At the other extreme, in one study the estimated ICC was 0.137 larger than the assumed value [187]. The intra-class correlation coefficient of agreement between the estimated and assumed ICCs was 0.38, indicating poor agreement.

Of the 7 studies [145, 149, 152, 183, 185, 187, 191] that reported ICCs estimated for a binary primary outcome, none stated whether the ICC was calculated on the proportions scale or the logistic scale [10]. Of these studies, 5 [149, 152, 183, 187, 191] used mixed effects models [23] to analyse the data, and it could be

assumed that they reported the ICC on the logistic scale. This could potentially account for some of the differences between the estimated and assumed ICCs. However, further examination of these ICCs showed clear differences for only two of these studies; 0.21 for the observed ICC versus 0.05 for the assumed ICC in Mulvaney et al [187], and 0.028 versus 0.1, respectively, in Obsuth et al [191].

Of the 29 studies that reported the estimated ICCs, unadjusted analyses were used to calculate the ICC in almost half of studies (n=14; 48%). Analyses adjusted for potential prognostic factors were used to estimate the ICCs in 6 (21%) studies [160, 167, 174, 185, 187, 189]. One (3%) study estimated the ICC separately from the main analysis [175]. In the remaining 8 (28%) studies it was unclear whether the ICC was estimated from an adjusted or unadjusted analysis [153, 155, 172, 176, 177, 180, 194, 202].

3.6 Discussion

This is the first systematic review to summarise the methodological practices and characteristics of school-based CRTs used to evaluate interventions for improving health outcomes on pupils in the UK. This section summarises the results and discusses the findings of this systematic review. This is followed by an assessment of the overall strengths and limitations of the systematic review, implications for the planning and conduct of future CRTs and areas in need of further research.

The rate of publication of school-based CRTs indexed in MEDLINE has increased since the first UK study was published in 1993 [190]. The review identified a specific increase since the publication of the CONSORT-CRT extension in 2004, which has been noted by others [53]. This may partly be due to better reporting and subsequent detection of studies using the established search strategy [127]. The CRT design lends itself well to the school setting [50], and the increase in publications also reflects the design's increasing popularity in school-based health research and the growing recognition of the role that schools can play in improving the health of children and young people [53, 115, 208]. This is apparent in health areas, such as physical activity, nutrition and social emotional functioning and its influences, where publications have increased markedly over the last 10 years. The increase in such studies may be due to the UK Government viewing schools as central to tackling issues, such as the obesity crisis, because

they are an ideal setting in which to actively engage children and their families across the socio-economic spectrum [46, 208].

Of the studies included in this systematic review, 78% reported sample size calculations that accounted for clustering. This is a higher percentage than systematic reviews of CRTs examining interventions for improving health outcomes in primary care settings [110, 112, 113, 209]. Of the 46 articles that reported the level of clustering assumed in the sample size calculation, almost all used the ICC to calculate the design effect (DE) to inflate their sample size calculation. This highlights the importance of reporting ICC estimates for use in future studies. On average, based on the DE in the sample size calculation, the studies in this review required just over twice as many pupils as would have been needed if individuals had been randomised rather than clusters. This empirical knowledge may be useful to make informed adjustments to precision estimates in meta-analyses of intervention effects, where the included studies have not allowed for clustering.

Only three (5%) studies allowed for loss to follow-up of clusters in their sample size calculation. This is far lower than a previous review that found 38% of CRTs reported allowing for cluster-level attrition in their sample size calculation [209]. Thirty studies in the current systematic review reported that at least one cluster was lost to follow-up, demonstrating the need to allow for cluster-level attrition in the sample size calculation, despite very few studies reporting actually doing so. Additionally, missing data resulting from entire school drop-out was highlighted as a problem by authors in 3 studies in this review [162, 199, 204]. Cluster-level attrition was higher in this systematic review than reported in a previous systematic review examining how missing data are handled in CRTs in primary care [210] (48% versus 18%, respectively).

This systematic review found the estimated ICCs from the study data often differed greatly from the ICC assumed in the sample size calculation. This will partially be due to sampling variation and that some studies adjusted for prognostic factors in the analysis. It could, however, also reflect the lack of availability of relevant and accurate ICC estimates at the time of sample size calculation. The current review found that less than half (45%) of studies reported the ICC for the primary outcome. This improves to 55% (18/33) for studies

published after 2012, an increase that may be attributed to dissemination of methodological guidance such as the CONSORT extension to CRTs [57].

The median estimated school-level ICC in this review was 0.039. ICCs for health outcomes in studies that randomise general practices are generally less than 0.05 [20]. In the current systematic review, just over half (54%) of school-level ICC estimates had values that were less than 0.05. Based on previous publication school-based ICC estimates for educational outcomes tend to be larger than for health outcomes, generally larger than 0.1 [88, 90, 96, 211]. This is to be expected as schools, by design, have the main function of educating rather than improving health [71, 88]. Several studies have provided lists of estimates of ICCs from school-based CRTs or surveys but tend to focus on specific health areas such as substance use [61, 71, 97-104], nutrition [105-107], physical activity [61, 107-109], and mental health and behaviour [61, 69, 96], and the majority summarise studies from the US. The distribution of school-level ICCs from UK school-based CRTs pupil health outcomes was found to be broadly similar to the previously reported summaries worldwide [69, 96-103, 105-109].

Only 5 of the studies that reported the school-level ICC estimates reported 95% CIs with these. Summarising the precision of ICC estimates is important as it provides a plausible range of values to help researchers to make an informed choice of ICC value for use in their sample size calculations. The marked differences for some studies in this review between the ICC assumed in the sample size calculation and the ICC estimated from the analysis demonstrates the need for better information on ICCs in this context. Furthermore, more research is needed in this area to understand the factors that influence the size of ICCs in school-based health research as previous research has found that school-level factors such as low socio-economic status and low academic achievement can have an impact on the size of ICCs [88].

Representativeness of cluster-level and individual-level characteristics in CRTs is important in order to improve external validity and inclusiveness. The majority of studies recruited schools from only one or two geographic regions, and only recruited one type of school (e.g., state schools). There was little information provided in the papers to assess how representative the study schools were of the general population, and little detail on aspects of the recruitment strategy. There was a lack of information on the characteristics of schools that declined to

participate. This systematic review, however, found the median percentage of pupils from a minority ethnic background was around 10% lower than the national average [212].

Challenges of recruitment were identified in this systematic review and have been noted in the wider literature [33, 74-76]. Just over half of the studies administered the intervention components as classroom lessons, often with teachers delivering the intervention. In addition, teachers were required to report primary outcomes on the pupils in five studies in this review. These activities require additional resources and time. The burden to teachers and schools may be a barrier to participation in such studies and result in lack of representation of certain types of schools. The fact that no sixth forms and colleges were eligible to participate in any of the studies in this review is perhaps due to the burden on timetabling during this stage in education.

In this systematic review, 80% of studies used some form of restricted randomisation to balance cluster-level characteristics, which is in keeping with recommendations from the methodological literature to use restricted randomisation [4, 6, 213]. The review found greater use of restricted randomisation than previously seen in other reviews of CRTs in primary care settings [83, 110, 112, 113, 209, 214], stating around half of studies used restricted randomisation [83, 113]. The current review also found that school-level deprivation was the factor that was most commonly used to balance the randomisation, particularly the percentage of children in the school that were eligible for free school meals. This may be in part due to this information being readily available from the UK Department for Education [215]. However, little is known about which cluster-level characteristics are prognostic for pupil health outcomes and few studies gave reasons for their choice of balancing factors. The best candidates for the balancing randomisation will be school-level characteristics that are predictive of the study outcomes, account for within-cluster correlation, or influence effectiveness of the intervention [3, 216].

3.7 Strengths and limitations

This systematic review is the first to describe the characteristics and practices of school-based CRTs of interventions for improving health outcomes on pupils in the UK. The review used a clearly defined search strategy in order to identify

school-based CRTs and make the search reproducible. The strategy built on an existing search strategy tailored specifically to identify CRTs [127]. The screening procedures and data extraction were undertaken by two independent reviewers, helping to minimise errors and increase accuracy and reliability.

The review included papers spanning a wide range of different interventions investigating many different health conditions/areas. This not only adds originality to this review but provides a broad coverage of public health areas of interest in both the UK and worldwide.

The review focussed on CRTs undertaken in the UK, resulting in rich data on CRT methodology in a single education system. As a result, the findings are readily applicable to this specific context. The findings of this review will still be of global interest as some countries, such as Australia, have a similar school system to the UK, and many of the findings may be applicable in those settings. The decision to focus the review on the UK was also a pragmatic one. There was an abundance of international literature in the field, and the time and resources for undertaking the review were limited.

The systematic review has some limitations. Due to time constraints and to make the review more focused, a pragmatic decision was made to restrict the search to one electronic database which focused on health interventions. Further scoping searches of EMBASE, DARE, PsycINFO and ERIC databases, however, only identified one additional eligible school-based CRT that was not identified by the MEDLINE search. Grey literature was not included due to time constraints, but it was also unlikely that any school-based CRTs would not be indexed in databases due to the incurred cost and time in delivering these studies.

Meta-analysis was not used in the review to pool ICCs, as a single pooled value from a meta-analysis would not be meaningful and would not take into account the methodological nuances of each study. It is more useful to summarise the variability in the estimated ICCs as this provides a plausible range of values which can then be used to inform sample size calculations in future CRTs [121]. Furthermore, articles included in this review were diverse in terms of methodology and health area and, therefore, there was no single underlying ICC as each scenario was different.

3.8 Implications

This chapter has identified key implications for school-based CRT methodology. The results provide a comprehensive summary of the common methodological characteristics and challenges faced by researchers conducting school-based CRTs in the UK. The increase in the number of published school-based CRTs over recent years provides a pool of knowledge to aid the design and conduct of future studies and identifies areas where improvements can be made.

There is evidence that the assumed ICC of the outcome in the sample size calculation is often quite different from that observed in the study data, which mirrors other settings where there have been renewed calls for better reporting of ICCs [53, 110-112]. The review has highlighted the need for better reporting of ICCs for health outcomes in order to establish plausible values to assume for sample size calculations for CRTs in the school setting and avoid poor estimates of the number of schools and pupils required in such studies.

More research is needed regarding whether the size of the ICC differs across disease areas and whether ICCs are transferable across different countries and contexts in school-based CRTs. There is a lack of published ICC estimates relevant to school-based CRTs in the UK. The dissemination of ICCs based on school-related clusters would greatly aid the planning of future school-based CRTs. Additionally, further work is needed to replicate this systematic review outside of the UK to see if the results are similar and to explore patterns in design features of school-based CRTs and ICCs across different world regions.

Given the high number of school-based CRTs that use some form of restricted randomisation to balance cluster-level factors across the trial arms, it would be informative to examine the strength of association between school characteristics and specific pupil health outcomes and the extent to which those characteristics account for the between-cluster variation. This knowledge would aid researchers to identify the best candidates on which to balance the randomisation of school-related clusters in CRTs.

The review also identified difficulties in obtaining representative samples of schools in school-based CRTs. Recruitment was often limited to one or two geographical regions and one school type. A representative sample of schools

will improve the generalisability of the findings to a wider range of schools in the study population. More research is needed, as there is inadequate knowledge regarding the barriers to successful recruitment of a representative sample of schools despite this being noted as a common challenge in the literature [3, 4, 81, 183]. Identifying these challenges will assist researchers to improve the diversity of their recruitment.

The findings of this systematic review will also help to inform the design of simulation studies for evaluating the properties of statistical methods for calculating sample size and analysing data from school-based CRTs.

3.9 Conclusions

Recent years have seen an increase in the rate of publication of school-based CRTs examining the impact of health interventions on pupils in the UK. The results of this systematic review provide researchers with data on relevant parameters to inform simulation-based studies for evaluating the performance of statistical methods in scenarios typical of school-based studies. The review illustrates key methodological challenges faced when undertaking school-based CRTs in the UK and will help future researchers to better plan for these challenges. The review provides ICC estimates for use in the sample size calculation of similar future school-based CRTs in the UK, but also highlights the need for more information on the ICCs to enable better of such studies. Better reporting of the recruitment process in CRTs will help to identify common barriers to obtaining representative samples of schools. Finally, previous school-based CRTs may provide useful sources of data to identify the school-level characteristics that are strong predictors of pupil health outcomes and that, particularly, account for the variation across schools in those outcomes. Such characteristics would, therefore, be potentially good factors on which to balance the randomisation and adjust for in the analysis of the intervention effect.

3.10 Chapter summary

This chapter presented findings from a systematic review of the characteristics and methodological practices of UK school-based CRTs used to evaluate interventions for improving health outcomes on school pupils. Following this

review of definitive studies, Chapter 4 will examine the characteristics and methodological practices from school-based feasibility CRTs.

Page intentionally left blank

Chapter 4: Systematic review of the characteristics of school-based feasibility cluster randomised trials of interventions for improving the health outcomes of pupils in the United Kingdom

4.1 Summary

This chapter presents background information specific to the methodological challenges of designing and conducting feasibility studies with a CRT design in schools in the UK. The chapter then outlines the aims and objectives and describes the methods and results of a systematic review examining the methodological characteristics of such studies. The chapter then concludes by discussing the results, strengths and limitations, implications and areas identified for further research. A peer-reviewed journal article has been published of the systematic review [122] (Appendix 3). The entire systematic review is reported here in detail.

4.2 Background

Prior to a definitive trial, a feasibility study may be used to determine whether the research is something that can be done, whether it should be done and how it should be done [43]. Feasibility studies focus on areas of uncertainty in trial delivery, such as: the randomisation process, recruitment and follow-up rates, acceptability to the participants of the trial processes and the intervention itself, implementation of the intervention, data collection processes, selection of outcome measures, potential harms related to the intervention and trial, knowledge of parameters that inform the sample size calculation for the definitive trial, and potential effectiveness of the intervention. The randomised pilot trial is a type of feasibility study that involves conducting a smaller version, or part of the future definitive trial [43]. Feasibility studies may also use a single-arm or non-randomised parallel group design, which can be used to develop interventions and trial methods, and test them prior to a full-scale trial [43, 217]. However, these designs are unable to test uncertainties related to the randomisation process, such as participants' willingness to be randomised.

Feasibility CRTs differ from those conducted in advance of individually RCTs in that they may be used to address concerns that are specific to the CRT design. These include challenges such as, evaluating the possibility for recruitment bias in studies where clusters are randomised before individual participants are recruited [29], and obtaining estimates of the intra-cluster correlation coefficient (ICC) of the primary outcome to inform the calculation of the sample size for the definitive trial. Other general feasibility considerations apply at both the cluster and individual levels, such as ease of recruitment, rate of loss to follow-up, and acceptability of the intervention. Methodological considerations that are unique to the conduct of feasibility CRTs include the need to take account of clustering when calculating the sample size for and reporting precision in feasibility parameter estimates from such studies [44].

In recent years, CRTs have been increasingly used to evaluate interventions for improving educational outcomes in schools [59] and complex interventions for improving child health outcomes [50, 53, 117]. Schools provide a natural environment in which to recruit and deliver public health interventions to children due to the amount of time they spend there [50]. The CRT design is suited to the natural clustered structure found in schools (pupils within classes, within schools), but there are challenges in delivering trials in this setting that mean feasibility studies are essential ahead of a definitive trial. For example, schools and teachers often have stretched and limited resources, and implementing an intervention and participating in a trial can be challenging, given that the primary focus of schools is the education of pupils. The systematic review of definitive school-based CRTs, described in Chapter 3, found that 52% of the studies required a member of school staff to deliver components of the intervention [117]. Feasibility trials could be used to explore issues regarding which type of cluster to randomise in the school setting for a given trial. For example, there may be a choice between randomising schools and randomising classes. Randomising schools is better for minimising the chance of contamination between trial arms, as individual pupils will interact across classes within schools. However, randomising classes would have the advantage of a smaller design effect and, therefore, greater power for a fixed total number of recruited pupils compared with schools [218]. Compared with other settings such as primary care, CRTs for evaluating health interventions have only relatively recently been used in schools

in the UK and, therefore, there is a smaller pool of experience available from previous studies to draw from [50, 117]. Given these uncertainties, feasibility trials have an important role to play in the design and execution of definitive school-based CRTs.

Authors have previously discussed the growing literature described as ‘feasibility’ or ‘pilot’ studies, and the associated methodological challenges [43]. The characteristics of feasibility studies generally [44, 219, 220] and cluster randomised feasibility studies specifically [221, 222] have been summarised. However, to date, no systematic review has focussed on the characteristics of school-based feasibility CRTs of interventions for improving pupil health outcomes. By summarising the design features of school-based feasibility CRTs, the results of this systematic review will identify areas for improvement in the conduct of such studies. Through reporting the feasibility objectives of the included studies, this review will help to identify aspects in which better use of feasibility studies could be made to explore uncertainties specific to the CRT design.

4.3 Aims and objectives

The aim of this systematic review was to report the key design features, methodological characteristics and feasibility-related objectives of school-based feasibility CRTs measuring pupil health outcomes in the UK.

The objectives were to:

- Describe the methodological characteristics and challenges of school-based feasibility studies with a CRT design in the UK measuring health outcome on pupils.
- Describe the feasibility-related objectives of school-based feasibility CRTs in the UK that measure health outcomes on pupils.

4.4 Methods

The systematic review has been reported in accordance with the PRISMA statement [125]. The review protocol was registered on PROSPERO (Registration number: CRD42020218993), an international register for systematic reviews.

4.4.1 Search strategy

The search strategy used in the systematic review was identical to the one previously used in Chapter 3 to find definitive trials, as described in Section 3.3.1. A brief summary of the search strategy is provided here for clarity.

Peer-reviewed school-based feasibility CRTs, indexed on MEDLINE (via Ovid), were the source of data for the review. MEDLINE was systematically searched from inception to 31st December 2020. The search strategy (Table 3.1) was developed using terms from the MEDLINE search strategy by Taljaard et al [127] to identify CRTs, and this was combined with *school* concept terms, including the 'Schools' MeSH term. The search was limited to the English language.

4.4.2 Eligibility criteria

The systematic review included school-based feasibility CRTs that measured health outcomes on pupils and were conducted in the UK. The population of included studies was pupils attending pre-school, primary school, secondary school, sixth form or college settings in the UK. 'Pre-school' was defined as an organisation offering early childhood education (e.g., pre-school, nursery school and kindergarten) prior to the child beginning compulsory (primary school) education [130].

Eligible clusters were any school-related unit (e.g., schools, classes, year groups). Studies that randomised school-related units as well as other types of clusters (e.g., towns, households) were only eligible for inclusion in the review if results of the study were shown separately for the school-related clusters (i.e., the authors did not pool results across different cluster types).

Any health-related interventions were eligible. The primary outcome had to be measured on pupils and be health related. Studies with education-related primary

outcomes were excluded. All types of CRT design were eligible, including parallel arm, crossover, factorial and stepped-wedge trials.

Only randomised, external feasibility studies (i.e., where outcome data from the feasibility study are not included as part of a main definitive trial analysis [223]) were included in this systematic review. The definition of feasibility study used to identify eligible papers was that used by Eldridge and colleagues [43], which states: “A feasibility study asks whether something can be done, should we proceed with it, and if so, how”. Thus, eligible studies had to be assessing some element of feasibility in the intervention and/or trial methodology, ahead of a definitive trial. This was determined by looking for the terms, ‘pilot’, ‘feasibility’ or ‘exploratory’ in the title and abstract, and by examining the aims and objectives of each study. Internal pilot studies that are part of definitive trials, where the data from the pilot phase are included in the main analysis [224], were excluded. Non-randomised feasibility studies and single-arm feasibility studies were excluded as randomisation was one of the aspects that the systematic review aimed to examine.

If there was more than one publication of the results for an eligible feasibility CRT, the first paper presenting quantitative results related to the feasibility objectives was designated the key study report (index paper) and used for data extraction. Articles that did not report the results of the feasibility objectives were excluded along with articles only reporting protocol/design information, cost-effectiveness/economic evaluations, and process evaluations.

4.4.3 Screening and selection

After the search strategy was run in MEDLINE, all titles and abstracts were downloaded into Endnote X9 [133]. Duplicate citations were removed, and remaining titles and abstracts were screened by two independent reviewers (KP and OU) for eligibility against inclusion criteria. Citations were coded (1) if they were thought to be eligible, or (2) if not. Once coded, the two Endnote libraries were merged to identify citations where coding differed between reviewers. Articles for which inclusion status was uncertain, and consensus could not be reached through discussion, were included for full text evaluation.

A new Endnote library was created with all potentially eligible articles. PDF versions of each article were obtained for full text screening. Full text screening was first piloted on a random sample of 10 articles. Endnote was used to code each article with a letter to indicate the reason for inclusion/exclusion. The reasons for exclusion are listed in Table 4.1.

Table 4.1. Reasons for exclusion at full text screening

Code	Reason
a	Include
b	Exclude - Not randomised
c	Exclude - Not a CRT
d	Exclude - Not undertaken in the UK
e	Exclude - Not school-based/school unit not randomised
f	Exclude - Primary outcome not reported on pupils
g	Exclude - Sibling paper of index studies
h	Exclude - Not feasibility studies
i	Exclude - Protocol/Design
j	Exclude - Cost-effectiveness/economic evaluation
k	Exclude - Process evaluation

In parallel, two independent reviewers (KP and SEd) screened articles for eligibility based on the inclusion criteria using this coding method. Once all texts had been screened and coded, the two Endnote libraries were merged to identify the articles where coding differed. Any disagreements that could not be resolved through discussion were sent to a third reviewer (OU) for a decision.

4.4.4 Data extraction

Prior to data extraction, each article was assigned a unique study ID number. The study characteristics (variables) to extract were chosen after consultation with experts in the field and examining similar methodological systematic reviews [53, 117, 221]. A bespoke data extraction form was developed using Microsoft Excel and the data extraction piloted using 5 eligible studies. Modifications were made following the pilot, and the final list of variables extracted is presented in Table 4.2.

Table 4.2. Data extracted from school-based feasibility CRTs

Variable category	Variable
<i>Publication details</i>	First author's name; Year of publication; Journal name; Main funding source; Trial registration status (yes – prospectively, yes – retrospectively, no).
<i>Setting and participant characteristics</i>	Country (e.g., England); School level (e.g., primary school); School type (e.g., state school, faith school); Co-educational status (co-ed, female only, male only); Age(s) of pupils; Year group(s) of pupils.
<i>Intervention and primary outcome</i>	Health area (e.g., smoking); How was the intervention delivered? (e.g., through classroom lessons); Name of primary outcome (e.g., body mass index (BMI) z-score); Intervention typology classification (using typology in Eldridge and Kerry [3] (p25-29) – individual-cluster, professional-cluster, cluster-cluster, external-cluster, multifaceted); Type of control group (e.g., usual care, active control).
<i>Study design</i>	Justification provided for using cluster trial design (Yes/No, if 'Yes' extract justification); Unit of randomisation (i.e., type of cluster that was randomised); Method used to balance the randomisation (e.g., completely randomised, matched pair, stratified, constrained, minimisation); Timing of randomisation of clusters relative to recruitment of pupils (recruitment of pupils before randomisation, recruitment of pupils after randomisation, unclear); Number of trial arms; Allocation ratio; Length of follow-up.
<i>Sample size information</i>	Justification for target sample size (extract direct quote); If the sample size was calculated, did the calculation account for clustering? (yes/no); Targeted number of schools, clusters and pupils; Number of recruited schools, clusters and pupils.
<i>Ethics and consent procedures</i>	Was ethical approval obtained? (yes/no)

Variable category	Variable
<i>Objectives of feasibility study</i>	<p>Test randomisation process (yes/no); Test willingness to be randomised (at cluster and/or individual levels) (yes/no); Estimate recruitment rate (at cluster and/or individual levels) (yes/no); Estimate retention/follow-up rate (at cluster and/or individual levels) (yes/no); Test implementation of the intervention (yes/no); Test compliance with the intervention (yes/no); Assess acceptability of the intervention (at cluster and/or individual levels) (yes/no); Assess acceptability of trial procedures (at cluster and/or individual levels) (yes/no); Test the feasibility of blinding procedures (yes/no); Test data collection process (yes/no); Test outcome measures (yes/no); Estimate standard deviation for continuous outcomes (or estimate control arm percentage for binary outcomes) (yes/no); Identify potential harms (yes/no); Estimate potential effectiveness of intervention (yes/no); Estimate costs of delivering the intervention (yes/no); Estimate the ICC of the primary outcome (yes/no); Calculate the sample size required for the definitive trial (yes/no); Any other feasibility objectives not listed above.</p>
<i>Other design characteristics of methodological interest</i>	<p>Analysis method used to estimate potential effectiveness of the intervention (Cluster-level analyses/ Individual-level analysis that allows for clustering/ Did not account for clustering/ N/A); Was a p-value for effectiveness reported? (yes/no); Were baseline cluster-level characteristics presented? (yes/no); If so, what were the baseline cluster-level characteristics?; ICC estimates (and 95% confidence intervals); Did the study conclude that a definitive trial was feasible? (yes/ yes (with modifications)/ no).</p>

After the content of the data extraction form was finalised, the principal investigator (KP) extracted data from all included studies. A second reviewer (either SEd or OU) also independently extracted data from all included studies for validation. Missing information that was not available in the index papers was not sought elsewhere and was recorded as 'Not stated'. If there was uncertainty regarding a particular article, the data obtained were checked by another member of the team (MN) and resolved through further discussion.

4.4.5 Data coding and classification

Once the data were extracted for all included texts, some text data were coded for ease of analysis. Data were extracted exactly as provided in each article. This section discusses specific variables where coding was more challenging and specific decisions were made regarding coding for analysis.

4.4.5.1 Publication details

Trial registration status was determined from the paper and by using trial registration information obtained for International Standard Randomised Controlled Trial Number (ISRCTN) registry [225].

4.4.5.2 Setting and participant characteristics

School type was recorded as stated in the article and then categorised as listed on the UK government website [134]. State schools (also called comprehensive, state-maintained, state-funded) receive funding through their local authority or directly from the government. The most common types of state school in the UK are local authority, foundation and voluntary-aided schools. Academies are schools run by government and not-for-profit trusts and are independent of local authority. Grammar schools are run by local authorities, but intake is based on assessment of the pupils' academic ability. Special schools cater for pupils with special educational needs. Faith schools follow the national curriculum but can decide what they teach in religious studies. Independent schools do not have to follow the national curriculum and charge fees for attending pupils.

Additionally, school level and year groups across the devolved nations in the UK were standardised in relation to their equivalent school level and year group in the English school system (i.e., *pre-school* (1 to 4 years), *primary school* (4 to 11

years), *secondary school* (11 to 16 years) and *sixth form/college* (16 to 18 years)). A table comparing school year groups across nations in the UK can be found in Appendix 5 [135].

4.4.5.3 Intervention and primary outcome

The intervention health area was categorised into broad health areas defined after consulting previous systematic reviews [53, 111]. For example, health difficulties such as mental health, behaviour, neurodiversity (e.g., attention deficit hyperactivity disorder (ADHD)), well-being, quality of life, bullying, social and emotional learning, and self-esteem were categorised under '*Social emotional functioning and its influences*'.

Intervention type was summarised using the typology described by Eldridge et al [3](p25-29). The categories were as follows: '*Individual-cluster*' interventions include components acting at the individual level (e.g., pupils); '*Professional-cluster*' interventions include components acting on the trained professionals delivering the intervention; '*External-cluster*' interventions involve using additional staff outside the cluster to deliver the intervention; '*Cluster-cluster*' interventions include components that have to be administered to entire clusters. '*Multifaceted*' interventions which include components across more than one of the '*individual-cluster*', '*professional-cluster*', '*external-cluster*' and '*cluster-cluster*' categories. (see Section 3.4.5.2).

Type of control group was recorded as 'usual care' or 'active'. An active control was defined as, '*a control group in which participants engage in some task during the intervention period that differs from normal practice*' [3](p88-89). If the study had more than one control group this was recorded.

The primary outcome was identified as the health outcome stated in the paper as being the primary outcome. If there were multiple primary outcomes or the primary outcome was unclear, then the outcome presented in the title, first outcome presented in the 'Outcomes' section in the *Methods*, or first outcome presented in the *Results* section was taken as the primary outcome (in this order of priority).

Primary outcome health area was categorised into broad health areas defined after consulting previous systematic reviews [53, 111]. For example, primary

outcomes in the health area of mental health, well-being or behaviour were categorised into '*Social emotional functioning and its influences*'.

4.4.5.4 Study design

Justifications for the use of the CRT design were categorised into common justifications based on potentially reasons established in the literature [3](p10-13). If the study provided more than one reason, these were recorded.

It was assumed that a completely randomised design (i.e., without balancing the randomisation on cluster-level characteristics) was used unless otherwise stated.

Timing of randomisation of clusters relative to recruitment of pupils was determined using the CONSORT flow diagram. If it was not clear from the CONSORT flow diagram then information was extracted and verified from the main text of the article.

4.4.5.5 Sample size information

The target mean cluster size was calculated by dividing the target number of pupils at follow-up by the target number of clusters.

The number of schools, clusters and pupils recruited and followed up was determined using the CONSORT flow diagram. If it was not clear from the CONSORT flow diagram then information was extracted and verified from the main body of the article.

4.4.5.6 Feasibility objectives

A list of common feasibility objectives was made based on those extracted in previous systematic reviews and expert knowledge of authors involved in this review [221, 222]. An 'Other' category was used so that any other feasibility objectives could be extracted. Only formal feasibility objectives were extracted from each article. These were obtained from the *Background* and *Methods* sections of the included articles.

4.4.5.7 Other design characteristics of methodological interest

The method of analysis used in each study to compare the primary outcome data between the trial arms was recorded into general categories for methods of

analyses. For example, Generalised Estimating Equation (GEEs) was categorised as '*Adjusted individual-level analysis*'.

The *Discussion* and *Conclusion* sections of the included papers were searched for information to help determine whether the study had concluded that a definitive trial was feasible.

In order to determine if the feasibility CRTs included in this review had progressed to a definitive trial, further web and literature searching was undertaken. If no article or website was found stating whether a definitive trial had commenced, authors were contacted to ask if their feasibility CRT had progressed to a definitive CRT.

4.4.6 Assessment of study quality

A formal quality/risk of bias assessment was not undertaken for this review as it was not necessary for summarising characteristics of studies. However, some of the data extracted and summarised are indicative of reporting quality of included studies based on the CONSORT extension for both CRTs [57] and pilot and feasibility studies [226]. These include provision of details on the rationale for using the CRT design, the rationale for the target sample size and specifying objectives related to the feasibility of a study.

4.4.7 Data analysis

Results of the search process were reported using a PRISMA flow diagram [125]. Study characteristics were described using means and standard deviations, or medians and interquartile ranges for continuous variables, and numbers and percentages for categorical variables. A histogram was used to illustrate how the rate of publication of such studies has changed over time. Statistical analysis was undertaken using Stata 17 software [227].

4.5 Results

4.5.1 Study selection and PRISMA flow diagram

3,285 articles were identified through searching the MEDLINE database. Following deduplication, 3,247 titles and abstracts were screened for eligibility. Sixty-two articles were eligible for full text screening, of which 24 were included in the review [228-251]. Agreement between reviewers on which articles should be included was 100%. Reasons for exclusion are shown in the PRISMA flow diagram [125] (Figure 4.1).

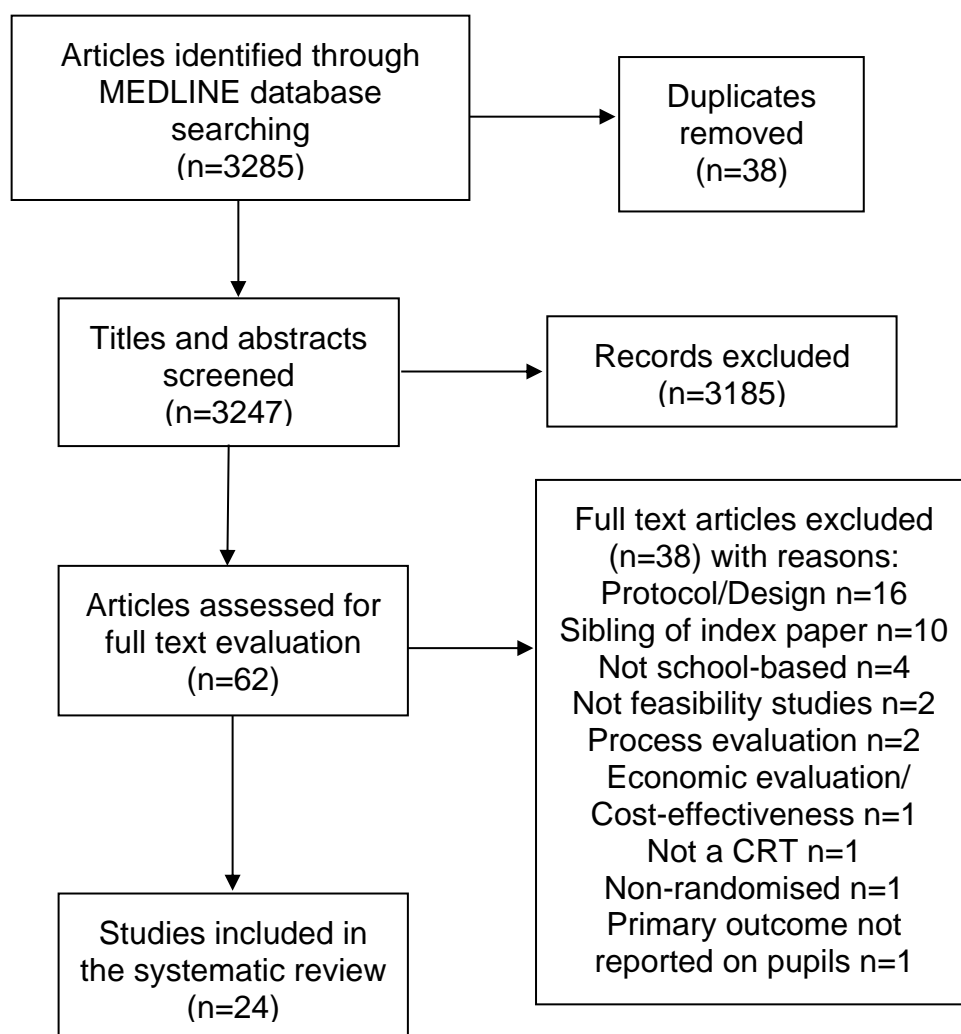


Fig. 4.1. PRISMA flow diagram summarising the results of the literature search and screening for eligibility

4.5.2 Publication details

The first publication of a school-based feasibility CRT for health interventions on pupils in the UK indexed on MEDLINE was in 2008; and the rate of publications of such studies has increased since then (Figure 4.2).

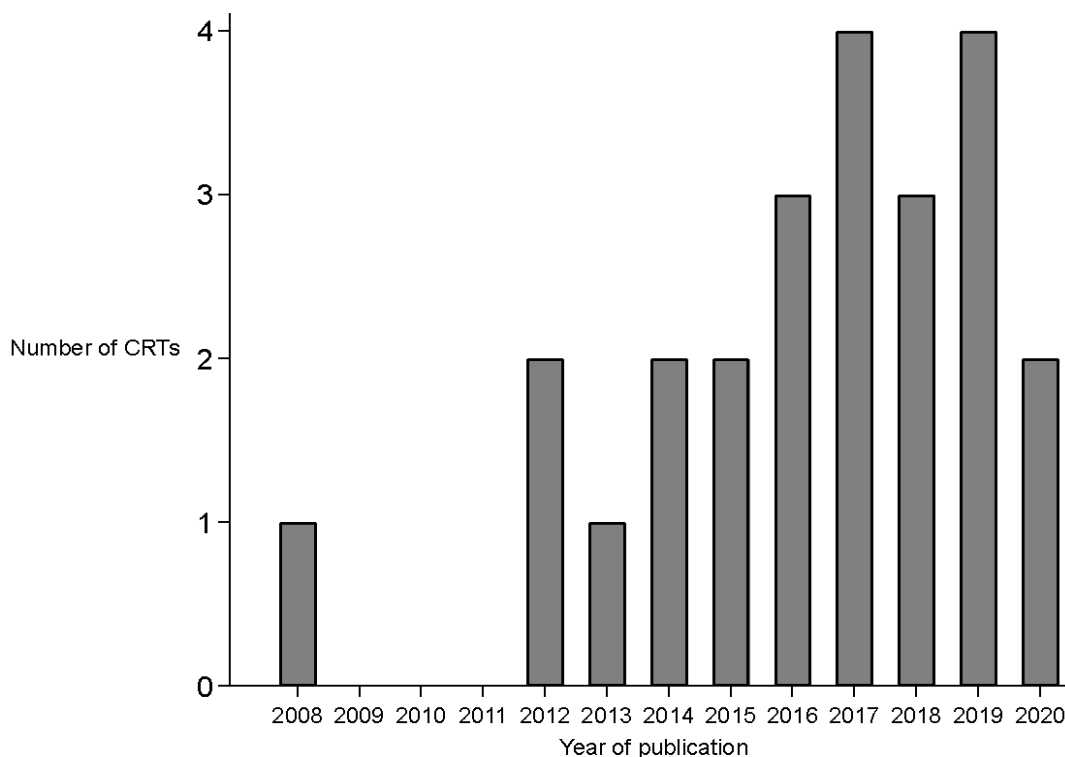


Fig. 4.2. Published feasibility cluster randomised trials indexed on MEDLINE from inception to 31st December 2020 (N=24)

The 24 included articles in this systematic review were published across 11 different journals. For 12 (50%) studies the main funding source was the *National Institute for Health Research (NIHR)*. Further details on journals and funding sources are reported in Appendix 8.

Ten (42%) articles described their study as a 'pilot trial', six (25%) as a 'feasibility trial' [237-239, 242, 246, 247], four as a 'feasibility study' [233, 234, 243, 251], two (8%) as an 'exploratory trial' [241, 248], one (4%) as a 'pilot feasibility trial' [245], and one (4%) as a 'pilot study' [230].

Of the 24 feasibility CRTs, eight (33%) were registered prospectively [229, 232, 235, 236, 244, 245, 247, 251], thirteen (54%) were registered retrospectively, and three (13%) did not state registration status [237, 240, 241].

4.5.3 Setting and participant characteristics

Three quarters of studies (n=18; 75%) took place in England, 13% (n=3) in Northern Ireland [230, 233, 242], 8% (n=2) in Wales [248, 250], and 4% (n=1) in Scotland [239].

Studies in this systematic review were most likely to take place exclusively in secondary schools (n=13; 54%). Eight (33%) studies took place exclusively in primary schools [231, 236, 238-241, 246, 248], 2 (8%) exclusively in pre-schools [228, 243], and 1 (4%) study included both primary and secondary schools [251]. No studies in this review took place in sixth form or college settings.

Fifteen (63%) of the 24 studies provided information regarding the types of school included in their sample. "State" schools were most commonly included (n=14; 93%).

Only seven (30%) studies reported the co-ed status of the schools sampled. Of these, four studies recruited only co-ed schools [229, 235, 242, 244], one study recruited co-ed schools and single sex schools of either gender [233], one study recruited co-ed schools and girl-only schools [234], and one study recruited girl-only schools [249].

Pupils of early teenage years were most commonly recruited (12 years (n=11; 46%) and 13 years (n=11; 46%)). No studies recruited pupils aged 16 or over.

4.5.4 Intervention and primary outcome

The interventions described in these studies aimed to improve outcomes across nine different health areas. Almost half aimed to improve physical activity in school pupils (n=11; 46%). Most often resources and materials provided to schools (n=11; 46%) or classroom lessons (n=10; 42%) were used to deliver the intervention.

Of the 24 studies in the systematic review, 23 (96%) had intervention components that were designed to be delivered to entire clusters ("cluster-cluster" interventions [3](p25-30)). This included components such as classroom lessons [249] and physical activity sessions [228]. Only 2 (8%) studies had intervention components that were aimed at individual pupils ("individual-cluster" interventions [3](p25-30)); the component for both studies was goal-setting [241, 251]. Three

quarters of studies (n=18) had intervention components that were delivered by a professional or person internal to the cluster (“professional-cluster” interventions [3](p25-30)). This included teachers [235], other members of school staff [228], fellow pupils/peers [247]. Eight studies (33%) had intervention components that were delivered by someone external to the cluster (“external-cluster” interventions [3](p25-30)), such as ‘active play practitioners’ [239], researchers [242] and dance teachers [237].

Most studies used a usual care control group (n=21; 88%). Two (8%) studies used an active control group. One of these studies used a goal-setting session followed by an attention control [251]. The other study delivered personal, social, health and economic education (PSHE) sessions which also included the young person receiving feedback that he/she was drinking in a way that may be harmful and provided them with an advice leaflet [245]. One (4%) study had two control arms (a usual care group and an active control group, where clusters received part of the intervention) [237].

A third of the primary outcomes involved measuring moderate-to-vigorous physical activity (MVPA) (n=8; 33%). This included: MVPA (min/week-day) (n=3; 13%) [237, 238, 247]; MVPA (min/day) (n=3; 13%) [229, 232, 233]; MVPA (min/school-day) (n=1; 4%) [239]; and mean minutes of MVPA in the hour before the start of school in the post-baseline week (n=1; 4%) [236].

Five (21%) studies only measured outcomes on female students [230, 234, 237, 247, 249]. In 1 (4%) study the intervention was only delivered to and outcomes measured on pupils who screened positive on an alcohol screening question to identify individuals whose consumption level or pattern is likely to be harmful to their health or well-being [245].

4.5.5 Study design

Only five (21%) of the 24 studies included in the review provided justification for the use of the CRT design. Three studies justified their choice of study design based on the intervention being designed to be delivered to entire clusters [231, 248, 249]. The other 2 studies stated that they chose the CRT design in order to minimise contamination between trial arms [245, 251].

Almost all studies randomised schools as the cluster unit (n=23; 96%). Only 1 study [249] randomised classrooms, with the authors' stating that random allocation was carried out at the level of the classroom for "pragmatic considerations".

Just over half the studies (n=13; 54%) used a form of restricted allocation to balance cluster characteristics between the trial arms. Of these, 5 used minimisation [228, 246, 248, 250, 251], 4 used stratification [231, 233, 237, 244], 3 used matched pairing [229, 239, 242], and 1 used constrained randomisation [238]. The other 11 (46%) studies used unrestricted randomisation methods.

Most studies (n=21; 88%) had two trial arms, 2 (8%) studies had three trial arms [237, 245], and 1 (4%) study had four trial arms [250]. Seventeen (71%) studies allocated clusters in a 1:1 ratio. Three (13%) studies used a 2:1 ratio [232, 244, 247] in favour of the intervention group, 1 (4%) study used a 2:3 ratio [233] in favour of the control group, and 1 (4%) three-arm study used a 2:3:2 ratio [245]. The allocation ratio used in 2 (8%) studies was unclear [230, 237].

Length of follow-up ranged from 2 to 24 months across studies. The median (IQR) length of follow-up was 7 (3 to 12) months. One (4%) study did not state the length of follow-up [234].

Half of the studies (n=12; 50%) recruited pupils before randomisation of clusters. Thirteen (54%) studies reported the baseline characteristics of the schools included in their sample.

4.5.6 Sample size information

Only 3 (13%) studies provided details of a formal sample size calculation. Of these only 1 (4%) study based their sample size on being able to estimate feasibility parameters (e.g., follow-up rates) with a specified level of precision [234]. The remaining 2 (8%) studies based their sample size on power to detect a specified intervention effect [230, 249]. Only 1 (4%) study accounted for clustering in their sample size calculation (to evaluate the intervention effect), which was done by using an ICC estimate to calculate the design effect (DE) [249].

Almost all studies (n=22; 92%) gave justification(s) for their choice of sample size. Of these, 7 (32%) studies based their target sample size on recommendations

from previous articles [228, 231, 235, 239, 243, 245, 246]. Six (27%) studies stated that a formal sample size calculation was not needed [229, 244, 247, 248, 250, 251]. In 5 (23%) studies, the target sample size was determined by resource and/or time constraints [233, 236-239]. Three (14%) studies provided a general statement that their sample size was considered sufficient to address the objectives of the feasibility CRT [236, 242, 251], and 1 (5%) study aimed to recruit as many clusters and participants as possible [240].

The median (IQR; range) target sample size was 7.5 (5 to 8; 2 to 20) schools, 7.5 (5 to 8; 2 to 20) clusters and 320 (150 to 1200; 50 to 1852) pupils. The median (IQR; range) sample size achieved was 7.5 (4.5 to 9; 2 to 37) schools, 8 (5.5 to 9.5; 2 to 37) clusters and 274 (179 to 557; 29 to 1567) pupils. The median (IQR; range) number of pupils per cluster was 35.9 (24 to 89.4; 1.4 to 237.7) and the median (IQR; range) number of pupils per school was 40.4 (24 to 109.3; 1.4 to 237.7). Of the 24 studies in this review, 2 (8%) included just two schools, and allocated one to each trial arm [235, 236]. Eighteen (75%) studies reported both targeted and achieved recruitment numbers at the cluster level, and of these 17 (94%) achieved their target. Thirteen (54%) studies reported both targeted and achieved recruitment numbers at the pupil level, and of these 6 (46%) of studies achieved their target. Two (8%) studies [234, 248] reported losing at least one cluster to follow-up.

The median (IQR; range) sample size at follow-up was 6.5 (4 to 8; 2 to 19) schools, 7 (5 to 8; 2 to 19) clusters and 197 (118 to 409; 17 to 1460) pupils.

4.5.7 Ethics and consent procedures

Twenty-two (92%) studies reported obtaining ethical approval for their study. One study did not state whether they had obtained ethical approval or not [249]. In another study the authors' reported that they "did seek ethical approval but the local research committee said it was not required as the study did not involve patients or National Health Service (NHS) staff" [240].

4.5.8 Analysis methods

Twenty (83%) of the 24 studies reported intervention effect estimates. Of these, 9 (45%) used an individual-level analysis method that allowed for clustering [229, 234, 237, 238, 240, 247-250], 4 (20%) used a cluster-level analysis method [228,

231, 241, 246], 4 (20%) did not allow for clustering in their analysis [230, 232, 236, 242], and 3 (15%) did not state which analysis method they used [235, 239, 245].

Eight (33%) studies reported p-values with the intervention effect estimate, which is at odds with published guidance for feasibility studies [57, 226].

4.5.9 Feasibility objectives

All 24 studies included in the review stated their formal feasibility objectives. The most common objective was to estimate the potential effectiveness of the intervention (n=17; 71%), including 2 studies that sought to undertake a definitive test of effectiveness [230, 249]. Other common objectives included assessing acceptability of the intervention (n=16; 67%), estimating the recruitment rate (n=15; 63%), estimating the retention/follow-up rate (n=15; 63%), and testing outcome measures (n=14; 58%).

The following feasibility objectives were stated specifically at the cluster-level: 10 (42%) studies sought to assess acceptability of the intervention [228, 229, 231, 233, 235, 241, 244, 245, 248, 250], 7 (29%) studies sought to estimate retention/follow-up rate at the cluster-level [231, 233, 236, 241, 244, 248, 250]. A quarter of studies (n=6; 25%) sought to estimate the recruitment rate at the cluster-level [233, 236, 241, 244, 248, 250]. Four (17%) studies sought to assess the willingness of clusters to be randomised [229, 231, 235, 244], and 3 (13%) sought to assess acceptability of the trial procedures [229, 244, 245].

Only 1 (4%) study sought to assess the appropriateness of cluster randomisation as a formal objective [251]. No studies sought to assess which type of cluster was most appropriate to randomise. Four (17%) studies randomised clusters before recruiting pupils [239, 248-250], and of these none investigated the possibility of recruitment bias. Only 2 (8%) studies sought to estimate the ICC of the primary outcome for use in the sample size calculation of the planned definitive study [241, 244]. All 24 studies reported the results for additional feasibility outcomes beyond those that they formally listed as objectives of the study.

No studies quantified the precision of their estimates of feasibility parameters, other than for potential intervention effectiveness, cost-effectiveness and the ICC.

Analyses were undertaken to investigate if the target sample size differed according to whether or not the studies addressed specific feasibility objectives. It was hard to identify clear patterns in the data as some formal objectives were only stated in a small number of the included studies. The twelve studies that assessed potential effectiveness aimed to recruit a median (IQR; range) of 7 (3.5 to 8; 2 to 20) schools, similar to the targeted recruitment in the remaining studies (7.5 (6 to 8; 5 to 12)).

Table 4.3 summarises the methodological and design characteristics of the studies included in this systematic review.

Table 4.3. Summary of methodological characteristics of included studies (N=24)

Characteristic	N	Statistic (n (%))
Setting		
<i>Country</i>	24	
England, n (%)		18 (75)
Scotland, n (%)		1 (4)
Wales, n (%)		2 (8)
Northern Ireland, n (%)		3 (13)
 <i>School types that were included [134] [Accessed 1st September 2021]¹</i>	15	
State, n (%)		14 (93)
Academy, n (%)		3 (20)
Voluntary aided, n (%)		1 (7)
Foundation, n (%)		1 (7)
Faith, n (%)		1 (7)
Grammar, n (%)		1 (7)
Independent, n (%)		1 (7)
 Intervention and primary outcome		
<i>Health area</i>	24	
Physical activity, n (%)		11 (46)
Physical activity and nutrition, n (%)		4 (17)
Alcohol misuse, n (%)		2 (8)
Sexual health, n (%)		2 (8)
Bullying, n (%)		1 (4)
Behavioural/social difficulties (Autism), n (%)		1 (4)
Body image, n (%)		1 (4)
Dating and relationship violence, n (%)		1 (4)
Illicit drug misuse, n (%)		1 (4)

Characteristic	N	Statistic (n (%))
<i>Type of intervention [3]²</i>	24	
Individual-cluster, n (%)		2(8)
Professional-cluster, n (%)		18 (75)
External-cluster, n (%)		8 (33)
Cluster-cluster, n (%)		23 (96)
Multifaceted, n (%)		21 (88)
 <i>Intervention components</i>	 24	
Resources and materials for schools, n (%)		11 (46)
Classroom lessons, n (%)		10 (42)
Physical activity lessons, n (%)		5 (21)
Incentive scheme, n (%)		4 (17)
Change in school/class environment, n (%)		4 (17)
Peer support, n (%)		3 (13)
Support for parents/guardians, n (%)		3 (13)
Goal setting, n (%)		2 (8)
Staff training, n (%)		2 (8)
Home activities, n (%)		2 (8)
Extracurricular physical activity, n (%)		2 (8)
Parent's evenings, n (%)		1 (4)
Drama workshops, n (%)		1 (4)
Funding, n (%)		1 (4)
School action group formation, n (%)		1 (4)
School club sessions, n (%)		1 (4)
Screening, n (%)		1 (4)
Feedback, n (%)		1 (4)
Motivational interviews, n (%)		1 (4)
Interactive sessions, n (%)		1 (4)

Characteristic	N	Statistic (n (%))
Discussions with parents/guardians, n (%)		1 (4)
Gamification (competitive) techniques, n (%)		1 (4)
<i>Type of control group</i>	24	
Usual care, n (%)		21 (88)
Active, n (%)		2 (8)
Two control groups (one usual care and one active control), n (%)		1 (4)
<i>Primary outcome</i>	24	
Moderate-to-vigorous physical activity (min/day), n (%)		3 (13)
Moderate-to-vigorous physical activity (min/weekday), n (%)		3 (13)
Moderate-to-vigorous physical activity (min/school-day), n (%)		1 (4)
Mean minutes of moderate-to-vigorous physical activity in the hour before the start of school in the post-baseline week, n (%)		1 (4)
Avoidance of unprotected sexual intercourse, n (%)		1 (4)
Alcohol consumption 28-days before "Timeline Followback" questionnaire, n (%)		1 (4)
Body mass index (BMI) z score, n (%)		1 (4)
Body Esteem Scale, n (%)		1 (4)
Bullying victimisation scale (Gatehouse), n (%)		1 (4)
Drinking initiation, n (%)		1 (4)
Health, nutrition and physical activity knowledge, n (%)		1 (4)
Lifetime illicit drug use, n (%)		1 (4)
Minutes spent on screen-based activities, n (%)		1 (4)
Overweight status, n (%)		1 (4)
Returning of completed vaccination consent form , n (%)		1 (4)
Safe Dates and short Conflicts in Adolescent Dating Relationships Inventory, n (%)		1 (4)
School-time physical activity, n (%)		1 (4)
Total difficulties score (Strengths and Difficulties Questionnaire), n (%)		1 (4)
Sedentary activity (min), n (%)		1 (4)

Characteristic	N	Statistic (n (%))
Time spent sitting (min/day), n (%)		1 (4)
Study design		
<i>Justification for CRT design</i>	24	
Yes, n (%)		5 (21)
<i>Type of randomisation</i>	24	
Completely randomised, n (%)		11 (46)
Minimisation, n (%)		5 (21)
Stratified, n (%)		4 (17)
Matched pair, n (%)		3 (13)
Constrained [38, 39], n (%)		1 (4)
<i>Number of trial conditions</i>	24	
Two, n (%)		21 (88)
Three, n (%)		2 (8)
Four, n (%)		1 (4)
<i>Length of follow-up</i>	24	
Up to 6 months, n (%)		11 (46)
7 to 12 months, n (%)		8 (33)
13 to 18 months, n (%)		3 (13)
More than 18 months, n (%)		1 (4)
Not stated, n (%)		1 (4)
<i>Were pupils recruited before randomisation of clusters?</i>	24	
Pupils recruited before randomisation, n (%)		12 (50)
Pupils recruited after randomisation, n (%)		4 (17)
Unclear, n (%)		8 (33)

Characteristic	N	Statistic (n (%))
<i>Were baseline cluster-level characteristics reported?</i>	24	
Yes, n (%)		13 (54)
<i>Ethical approval</i>		
<i>Was ethical approval obtained?</i>	24	
Yes, n (%)		22 (92)
No, n (%)		1 (4)
Not stated, n (%)		1 (4)
<i>Sample size</i>		
<i>Type of justification for sample size</i>	24	
Formal sample size calculation ³ , n (%)		3 (13)
Other justification ⁴ , n (%)		19 (79)
Not stated, n (%)		2 (8)
<i>Target number of schools, median (IQR; range)</i>	18	7.5 (5 to 8; 2 to 20)
<i>Target number of clusters, median (IQR; range)</i>	18	7.5 (5 to 8; 2 to 20)
<i>Target number of pupils, median (IQR; range)</i>	13	320 (150 to 1200; 50 to 1852)
<i>Achieved number of schools, median (IQR; range)</i>	24	7.5 (4.5 to 9; 2 to 37)

Characteristic	N	Statistic (n (%))
<i>Achieved number of clusters, median (IQR; range)</i>	24	8 (5.5 to 9.5; 2 to 37)
<i>Achieved number of pupils, median (IQR; range)</i>	24	274 (179 to 557; 29 to 1567)
<i>Achieved mean cluster size, median (IQR; range)</i>	24	35.9 (24 to 89.4; 1.4 to 237.7)
Objectives of the feasibility study		
<i>Feasibility objectives</i>	24	
Test randomisation process, n (%)		3 (13)
Test data collection process, n (%)		8 (33)
Test willingness to be randomised (at cluster level and/or individual level), n (%)		4 (17)
Estimate recruitment rate (percentage) (at cluster level and/or individual level), n (%)		15 (63)
Estimate follow-up rate (percentage) (at cluster level and/or individual level), n (%)		15 (63)
Test implementation of intervention, n (%)		10 (42)
Test compliance with intervention, n (%)		6 (25)
Assess acceptability of intervention (at cluster level and/or individual level), n (%)		16 (67)
Assess acceptability of trial procedures (at cluster level and/or individual level), n (%)		6 (25)
Test the feasibility of blinding procedures, n (%)		0 (0)
Test outcome measures, n (%)		14 (58)
Estimate standard deviation of continuous outcomes or control arm rate for binary outcomes, n (%)		1 (4)
Identify potential harms, n (%)		3 (13)
Assess potential effectiveness of intervention, n (%)		17 (71)
Estimate intervention cost, n (%)		7 (29)
Estimate the ICC of the primary outcome, n (%)		2 (8)

Characteristic	N	Statistic (n (%))
Calculate sample size for definitive trial, n (%)		5 (21)
Analysis methods		
<i>Analysis method for estimating potential effectiveness</i>	24	
Individual-level analysis that allows for clustering, n (%)		9 (38)
Cluster-level analysis, n (%)		4 (17)
Did not allow for clustering, n (%)		4 (17)
Not stated, n (%)		3 (13)
Did not estimate potential effectiveness, n (%)		4 (17)
<i>p-value reported for effectiveness</i>	24	
Yes, n (%)		8 (33)

¹ Some studies included more than one school type. This is the number of studies that included specific types of school. State schools receive funding through their local authority or directly from the government; the most common ones are local authority, foundation and voluntary aided school which are all funded by the local authority. Academies are run by government and not-for-profit trusts, and are independent of local authority. Grammar schools are run by local authorities, but intake is based on assessment of the pupils' academic ability. Special schools cater for pupils with special educational needs. Faith schools follow the national curriculum but can decide what they teach in religious studies. Independent schools follow the national curriculum but charge fees for attending pupils.

² Intervention type has been described using the typology of Eldridge and Kerry [3]. 'Individual-cluster' interventions contain components that are aimed at individuals. 'Professional-cluster' interventions contain components that are delivered by a professional or person internal to the cluster (e.g. teacher, pupils). 'External-cluster' interventions contain components that require people external to the cluster to deliver the intervention (e.g. research staff, community support consultant). 'Cluster-cluster' interventions contain components that have to be delivered at the cluster level (e.g., classroom lessons). 'Multifaceted' interventions contain components across more than one of the 'individual-cluster', 'professional-cluster', 'external-cluster' and 'cluster-cluster' categories.

³ In one study, the sample size was based on being able to estimate feasibility parameters with a pre-specified level of precision. Two studies based their sample size on a definitive test of intervention effectiveness.

⁴ Other reasons were: based their target sample size on recommendations from previous articles; stated that a formal sample size calculation was not needed; the target sample size was determined by resource and/or time constraints; provided a general statement that their sample size was considered sufficient to address the objectives of the feasibility CRT; aimed to recruit as many clusters and participants as possible.

4.5.10 Estimated intra-cluster correlation coefficients

One third (n=8; 33%) of studies reported estimates of the ICC for the provisional primary outcome of the planned definitive study. Of these, 5 (63%) studies also provided 95% confidence intervals (CIs) for these estimates. Table 4.4 reports the ICC estimates.

Most of the reported 95% CIs for the ICCs were wide as the sample size was too small to precisely estimate the ICC. However, two ICC estimates did have an upper bound of 0.03 despite those studies having only 6 [247] and 19 [240] clusters. This still provides information of practical use regarding plausible true values of the ICC.

Table 4.4. Reported intra-cluster correlation coefficients (ICCs) for primary outcomes (N=8)

Author (Year)	Cluster unit	Health area	Outcome	Outcome type	ICC (95% CI¹)
Jago (2012) [237]	Schools	Physical activity	MVPA ² (mins/weekday)	Continuous	0.018 (<0.001 to 0.087)
Jago (2014) [238]	Schools	Physical activity	MVPA (mins/weekday)	Continuous	0.0653 (0.00091 to 0.12977)
Sebire (2018) [247]	Schools	Physical activity	MVPA (mins/weekday)	Continuous	<0.0001 (0.0 to 0.03)
Kipping (2008) [240]	Schools	Physical activity and nutrition	Time spent on screen-based activities (mins)	Continuous	0.01 (0 to 0.03)
Lloyd (2012) [241]	Schools	Physical activity and nutrition	BMI ³ z score	Continuous	0.04 (0 to 0.15)
Sahota (2019) [246]	Schools	Physical activity and nutrition	Healthy nutrition and physical activity knowledge	Continuous	0.07
Segrott (2015) [248]	Schools	Alcohol misuse	Drinking initiation	Binary	0.112
White (2017) [250]	Schools	Illicit drug misuse	Lifetime illicit drug use	Binary	0.003

¹ 95% confidence intervals² Moderate-to-vigorous physical activity³ Body Mass Index

4.5.11 Feasibility study conclusions

Of the 24 studies included in this review, 11 (46%) concluded that the definitive trial was feasible [228, 230, 231, 233, 235, 236, 239, 243, 246, 249, 251]. A further 11 studies (46%) said the definitive trial would be feasible with some modifications [229, 232, 234, 237, 238, 240-242, 245, 247, 250], and 2 (8%) said that the planned study was not feasible [244, 248]. Through searching the literature and via personal correspondence with the authors, it has been confirmed that 11 (46%) of the 24 feasibility CRTs have advanced to definitive trials [229, 230, 232, 237, 240-242, 245, 247, 250, 251]. Of these, 9 (82%) had concluded that the definitive trial was feasible, and 2 (18%) had concluded that the definitive trial would be feasible with modifications.

4.6 Discussion

This is the first systematic review to summarise the characteristics and objectives of school-based feasibility CRTs of interventions to improve pupil health outcomes in the UK. The review found an increase in the rate of publication of school-based feasibility CRTs since the earliest included paper was published in 2008 [240]. This mirrors the increase in definitive CRTs in this area reported in the parallel systematic review in Chapter 3 [117], and highlights the growing use of the CRT design in health-based research in the school-setting. The increase in feasibility CRTs may partly be due to the publication of the 2006 MRC guidelines for the evaluation of complex interventions [252] which emphasises the importance of conducting feasibility studies ahead of full-scale trials. The relatively large number of feasibility CRTs with interventions for increasing physical activity indicates the growing importance of adolescent physical activity as a public health priority and the use of schools as a setting to deliver these types of intervention [253]. The review of school-based definitive CRTs also found that the design is increasingly used to evaluate physical activity interventions [117]. Based on what was observed in the review of definitive school-based CRTs, there were fewer than expected feasibility studies in the area of social emotional functioning and its influences. This is despite the increased awareness of the prevalence of these health conditions and research funding in this area [254].

The studies included in the current review sought to address a range of feasibility objectives, most commonly estimating potential effectiveness of the intervention. It was notable, however, that few studies formally stated objectives that were related to uncertainties that are unique to the cluster design. This finding is similar to another review of feasibility CRTs measuring health outcomes which also stated that few studies investigated issues specific to the complexities of the CRT design [222]. In this chapter, only one study assessed whether a cluster design was needed, and none used the research to decide on the type of school-based cluster (e.g., school versus classes) that was best to randomise. It may be the case that the need for cluster randomisation and the appropriate kind of cluster to allocate had a strong theoretical basis, negating the need for empirical justification, but only 5 of the 24 studies provided a rationale for the cluster design even though the CONSORT extension for CRTs [57] recommends reporting this. Finally, none of the 4 studies that randomised clusters before recruiting pupils investigated the potential for recruitment bias as a feasibility objective.

Only 3 studies in the review reported details of a formal calculation for the target sample size [230, 234, 249], with just 1 accounting for clustering [249]. These results are similar to that found in a previous systematic review of feasibility CRTs which reported that only 1 of the 18 studies reported a formal sample size calculation based on the primary feasibility objective [221]. A quarter of the included papers in this chapter stated that a formal sample size calculation was not needed, and some authors have argued that a formal sample size calculation is not always appropriate in feasibility studies [219]. In a recent review of current practice in feasibility studies, only 36% reported sample size calculations [255]. Also, when surveyed, some journal editors stated they were willing to accept pilot studies for publication that did not report a sample size calculation [255]. The precision with which parameters are estimated in feasibility CRTs should be reported, especially given the small number of clusters that are typically included in such studies, although this was not done by any papers in this review. Furthermore, a formal sample size calculation based on the feasibility objectives that allows for clustering [44] is appropriate to ensure the study is large enough to estimate parameters precisely and, therefore, minimise the uncertainty regarding the assumptions that are made for the subsequent definitive study [219, 255].

The systematic review found the median number of clusters recruited (8 clusters) was similar to a previous review [221]. Based on results from a simulation study, it has been suggested that as many as 30 or more clusters may be required to obtain accurate estimates of feasibility parameters in pilot CRTs, including estimating the number of clusters required to test the intervention effect in the subsequent definitive CRT, due to the imprecision of estimated ICC from feasibility studies [44]. The current review found only 2 studies recruited more than 20 clusters [238, 251] as it is difficult to achieve this level of recruitment due to funding and practical constraints of feasibility CRTs [44]. However, smaller feasibility CRTs may still produce informative estimates of many parameters. Two of the feasibility studies in the review were able to estimate the ICC with a 95% confidence interval upper bound of 0.03, despite including only 6 [247] and 19 [240] clusters, respectively. Such a finding could rule out the need for unachievable, large sample sizes in the definitive study. A large number of studies report feasibility objectives in the form of percentages (e.g., recruitment and follow-up). A formula for calculating the sample size required in feasibility CRTs to estimate percentages of individual-level characteristics (e.g., whether the pupil was followed up) with a confidence interval of specified width, whilst allowing for clustering, is provided in an article by Eldridge and colleagues [44]. Using the average sample size based on the findings in this review (i.e., 8 schools and 240 pupils), and assuming an ICC for the feasibility characteristic is 0.05, a study of this size would be large enough to estimate percentages for pupil-level characteristics with a margin of error no greater than 10 percentage points based on a 95% confidence interval. There will normally be little precision for estimating percentages based on cluster-level characteristics since this is determined by the typically small number of schools (clusters) in feasibility studies.

Another important reason to recruit sufficient clusters to feasibility CRTs is to assess how the intervention might be implemented and the trial delivered in a range of different types of cluster [221]. Parameter estimates will only be useful to the extent that the clusters and individuals in the feasibility study are broadly representative and reflect the diversity of the population from which the sample in the definitive trial will be drawn [221]. In the context of school-based trials, important aspects of representativeness include single sex versus co-educational schools, state versus independent schools, and deprived versus non-deprived

areas. In the current review, only 54% of studies reported baseline characteristics of the schools, although this is higher than found in a previous systematic review of feasibility CRTs where only 11% of studies reported baseline cluster-level characteristics [221].

In the current systematic review, of the 13 studies that reported both targeted and achieved numbers of pupils recruited, those targets were only achieved in 46% of studies. A previous systematic review of definitive school-based CRTs found that only 77% of studies achieved their recruitment target of pupils [117]. Facilitators and barriers to the recruitment and retention of pupils to school-based CRTs have been discussed in detail in the literature [81, 256, 257]. Challenges include: the type of intervention being offered (e.g., sexual education) [81, 257]; lack of time [81]; incompatibility of the intervention with the needs of pupils or parents or with the school's ethos [81]; and a lack of incentivisation [256].

4.7 Strengths and limitations

In terms of strengths, the review used a predefined search strategy to identify school-based CRTs that was previously used in a published systematic review [117]. The strengths and limitations of the search strategy have been discussed in Chapter 3 at length (Section 3.7). The search was not limited to articles that included terms such as 'feasibility' and 'pilot' in case eligible texts did not use the terms in their titles, abstracts or key words. The protocol was publicly available prior to conducting the review. Screening, piloting of the data extraction form, and data extraction were undertaken by two independent reviewers in order to minimise errors.

As well as a number of strengths, this systematic review also has some limitations. The search was limited to one database. However, MEDLINE was chosen as health-based studies were the focus of this review. Further articles may have been found by searching other databases, grey literature and through citation searching but this was considered unnecessary due to the precision of the search strategy used. While the approach was not comprehensive, it was used to efficiently identify studies of interest that were undertaken in advance of planned definitive CRTs.

Articles were only eligible for the systematic review if they reported the findings of a feasibility study that used a CRT design, therefore potentially missing out on relevant knowledge from other types of study design, such as non-randomised parallel arm and single-arm feasibility studies. These types of studies were excluded as the review was interested in studies that could be used to assess a wide range of uncertainties for definitive CRTs, including those related to the randomisation process.

Data on consent procedures used by the included studies were not extracted. In the previous review of definitive school-based CRTs [117] this information was inconsistently reported across studies making it difficult to summarise, and highlights the need for more comprehensive reporting of the consent procedures in CRTs.

4.8 Implications

The systematic review has identified a number of important implications for the planning and conduct of school-based feasibility CRTs. The findings of the systematic review show that few studies performed a formal sample size calculation or gave statistical justification for their choice of sample size in their feasibility study. Despite this and the fact that studies usually include few schools, the median sample size of the studies included in the review was large enough to estimate pupil-level feasibility parameters (e.g., percentage followed up) with reasonable precision. This information adds to the growing literature on sample size for feasibility CRTs.

Besides potential effectiveness of the intervention, ICC and cost-effectiveness, no studies reported the precision with which feasibility parameters were estimated. If future researchers address these issues they will minimise uncertainty regarding assumptions that are made when planning the subsequent definitive trial.

The characteristics of the recruited schools in feasibility CRTs could be better described to help understand the extent to which the feasibility parameter estimates are applicable to the planned definitive trial.

The review also presented the formally stated feasibility objectives of school-based feasibility CRTs, the findings of which highlight that few studies, if any, explored challenges specific to the CRT design. For example, which type of cluster (e.g., school vs class) would be best to randomise and the possible existence of recruitment bias in studies that randomise clusters before recruiting pupils. Greater use should be made of feasibility studies to explore uncertainties specific to the CRT design.

The findings of this systematic review will also provide relevant parameter values for simulation studies that use synthetic data to assess the statistical properties of methods used to analyse data from school-based feasibility CRTs.

4.9 Conclusions

Feasibility CRTs are being increasingly used in the school setting to test feasibility prior to definitive trials. The findings from the review emphasise a need for clearer justification regarding the sample size of school-based feasibility CRTs, and that authors should report the precision with which feasibility parameters are estimated. Despite the studies included in the review usually randomising a small number of schools, the median sample size (8 clusters) would be large enough to estimate pupil-level feasibility parameters in the form of percentages (e.g., follow-up rates, intervention adherence rates) with a reasonable level of precision. Better reporting of the characteristics of the recruited schools in feasibility CRTs could help researchers to understand the extent to which the feasibility parameter estimates are appropriate for use in the planning of definitive trials. Finally, better use could be made of feasibility CRTs to explore challenges that are specific to the CRT study design.

4.10 Chapter summary

The chapter presented findings from a systematic review of the characteristics, methodological practices and objectives of school-based feasibility CRTs undertaken in advance of definitive CRTs to evaluate interventions aimed at improving health outcomes on pupils and adds to the knowledge reported in Chapter 3. Chapters 5 and 6 will present the findings from a programme of

research aimed at addressing the lack of suitable ICC estimates, as identified by the systematic reviews presented in Chapter 3 and 4.

Page intentionally left blank

Chapter 5: Intra-cluster correlation coefficients from school-based cluster randomised trials of interventions for improving health outcomes on pupils

5.1 Summary

This chapter presents a background highlighting the need to collate and summarise intra-cluster correlation coefficient (ICC) estimates for pupil health outcomes from school-based CRTs. The chapter then summarises the aims, objectives, methods and results of a literature search and results of analyses examining the ICCs from school-based CRTs worldwide. The chapter concludes by discussing the results, strengths and limitations, implications and areas identified for further research. A peer-reviewed journal article has been published of the work in the chapter [123] (Appendix 4).

5.2 Background

CRTs require more participants than individually RCTs because observations on individuals in the same cluster are usually more similar than those from different clusters [4](p6-7). Researchers need to take into consideration the correlation between individuals within clusters when calculating the sample size for a CRT, otherwise the study will be underpowered [4](p6-7). This is done by inflating the sample size that would be required for an individually randomised trial by the design effect (DE):

$$DE = 1 + (\bar{m} - 1)\rho$$

where \bar{m} is the mean number of participants providing outcome data in each cluster (cluster size) and ρ is the intra-cluster correlation coefficient (ICC) of the outcome [3](p142). The ICC quantifies the similarity of observations on individuals within clusters. For continuous outcomes, the ICC (ρ) is the proportion of the total variation in the outcome that is between clusters (σ_b^2) as opposed to between individuals within clusters (σ_w^2)[10].

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

where σ_b^2 is the between-cluster variance component and σ_w^2 is the within-cluster variance component [3](p174). Under this definition, the ICC can take values between zero and one. The larger the ICC, the greater the sample size required.

The proportion of variance definition of the ICC is expressed differently for binary outcomes [11], for which the overall variance of the outcome, $\pi(1 - \pi)$, depends on the outcome prevalence, π [13]. The definition for the ICC, $\rho_{b(linear)}$, for a binary outcome is:

$$\rho_{b(linear)} = \frac{var(\pi_i)}{\pi(1 - \pi)}$$

where π_i is the proportion with the binary trait in the i th cluster and $var(\pi_i)$ is the variance of the cluster proportions (between-cluster variation). Under this definition the total outcome variance is expressed on the linear (proportions) scale.

There is different definition of the ICC for binary outcomes where the between-cluster variation is expressed on the logit, or log odds, transformation of π_i :

$$logit(\pi_i) = \ln(\pi_i/(1 - \pi_i))$$

This definition of the ICC assumes that the binary outcome is the dichotomised version of an underlying latent continuous variable that represents the tendency of an individual level cluster member to have the binary trait [10]. Individuals for whom the value of this latent variable is over a certain threshold, have the binary trait (coded 1) while the remaining individuals do not have the trait (coded 0). The underlying continuous variable is assumed to follow a logistic distribution. The definition for the ICC, $\rho_{b(logit)}$, for a binary outcome is then:

$$\rho_{b(logit)} = \frac{var(logit(\pi_i))}{var(logit(\pi_i)) + (\pi^2/3)}$$

where $var(logit(\pi_i))$ is the between-cluster variance on the logit scale, π is the mathematical constant (~ 3.141592654), and $\pi^2/3$ is the within-cluster variance on the logit scale [10].

In the context of sample size calculations for binary and count outcomes, similarity between participants from the same cluster can also be quantified by the between-cluster coefficient of variation of the outcome (CV) (the ratio of the between-cluster standard deviation to the outcome mean [5]):

$$CV = \frac{\sigma_b}{\mu}$$

where σ_b is the between-cluster standard deviation and μ is the mean outcome across the clusters [5]. The CV can be incorporated into a modified design effect formula. The ICC/CV estimate is usually unknown at the time of sample size calculation for a CRT and ideally should be obtained from previous studies that randomised or sampled the same type of cluster and measured the same or a similar outcome as the one in the planned study [3](p172-173).

Previous studies have reported that ICC estimates for pupil health outcomes are usually smaller than for educational outcomes in school settings [68-70]. ICCs for educational outcomes from school settings might be expected to be higher than for health outcomes as the main purpose of schools is to provide education [71]. Although, ICCs for health outcomes in health care settings are well established, particularly in primary care [3, 26, 72], it is not known how these estimates translate into the school setting and there is a comparative lack of reported ICC estimates of health outcomes from school-based CRTs.

The systematic review in Chapter 3 highlighted the poor reporting of ICC estimates, finding that only 45% of UK school-based definitive CRTs evaluating interventions for improving pupil health outcomes reported the ICC for their primary outcome [117]. Much of the existing literature summarising ICCs from CRTs or surveys in school settings is from the US, meaning these ICCs may not be generalisable to other countries and school systems. Existing summaries generally focus on specific outcomes such as substance misuse [61, 71, 97-104], nutrition [105-107], physical activity [61, 107-109, 258], and mental health and behaviour [61, 69, 96, 259], and there is a lack of estimates for other health outcome areas, such as infectious diseases and dental health. Additionally, it has been suggested that ICCs for pupil health outcomes need to be both country- and outcome area-specific [71], but this hypothesis needs further investigation in the school setting. Patterns in the size of the ICC have been explored [13, 26, 71, 98-

100, 102, 260-262], but little is known about the extent to which the size of ICC estimates from school-based CRTs differ by study characteristics.

To date, no study has summarised ICC estimates from a range of health outcome areas in different settings. Such a study would provide information on plausible ICC values for use in sample size calculations and aid the design of future school-based CRTs. Summarising ICC estimates specifically from CRTs, rather than data from surveys, is potentially more relevant for use in future sample size calculations as such estimate may better reflect the level of variation in outcomes amongst the types of schools that participate in such trials [3](p177).

5.3 Aims and objectives

The aim of this study was to collate ICC estimates for pupil health outcomes from school-based CRTs across different countries, educational systems and outcome areas. Additionally, the study sought to better understand the relationships that design, and contextual factors have with the size of ICC estimates.

The objectives were to:

- Collate and summarise ICC estimates for pupil health outcomes from school-based CRTs worldwide.
- Examine the relationship between the size of the ICC and study characteristics.

5.4 Methods

In order to find articles reporting estimates of ICCs from school-based CRTs, a systematic searching approach was used. This method used a previously developed search strategy (see Section 3.3.1) for identifying school-based CRTs for evaluating the effect of interventions on pupil health outcomes. The use of a systematic approach makes it possible to replicate the study in the future.

5.4.1 Search strategy

The study used the same systematic search strategy used in Chapters 3 and 4 (see Section 3.3.1). A brief description of the search strategy is provided here.

Articles, indexed on the MEDLINE (via Ovid) database, that reported the results of peer-reviewed CRTs that randomised school related units and measured at

least one health outcome on pupils were the source of data for the study. MEDLINE was systematically searched from inception to 18th October 2021 for eligible articles. The search strategy (Table 3.1) was developed based on a search strategy by Taljaard and colleagues [127] for identifying CRTs, with the addition of school-related terms. The search was limited to articles written in English.

5.4.2 Eligibility criteria

Eligible articles reported the findings from CRTs that randomised school-related units, measured at least one health outcome on school pupils, and presented at least one estimate of the ICC or between-cluster CV.

5.4.2.1 Population

The eligible study population was pupils attending full-time education at either pre-primary, primary, or lower and upper secondary educational settings according to the United Nations Educational, Scientific and Cultural Organisation (UNESCO) International Standard Classification of Education (ISCED) system [263].

The term 'pre-school' was defined as an organisation offering early childhood education prior to compulsory (primary) education [130]. This included nursery schools, *Head Start* schools, educational childcare centres and kindergartens. Studies that took place in higher/tertiary education settings were excluded. Studies were excluded if they randomised after-school clubs, school-based health centres or childcare centres (i.e., provided childcare only).

The types of eligible clusters included schools, year-groups, classes/classrooms, teachers or any other relevant school-related unit of randomisation. Studies that randomised both school-related units and other types of clusters (e.g., villages, households) were eligible for inclusion in this review as long as the results of the study were shown separately for the school-related clusters (i.e., the authors did not pool results across the different types of clusters).

5.4.2.2 Intervention

All interventions were considered, including educational interventions. Interventions had to be administered to the pupils or administered to an individual

who interacted with the pupil (e.g., parents/carers, school staff) to be eligible. Interventions could be universal (i.e., aimed at all participants) or targeted (i.e., aimed at a specific group of participants) in the way they were administered.

5.4.2.3 Comparison

There had to be at least one control/usual care comparison group in the study. Any type of comparator was eligible including active control group(s). An active control was defined as, '*a control group in which participants engage in some task during the intervention period that differs from normal practice*' [3](p88-89).

5.4.2.4 Outcome

Eligible articles reported at least one health-related outcome that was measured on the school pupils. Articles reporting primary outcomes that were not related to health (e.g., improvements in literacy), and/or primary outcomes that were not measured on pupils (e.g., parents/carers) were still included as long as there was at least one secondary outcome of interest related to pupil health.

5.4.2.5 Study type

All types of CRT design were eligible, including parallel arm, crossover, factorial and stepped wedge studies. Feasibility CRTs were eligible. Non-randomised trials, single-arm trials and quasi-experimental designs (no random assignment) were excluded.

5.4.2.6 Other eligibility criteria

Articles had to present the components of variance and/or the ICC and/or the between-cluster coefficient of variation of the outcome for at least one pupil health outcome. If the ICC estimate for the outcome was provided as a range (e.g., " <0.1 "), the article was excluded.

If there were multiple articles reporting the ICC estimates from the same study (i.e., sibling articles), the article that first presented the ICC (i.e., earliest publication) was designated the index article and used for data extraction.

Articles specifically reporting protocol/designs, process evaluation findings, economic evaluations/cost-effectiveness findings, statistical analysis plans or commentaries were excluded.

5.4.3 Screening and selection

MEDLINE (via Ovid) was searched from inception until 18th October 2021 and all titles and abstracts were downloaded into Endnote 20 [264]. After deduplicating the citations, the remaining titles and abstracts were screened for eligibility by two independent reviewers (KP and OU). Articles were coded (1) if they were thought to be eligible, or (2) if they were not. Once completed, the two Endnote libraries were merged to identify articles where the coding differed. If there was uncertainty about inclusion and consensus could not be reached through discussion, articles were included in full text screening.

PDF versions of the full texts of potentially eligible articles were obtained and screened using Endnote. A coding scheme was developed using Endnote in order to identify reasons for inclusion/exclusion and was piloted on a random sample of 10 articles. The coding scheme for inclusion/exclusion reasons is in Table 5.1.

Table 5.1. Reasons for exclusion at full text screening

Coding scheme	Reason
<i>a</i>	Include – with ICC estimate for health-related outcome
<i>b</i>	Exclude – No ICC estimate provided
<i>c</i>	Exclude – ICC estimate given as a range
<i>d</i>	Exclude – Not a CRT
<i>e</i>	Exclude – Not school-based/school unit not randomised
<i>f</i>	Exclude – Protocol/design article
<i>g</i>	Exclude – No pupil health outcome
<i>h</i>	Exclude – Process evaluation paper
<i>i</i>	Exclude – Commentary
<i>j</i>	Exclude – Mediators/Moderators paper
<i>k</i>	Exclude – Cost-effectiveness/Economic evaluation paper
<i>l</i>	Exclude – Other

In parallel, two independent reviewers (KP and OU) screened articles for inclusion based on the coding method. After the articles had been coded, the two Endnote libraries were merged. Any articles where coding differed were identified and any disagreements that could not be resolved via discussion were sent to a third reviewer (MN) for a decision.

5.4.4 Data extraction

After screening, selected articles were each given a study ID number for data extraction. The variables chosen for data extraction were identified and informed by previous similar reviews of ICCs [71, 105] and the findings and knowledge from the studies presented in Chapters 3 and 4 [117, 122]. Using Microsoft Excel, a bespoke data extraction form was developed and piloted on 20 eligible articles. Modifications were made to the data extraction form following piloting. The variables extracted are listed in Table 5.2.

Table 5.2. Data extracted from included articles

Aspect	Information extracted
<i>Publication details</i>	Author surname, year of publication, title of article, type of study (i.e., definitive or feasibility study).
<i>Setting information</i>	Country in which the study took place (e.g., France), stage of education (e.g., primary, secondary), gender of pupils, age(s) of pupils at baseline.
<i>Study design</i>	Type of cluster unit allocated, cluster unit of ICC/CV estimate.
<i>Sample size information</i>	ICC/CV assumed in the sample size calculation, number of clusters and pupils that provided outcome data, number of classes per school.
<i>Health outcome information</i>	Health area of outcome (e.g., physical activity), outcome description (e.g. amount of moderate-to-vigorous physical activity), outcome type (e.g., continuous, binary), timing (months post-randomisation) at which outcome was measured.
<i>ICC information</i>	ICC/CV of the outcome (and 95% CIs where provided), analysis method used to calculate ICC/CV (e.g., multilevel model [265], marginal model using Generalised Estimating Equations (GEEs) [24]), whether the ICC/CV estimate was pooled across trial arms, whether the ICC/CV estimate was unadjusted or adjusted for prognostic factors, whether the ICC/CV estimate was adjusted for the baseline value of the outcome, whether the ICC/CV was estimated from an analysis of change scores between baseline and follow-up, whether a repeated measures analysis was used to estimate the ICC.

The ICC/CV estimate(s) of one health outcome, measured on school pupils was extracted from each article. This method was used as estimates for multiple outcomes from the same study would likely be correlated and contribute relatively little additional information to the analyses which are focussed on comparing the ICC/CV across different study scenarios. A description and explanation of the criteria used to select the ICC/CV when multiple estimates were reported for a given article are presented in Table 5.3.

Where studies reported both unadjusted and adjusted ICCs, the former was extracted on the basis that this would be of more general use to future researchers who may want to adjust their estimate of the intervention effect for a specific set of prognostic factors. Where the ICC for a given outcome was reported for multiple time points, the ICC for the earliest study wave was extracted, as the ICC estimate would be less likely to be impacted by the intervention. For a similar reason, where the ICC was reported separately for the control and intervention arms the former was chosen. If estimates were reported at multiple levels (e.g., school and class) for the chosen outcome then all were extracted.

Table 5.3. Criteria used to select which intra-cluster correlation coefficient (ICC) or between-cluster coefficient of variation of the outcome (CV) to extract

Aspect	Criteria
<i>Outcome measure</i>	In the first instance, the ICC/CV for the primary health outcome was selected. If there was more than one primary health outcome, the ICC/CV for the first primary outcome presented in the Results section of the paper was selected. If no primary health outcome was declared, the ICC/CV for the health outcome on which the sample size calculation was based was selected. If no primary health outcome was declared and the sample size was not based on a health outcome, the ICC/CV for the first health outcome reported in the Results section of the paper was selected.
<i>Time point at which outcome was measured</i>	In the first instance, the ICC/CV from the baseline time point was selected. If this was not reported, the ICC/CV from the earliest time point of measurement was selected. This was because the estimate would be less likely to be impacted by the intervention.
<i>Unadjusted versus adjusted ICC/CV</i>	If the study presented both unadjusted ICCs/CVs estimates and estimates that are adjusted for prognostic factors, the unadjusted ICC/CV was extracted. This is because the unadjusted estimate would be of more use to future researchers who may want to adjust their estimate of the intervention effect for a specific set of prognostic factors.
<i>Control versus intervention arm</i>	If the ICC/CV was reported separately for the intervention and control arms, the ICC/CV from the control arm was selected. This was because the estimate would be less likely to be impacted by the intervention.

The principal investigator (KP) extracted data from all included studies. A second reviewer (OU) validated this data extraction by scrutinising the accuracy of the data that had been extracted by the principle investigator against the information in the articles. Missing information that was not available in the index articles was sought from protocol or sibling papers identified in the MEDLINE search. If missing information could not be obtained from the protocol or sibling papers it was recorded as 'Not stated'. Authors were not contacted for missing information. Any uncertainty regarding the data extracted was resolved through discussion with another member of the team (MN).

5.4.5 Data coding and classification

Data were extracted exactly as reported in each article. Once the data were extracted for all included texts, some data were coded for ease of analysis.

5.4.5.1 Publication details

Type of study (i.e., definitive or feasibility study) was determined by the title and objectives of the study. Type of outcomes paper being reported (i.e., follow-up outcome paper, secondary data analysis) was determined by the title or reference in the text to use of secondary data.

5.4.5.2 Participants and setting information

Countries in which the study took place were grouped into world regions based on a 7-continent system (Asia, Africa, North America, South America, Antarctica, Europe and Australia) [266]. A category for studies that took place in the UK was also used even though it is not a world region. This was due to large number of studies undertaken in the UK.

The variable 'School level' was first recorded as stated in the article (e.g., primary school, middle school). Some articles did not state a 'School level' but did provide the age (or range of ages) of pupils at baseline. Therefore, 'School level' and 'age of pupils at baseline' were used to categorise studies into 'pre-primary' 'primary' and 'secondary' educational systems according to the UNESCO ISCED system [263].

5.4.5.3 Study design and analytical methods

Number of classes per school, if not provided in the text, was calculated by dividing the number of classes by the number of schools providing outcome data in the study.

5.4.5.4 Sample size information

If the ICC assumed in the sample size calculation was given as a range (e.g., <0.005), it was recorded as such. The ICC/CV assumed in the sample size calculation was sought from protocol or sibling papers if not provided in the article. A note was also made if the ICC/CV assumed in the sample size calculation was for a different outcome from the ICC/CV estimated from the study data.

5.4.5.5 Health outcome information

The health area of the outcome was categorised using the same method described in Chapter 3 (Section 3.4.5.6).

If the measurement timing was reported at baseline, this was recorded as '0'. If repeated measures were used, all time points were recorded.

The number of clusters and pupils providing outcome data were initially determined from the CONSORT flow diagram in the papers. If it was not clear from the CONSORT flow diagram then information was extracted and verified from the main text.

5.4.5.6 ICC information

If the article only provided the components of variance then the ICC was calculated using the proportion of variance definition, as described in Section 5.2 and Table 5.3.

In some articles, it was not certain whether the ICC estimate was from an unadjusted or an analysis that was adjusted for prognostic factors and, therefore, the following categories were used: 'definitely adjusted', 'probably adjusted', 'definitely unadjusted', 'probably unadjusted' and 'unclear'.

5.4.6 Data analysis

A PRISMA flow diagram [125] was used to report the results of the screening process. Study characteristics were summarised using medians and interquartile ranges for continuous variables, and numbers and percentages for categorical variables. A histogram was used to describe the distribution of the ICCs. Mann-Whitney and Kruskal-Wallis tests were used to compare the ICC estimate across subgroups. Statistical analyses were undertaken using Stata 17 software [227].

5.5 Results

3632 articles were identified through searching MEDLINE. After deduplication, 3618 articles were title and abstract screened against inclusion criteria. 1590 were included in the full text screening stage and 246 articles were identified as eligible for inclusion in the review. Agreement between reviewers on which articles should be included was 99.6% (245/246). The PRISMA flow diagram is presented in Figure 5.1.

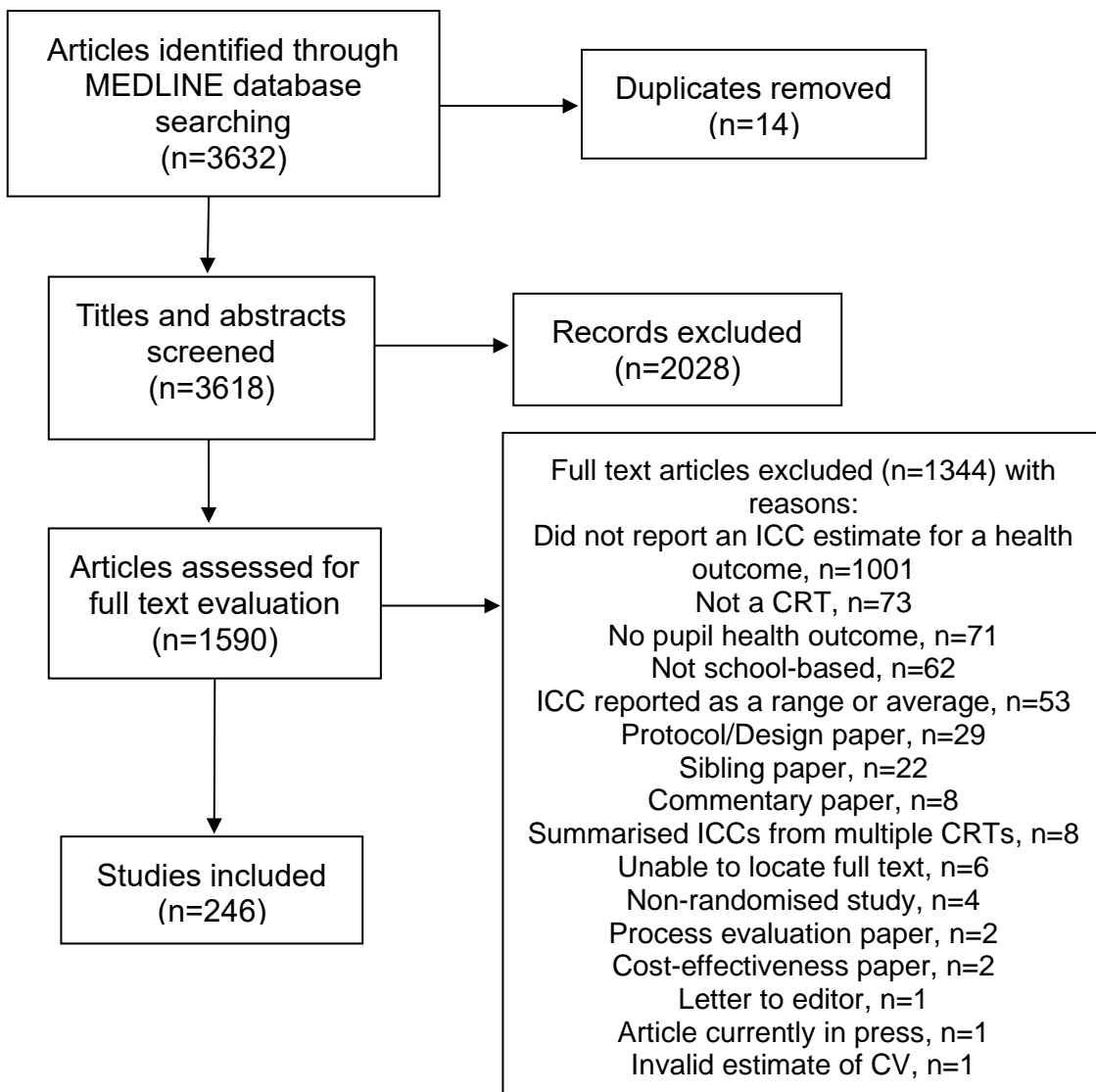


Fig. 5.1. PRISMA flow diagram summarising the results of the literature search and screening for eligibility

5.5.2 Publication characteristics

The rate of publication of papers reporting ICC estimates of pupil health outcomes in school-based CRTs has increased since the first included publication in 1999. Of the 246 articles that the search identified, 44 articles were published between 1999 and 2010, compared to 25 in 2021 alone. The majority of articles included were findings from definitive trials (n=226; 91.9%), while the remaining 20 (8.1%) articles summarised the findings from feasibility studies. 238 (96.7%) articles reported follow-up outcomes and the remaining 8 (3.3%) articles published the results of secondary data analyses.

5.5.3 Study features and design characteristics

Of the 246 eligible articles, 57 (23.2%) articles summarised studies that took place in Europe, with a further 44 (17.9%) taking place in the UK. 53 (21.5%) were undertaken in the USA or Canada, 33 (13.4%) in Australia and New Zealand, 23 (9.3%) in Asia, 19 (7.7%) in Central America and South America, and 17 (6.9%) in Africa.

Most commonly studies took place only in secondary educational settings (n=97; 39.4%). Eighty-eight (35.8%) articles reported studies that took place only in primary educational settings and 16 (6.5%) articles reported studies that took place exclusively in pre-primary educational settings. Almost all studies (n=227; 92.3%) included both male and female pupils.

In most of the articles schools were randomised as the cluster unit (n=220; 89.4%); classes/classrooms were randomised in 23 (9.3%) articles; and school buildings [267], student groups [268] and year groups [202] were randomised in one article each.

A range of different health outcome areas were spanned by the included articles, the most frequent being social emotional functioning (n=53; 21.5%), physical activity (n=34; 13.8%), adiposity (n=28; 11.4%) and smoking (n=21; 8.5%). In 163 (66.3%) articles, the outcome type was continuous, in 76 (30.9%) it was binary, in 6 (2.4%) it was count/rate data, and in 1 (0.4%) it was ordinal. Just over half the outcomes were reported by the school pupils themselves (n=139; 56.5%). An

objective measuring device (e.g., accelerometer, weighing scales) was used in 54 (22.0%) articles.

The median (IQR; range) assumed school-level ICC in the sample size calculation was 0.04 (0.02 to 0.09; 0.001 to 0.3) based on the 106 articles that provided these data. The median (IQR; range) assumed class-level ICC was 0.055 (0.03 to 0.1; 0.01 to 0.63) based on the 14 articles that provided these data. Four articles provided the between-cluster coefficient of variation (CV) of the outcome assumed in the sample size calculation; these were 0.1 [269], 0.15 [270], 0.2 [271], and 0.25 [272].

Altogether, 260 ICC estimates were identified and extracted from the 246 articles: 210 at school level; 46 at class/classroom level; and 1 each at the levels of school building [267], student group [268], year group [273] and sports-team [274]. Only 34 (13.8%) articles provided 95% confidence intervals for the ICC estimate. Forty-five (17.3%) ICCs were estimated using the baseline measurement of the outcome and 2 ICCs were for the control arm only [149, 275]. ICCs were extracted for 172 continuous outcomes, 78 binary outcomes, 6 count/rate outcomes and 2 ordinal outcomes. For 2 ICCs the outcome type was unclear.

Of the studies that reported school-level ICC estimates, the median (IQR) number of clusters and pupils were 22 (12 to 40) and 1110 (441 to 2443), respectively. Of the studies reporting class-level ICC estimates, the median (IQR) number of clusters and pupils were 47 (25 to 88) and 647.5 (288 to 1477), respectively. Sixty-eight articles provided enough information to determine the number of classes per school; in those studies, the median (IQR; range) number of classes per school was 3.4 (2 to 5.3; 1 to 61.3).

Of the 246 articles, 180 (73.2%) used mixed effects (“multilevel”) models to calculate the ICC estimate, 21 (8.5%) used marginal models using GEEs, 6 (2.4%) used random effects analysis of variance [150, 170, 175, 276-278], 4 (1.6%) used latent growth models [279-282], 1 (0.4%) used Flesiss-Cuzick estimators [283], 1 (0.4%) stated the ICC was “calculated from empirical design estimates” [284], and 1 (0.4%) used “an appropriate formulae from Hayes and Moulton [5]” [285]. In 32 (13.0%) articles, the analysis method used to calculate the ICC estimate was unclear.

In 71 (28.9%) articles, there was sufficient information to determine that the published ICC estimates were adjusted for prognostic factors. In 83 (33.7%) articles, there was sufficient information to determine that the published ICC estimates were not adjusted for prognostic factors.

The study features and design characteristics are summarised in Table 5.4.

Table 5.4. Summary of study features and design characteristics (N=246)

Characteristic	N	Statistic
<i>Region</i>	246	
Europe, n (%)		57 (23.2)
USA and Canada, n (%)		53 (21.5)
UK, n (%)		44 (17.9)
Australia and New Zealand, n (%)		33 (13.4)
Asia, n (%)		23 (9.3)
Central and South America, n (%)		19 (7.7)
Africa, n (%)		17 (6.9)
<i>Education level</i>	246	
Pre-primary Educational System, n (%)		16 (6.5)
Primary Educational System, n (%)		88 (35.8)
Secondary Educational System, n (%)		97 (39.4)
Pre-primary and Primary Educational Systems, n (%)		6 (2.4)
Primary and Secondary Educational System, n (%)		36 (14.6)
Pre-primary, Primary and Secondary Educational System, n (%)		3 (1.2)
<i>Gender of pupils on which outcome was measured</i>	246	
Male and female, n (%)		227 (92.3)
Female, n (%)		15 (6.1)
Male, n (%)		4 (1.6)
<i>Cluster unit allocated</i>	246	
Schools, n (%)		220 (89.4)
Classes/classrooms, n (%)		23 (9.3)
Year groups, n (%)		1 (0.4)
Student groups, n (%)		1 (0.4)

Characteristic	N	Statistic
School buildings, n (%)		1 (0.4)
Health area of outcome	246	
Social emotional functioning, n (%)		53 (21.5)
Physical activity, n (%)		34 (13.4)
Adiposity, n (%)		28 (11.4)
Smoking, n (%)		21 (8.5)
General health, n (%)		13 (5.3)
Alcohol misuse, n (%)		12 (4.9)
Sexual health and obstetrics, n (%)		11 (4.5)
Dental/oral health, n (%)		10 (4.1)
Infectious disease, n (%)		10 (4.1)
Nutrition, n (%)		10 (4.1)
Violence, n (%)		8 (3.3)
Injury, n (%)		6 (2.4)
Skin cancer		5 (2.0)
Safety, n (%)		4 (1.6)
Pain, n (%)		3 (1.2)
Anaemia, n (%)		2 (0.8)
Hearing, n (%)		2 (0.8)
Physical activity/nutrition, n (%)		2 (0.8)
Substance misuse, n (%)		2 (0.8)
Allergy, n (%)		1 (0.4)
Biomarkers, n (%)		1 (0.4)
Cancer, n (%)		1 (0.4)
Dating violence, n (%)		1 (0.4)
Epilepsy, n (%)		1 (0.4)
Heart disease, n (%)		1 (0.4)
Motor skills, n (%)		1 (0.4)

Characteristic	N	Statistic
Ophthalmology, n (%)		1 (0.4)
Organ donation, n (%)		1 (0.4)
Speech and language, n (%)		1 (0.4)
Reporter of the outcome	246	
Pupil, n (%)		139 (56.5)
Objective measuring device, n (%)		54 (22.0)
Researchers, n (%)		19 (7.7)
Teachers, n (%)		9 (3.6)
Health professionals, n (%)		6 (2.4)
Parents/carers, n (%)		6 (2.4)
Routine data, n (%)		5 (2.0)
Laboratory tests, n (%)		2 (0.8)
Other ¹ , n (%)		6 (2.4)
Outcome type	246	
Continuous, n (%)		163 (66.3)
Binary, n (%)		76 (30.9)
Count/rate, n (%)		6 (2.4)
Ordinal, n (%)		1 (0.4)
School-level ICC assumed in sample size calculation, median (IQR; range)	109	0.040 (0.020 to 0.090; 0.001 to 0.250)
Class-level ICC assumed in sample size calculation, median (IQR; range)	14	0.055 (0.030 to 0.100; 0.010 to 0.630)
Studies providing a school-level ICC estimate		

Characteristic	N	Statistic
Number of school clusters, median (IQR; range)	207	22 (14 to 40; 3 to 418)
Number of pupils, median (IQR; range)	210	1110 (441 to 2443; 34 to 92770)
<i>Studies providing a class-level ICC estimate</i>		
Number of class clusters, median (IQR; range)	41	47 (25 to 88; 4 to 385)
Number of pupils, median (IQR; range)	46	647.5 (288 to 1477; 75 to 4866)
<i>Average number of classes per school, median (IQR; range)</i>	68	3.4 (2.0 to 5.3; 1.0 to 61.3)
<i>Analysis method used to calculate the ICC estimate</i>		
Mixed effects (“multilevel”) models	246	180 (73.2)
Marginal models using GEEs		21 (8.5)
Random effects analysis of variance		6 (2.4)
Latent growth models		4 (1.6)
Other ²		3 (1.2)
Unclear		32 (13.0)

¹ Other includes certified athletic trainers, medical students, both parents and kindergarten doctors, pupil and parents/carers, sports team manager/physiotherapists, and “trained observers”.

² Included one article that used Fleiss-Cuzick estimators, one article that stated the ICC was “calculated from empirical design estimates”, and one article that used “an appropriate formulae from Hayes and Moulton[5]”.

5.5.4 Summary of ICC estimates

All ICC estimates are reported with articles referenced in Appendix 9. School- and class-level ICCs are reported side-by-side for the 14 studies that reported at both those levels in Appendix 10.

For the 210 articles that provided ICC estimates at the school level, the median (IQR; range) ICC was 0.031 (0.011 to 0.08; 0 to 0.47). Almost a quarter ($n=51$; 24.3%) of these estimates were less than or equal to 0.01, and just under two-thirds of estimates ($n=135$; 64.3%) were less than or equal to 0.05. The mean (SD) school-level ICC was 0.060 (0.076). Figure 5.2 summarises the distribution of the school-level ICC estimates. Both the beta distribution (with shape parameters 0.77 and 11.0) and the exponential distribution provided a good fit to the school-level ICC estimates.

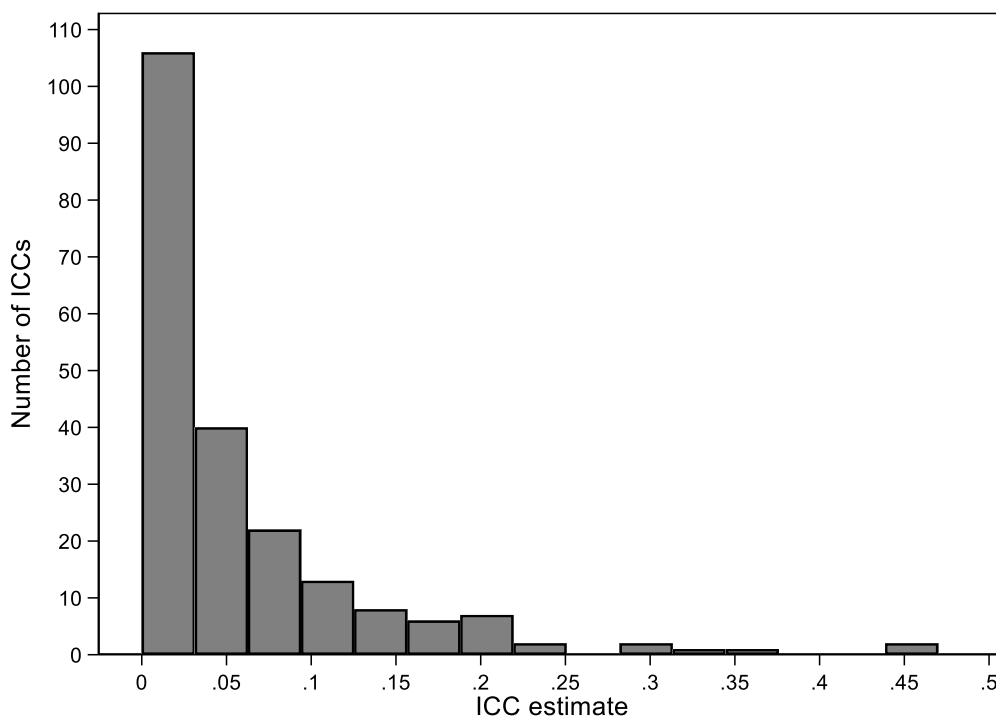


Fig. 5.2. The distribution of school-level intra-cluster correlation coefficients (ICCs) observed in school-based CRTs ($N=210$)

For the 46 articles that provided ICC estimates at the class level, the median (IQR; range) ICC was 0.063 (0.024 to 0.1; -0.009 to 0.262). Only one negative ICC estimate was reported, which was at the class level [175]. Figure 5.3 summarises the distribution of the class-level ICC estimates.

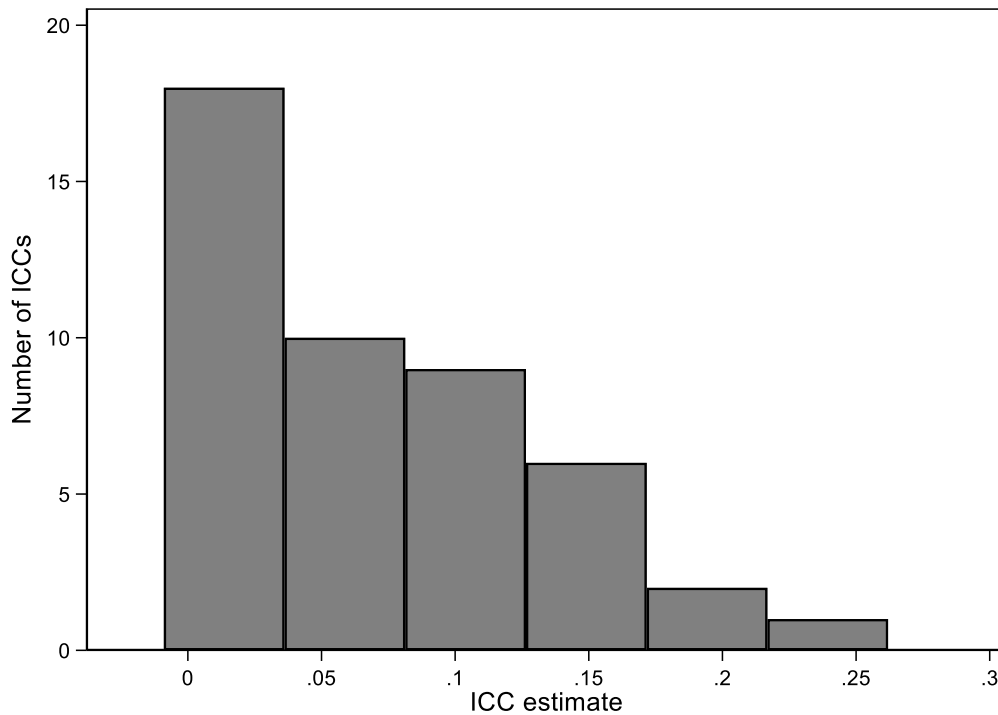


Fig. 5.3. The distribution of class-level intra-cluster correlation coefficients (ICCs) observed in school-based CRTs (N=46)

In Table 5.5, the median (IQR; range) school-level ICC is reported by categories defined by world region, health outcome area (for the 10 most frequently reported areas) and educational stage. Figure 5.4 describes the distributions of these ICC estimates using dot plots. Tests of significance showed little evidence of differences across subgroups defined by region, outcome area and education stage. The distribution of ICC estimates showed a fair amount of overlap across subgroups.

Regarding world region, the largest median ICC estimates were for Asia, Central and South America, and Africa (all 0.05). The smallest median ICC was for the estimates from Australia and New Zealand (0.02). The distribution of ICCs from the USA and Canada (median 0.033 and 75% of estimates being lower than 0.073) is similar to published findings from previous summaries of USA-based estimates [69, 96-103, 105-109]. There was reasonable overlap with the distributions in the other regions apart from Australia/New Zealand, for which the median and upper quartile were notably lower.

For the 10 most common health outcome area, the largest median ICC estimate was for studies in nutrition (0.06), and the smallest median ICC estimate was for

studies in general health (0.025). The distributions of school-level ICC estimates for adiposity, physical activity and general health were generally low compared with the 7 other most common health outcome areas. For two specific outcomes there were more than 10 estimates of the school-level ICC: Body mass index (BMI) was reported in 17 articles, across which the median (IQR) school-level ICC was 0.021 (0.015 to 0.04); Moderate-to-vigorous physical activity (MVPA) was reported in 11 articles, across which the median (IQR) school-level ICC was 0.018 (0.01 to 0.057).

For educational stage, the largest median ICC estimate was for those studies that took place in pre-primary educational settings (0.048), but only 13 ICC estimates were included. The median ICC estimate decreased in size for later educational stages (i.e., from pre-primary to primary to secondary educational settings), although there was little evidence of a true difference ($p=0.40$). The overall distribution of ICCs was also lower for the pre-primary stage compared to the later stages of education.

Table 5.5. Median (IQR; range) school-level intra-cluster correlation coefficient (ICC) by world region, outcome area and education stage (N=210)

Characteristic	N	Median ICC (IQR; range)	p-value
<i>Region</i>			0.26
Europe ¹	45	0.04 (0.014 to 0.08; 0 to 0.47)	
USA and Canada	44	0.033 (0.010 to 0.073; 0 to 0.286)	
UK ²	40	0.029 (0.01 to 0.106; 0 to 0.45)	
Australia and New Zealand	27	0.02 (0.01 to 0.03; 0 to 0.16)	
Asia ³	21	0.05 (0.013 to 0.118; 0 to 0.31)	
Central and South America ⁴	17	0.05 (0.016 to 0.09; 0.0001 to 0.36)	
Africa ⁵	16	0.05 (0.018 to 0.127; 0.0005 to 0.21)	
<i>Health outcome area</i>			0.76
Social emotional functioning ⁶	39	0.05 (0.02 to 0.097; 0 to 0.217)	
Physical activity	30	0.035 (0.013 to 0.059; 0 to 0.19)	
Adiposity	26	0.027 (0.014 to 0.041; 0.004 to 0.19)	
Smoking	19	0.055 (0.017 to 0.11; 0 to 0.286)	
Alcohol use	10	0.055 (0.02 to 0.098; 0 to 0.121)	
Dental/oral health	10	0.051 (0.027 to 0.119; 0 to 0.31)	
General health	10	0.025 (0.014 to 0.045; 0.001 to 0.18)	

Characteristic	N	Median ICC (IQR; range)	p-value
Infectious disease	9	0.042 (0.004 to 0.070; 0.0001 to 0.21)	
Nutrition	8	0.06 (0.010 to 0.097; 0 to 0.36)	
Violence	8	0.048 (0.014 to 0.085; 0.002 to 0.13)	
<i>Education stage</i>			0.40
Pre-primary education only ⁷	13	0.048 (0.03 to 0.063; 0 to 0.097)	
Primary education only ⁸	81	0.04 (0.013 to 0.094; 0 to 0.47)	
Secondary education only ⁹	81	0.03 (0.01 to 0.07; 0 to 0.31)	

¹ Included countries stated as: Finland, The Netherlands, Denmark, Belgium, Norway, Germany, Estonia, Poland, Spain, Switzerland, Cyprus, Italy, Greece, Hungary, Sweden, Austria, Majorca, France, Ireland, Romania, Slovenia.

² Included countries stated as: England, Northern Ireland, Scotland, Wales.

³ Included countries stated as: Israel, China, Iran, India, Japan, Bangladesh, Nepal, Taiwan, Peru, Pakistan, Thailand, Indonesia, Hong Kong.

⁴ Included countries stated as: Jamaica, Brazil, Ecuador, Chile, Haiti, Belize.

⁵ Included countries stated as: Uganda, South Africa, Kenya, Tanzania, Burundi.

⁶ Includes mental health, behaviour, neurodiversity, well-being, quality of life, bullying, social and emotional learning, body image and self-esteem., Ireland, Romania, Slovenia.

⁷ Includes pre-schools, kindergartens, educational childcare centres and head-start schools

⁸ Includes elementary schools, middle schools (Grade 6)

⁹ Includes secondary schools, middle schools (>= Grade 7), high schools, junior high schools, lower secondary schools, higher/upper secondary schools, vocational schools, intermediate vocational schools, secondary-level vocational schools and continuation schools.

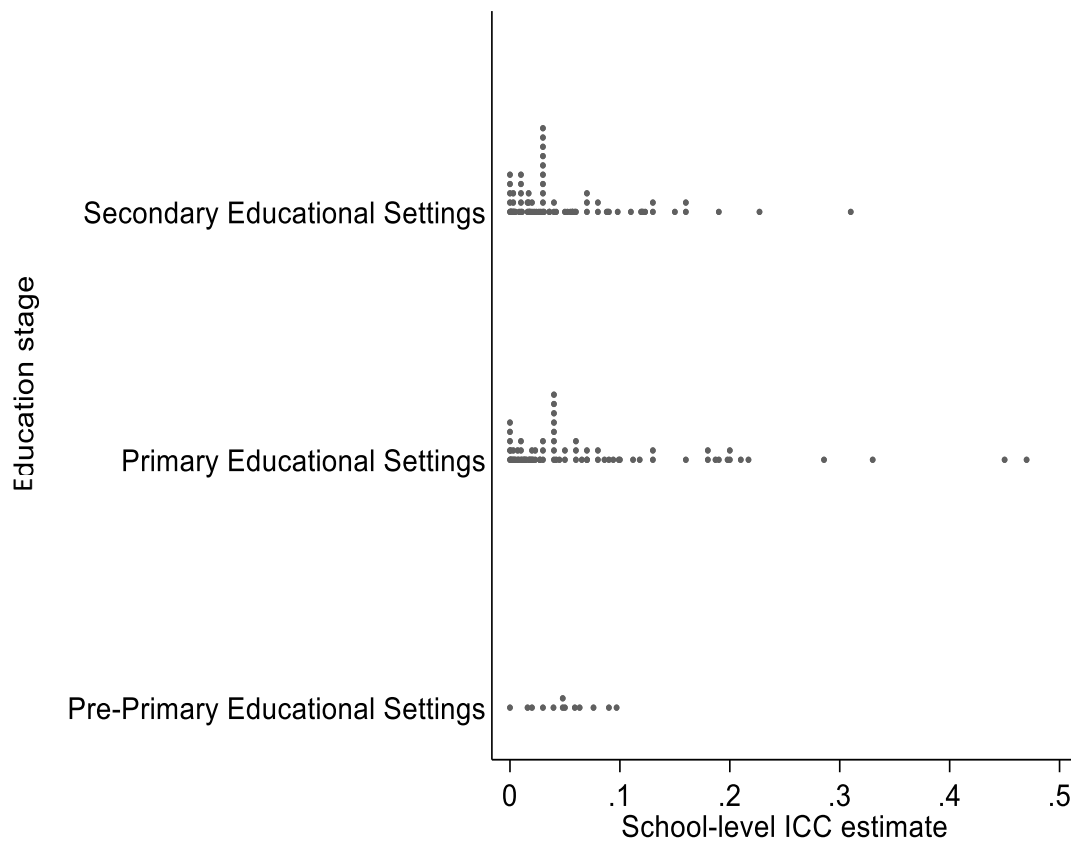


Fig. 5.4. Dot plots of school-level intra-cluster correlation coefficients (ICCs) by region, outcome area and education stage

The median (IQR) school-level ICC estimate was larger for definitive studies (0.038 (0.016 to 0.08); N=192) than feasibility studies (0.01 (0.0005 to 0.04); N=18) ($p=0.005$).

The median (IQR) school-level ICC was larger for continuous outcomes (N=135) than binary outcomes (N=68) although there was little evidence of a true difference in the distributions (0.04 (0.014 to 0.08) versus 0.025 (0.008 to 0.08); $p=0.21$). Summaries of the school-level ICCs are reported separately for continuous and binary outcomes in Table 5.6.

Table 5.6. Median (IQR; range) school-level intra-cluster correlation coefficient (ICC) by region, health outcome area and education stage summarised separately for continuous and binary outcomes (N=210)

Characteristic	Continuous outcomes		Binary outcomes	
	N	median ICC (IQR; range)	N	median ICC (IQR; range)
Region				
Europe ¹	25	0.04 (0.016 to 0.080; 0 to 0.33)	18	0.03 (0.011 to 0.08; 0 to 0.47)
USA and Canada	29	0.04 (0.02 to 0.076; 0.0005 to 0.16)	13	0.02 (0.002 to 0.06; 0 to 0.286)
UK ²	27	0.030 (0.01 to 0.12; 0 to 0.217)	13	0.027 (0.01 to 0.04; 0.003 to 0.45)
Australia and New Zealand	20	0.028 (0.01 to 0.055; 0 to 0.16)	6	0.014 (0.004 to 0.02; 0 to 0.03)
Asia ³	16	0.05 (0.012 to 0.102; 0 to 0.31)	4	0.111 (0.02 to 0.219; 0.017 to 0.24)
Central and South America ⁴	12	0.065 (0.027 to 0.114; 0.015 to 0.36)	4	0.026 (0.006 to 0.061; 0.0016 to 0.08)
Africa ⁵	6	0.038 (0.02 to 0.07; 0.0005 to 0.18)	10	0.065 (0.016 to 0.13; 0.007 to 0.21)
Health outcome area				

Characteristic	Continuous outcomes		Binary outcomes	
	N	median ICC (IQR; range)	N	median ICC (IQR; range)
Social emotional functioning	34	0.055 (0.02 to 0.126; 0 to 0.217)	4	0.020 (0.011 to 0.024; 0.003 to 0.028)
Physical activity	29	0.03 (0.013 to 0.059; 0 to 0.19)	1	0.040
Adiposity	24	0.03 (0.015 to 0.045; 0.004 to 0.19)	2	0.017
Smoking	2	0.04	16	0.043 (0.018 to 0.102; 0 to 0.286)
Alcohol use	3	0.03	7	0.088 (0.02 to 0.112; 0 to 0.121)
Dental/oral health	9	0.052 (0.03 to 0.119; 0 to 0.31)	1	0.027
General health	6	0.044 (0.02 to 0.063; 0.001 to 0.18)	3	0.014
Infectious disease	0	-	6	0.056 (0.04 to 0.13; 0.004 to 0.21)
Nutrition	8	0.06 (0.010 to 0.097; 0 to 0.36)	0	-
Violence	3	0.007	5	0.06 (0.06 to 0.109; 0.02 to 0.13)

Characteristic	Continuous outcomes		Binary outcomes	
	N	median ICC (IQR; range)	N	median ICC (IQR; range)
<i>Education stage</i>				
Pre-primary education only ⁷	13	0.048 (0.03 to 0.063; 0 to 0.097)	0	-
Primary education only ⁸	52	0.04 (0.014 to 0.097; 0 to 0.33)	26	0.04 (0.01 to 0.112; 0 to 0.47)
Secondary education only ⁹	49	0.036 (0.017 to 0.07; 0 to 0.31)	30	0.018 (0.004 to 0.055; 0 to 0.227)

¹ Included countries stated as: Finland, The Netherlands, Denmark, Belgium, Norway, Germany, Estonia, Poland, Spain, Switzerland, Cyprus, Italy, Greece, Hungary, Sweden, Austria, Majorca, France, Ireland, Romania, Slovenia.

² Included countries stated as: England, Northern Ireland, Scotland, Wales.

³ Included countries stated as: Israel, China, Iran, India, Japan, Bangladesh, Nepal, Taiwan, Peru, Pakistan, Thailand, Indonesia, Hong Kong.

⁴ Included countries stated as: Jamaica, Brazil, Ecuador, Chile, Haiti, Belize.

⁵ Included countries stated as: Uganda, South Africa, Kenya, Tanzania, Burundi.

⁶ Includes mental health, behaviour, neurodiversity, well-being, quality of life, bullying, social and emotional learning, body image and self-esteem., Ireland, Romania, Slovenia.

⁷ Includes pre-schools, kindergartens, educational childcare centres and head-start schools

⁸ Includes elementary schools, middle schools (Grade 6)

⁹ Includes secondary schools, middle schools (\geq Grade 7), high schools, junior high schools, lower secondary schools, higher/upper secondary schools, vocational schools, intermediate vocational schools, secondary-level vocational schools and continuation schools.

For continuous outcomes, the median (IQR) school-level ICC was higher for studies that adjusted for the baseline of the outcome at the pupil level (N=35) (0.045 (0.013 to 0.09)) compared with those that did not (N=95) (0.040 (0.016 to 0.07)). However, there was little evidence of a true difference ($p=0.50$). Further to this, the size of the median school-level ICC (0.04) was the same regardless of whether studies with continuous outcomes analysed change scores (N=11) or not (N=124) ($p=0.37$).

The median (IQR) school-level ICC for studies (N=37) that estimated the ICC from a repeated measures analysis was 0.027 (0.01 to 0.057). The median (IQR) school-level ICC those that did not use a repeated measures analysis (N=173) was 0.036 (0.013 to 0.088). Despite the median ICC estimate being larger for the latter there was little evidence of a true difference ($p=0.15$).

Lastly, for binary outcomes, the median (IQR) school-level ICC estimate for studies that used mixed effects (multilevel) logistic regression to estimate the ICC on the logistic scale (N=42) was 0.049 (0.014 to 0.109), which was larger than the 14 studies that used other methods to estimate it on the proportions (natural) scale (median (IQR) ICC was 0.014 (0.007 to 0.023)), although there was only weak evidence of a true difference between the analysis approaches ($p=0.08$). The higher median for ICCs estimated on the logistic scale is in keeping with the methodological literature on the ICC [10]. Figure 5.5 summarises the relationship between ICC estimates and the prevalence for binary outcomes. The size of the ICC increases as the prevalence reaches 50%.

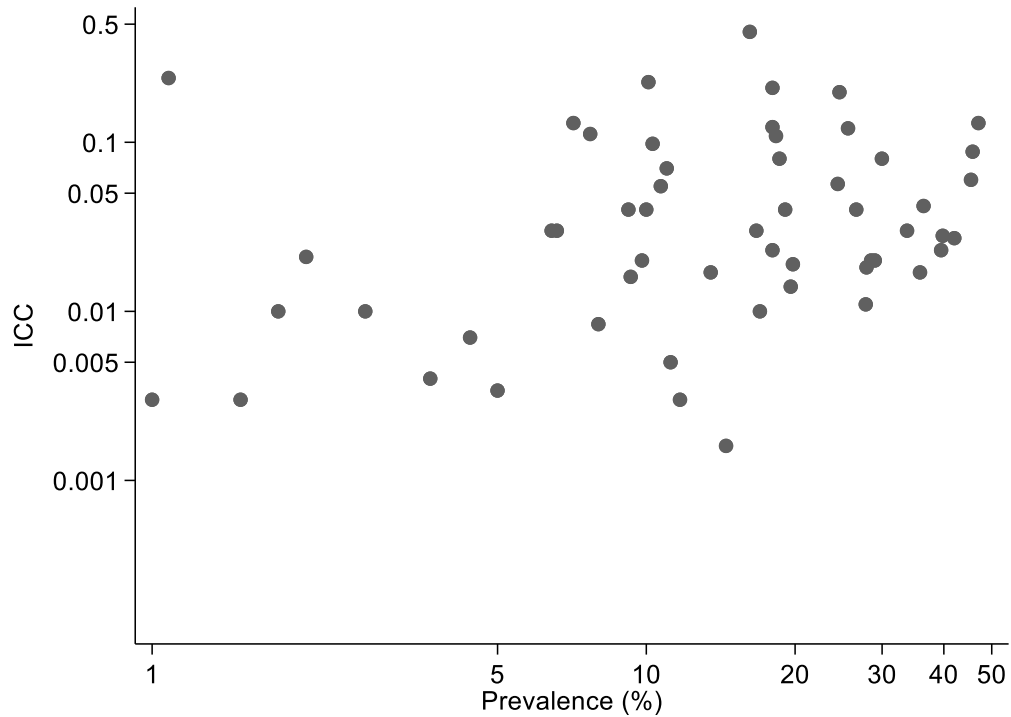


Fig. 5.5. Scatterplot of school-level intra-cluster correlation coefficient (ICC) estimates versus the prevalence for binary outcomes: both variables plotted on the logarithmic scale (N = 55) *

* The ICC (y axis) and binary outcome prevalence (x axis) are both plotted on the logarithmic scale. In total, 62 studies reported a school-level ICC estimate and the prevalence of the binary outcome. Seven studies are not included in the graph as the ICC estimate was zero for 6 studies and the prevalence was zero for one study.

5.6 Discussion

This chapter reports and summarises 260 ICC estimates from 246 articles reporting the findings from school-based CRTs worldwide for a range of different pupil health outcomes. Despite there being few clear patterns regarding the relationship of the ICC with aspects of the design and analysis, the summary has identified a number of results that are worthy of reflection. The summary of the ICC estimates across different study features was characterised by overlap in the distributions. However, although apparently small, the differences in median ICC estimates across subgroups are large in terms of the impact they would have on the sample size requirement for a CRT. Furthermore, differences between subgroups defined by study design and context may have not been detected due to reduced noise caused by imprecise ICC estimates.

An ICC estimate from a single CRT may be considered poorly generalisable due to the clinical and methodological heterogeneity across CRTs [108]. Therefore, it has been recommended that researchers use the distribution of ICC estimates from a range of different studies in order to model the sensitivity of sample size calculations [3, 26, 286]. Similar to the findings in this chapter, previous studies randomising primary care clusters have demonstrated that ICC estimates for health outcomes are well described by the beta distribution [26, 287]. Information regarding the distribution of ICC estimates is useful for constructing informative priors when using a Bayesian framework to incorporate uncertainty concerning the ICC in sample size calculations for and analysis of the intervention effect in school-based CRTs [121, 288].

A quarter of included studies were undertaken in Europe ($n=58$; 23.4%) (rising to 41.1% when including the UK), which is similar to a systematic review that found the greatest proportion (29%) of CRTs in children were undertaken in Europe (including the UK) [53]. The large proportion of studies in high-income regions, such UK, Europe and the US, may also be reflected by the type of health areas identified in the current study. The study found that the distribution of school-level ICCs worldwide was broadly similar to previously reported summaries of school-based ICCs for pupil health outcomes, the majority of which are from the US [69, 96-103, 105-109]. Most estimates were less than 0.05 and few were greater than 0.1.

The median ICC for pupil health outcomes was 0.031 at school level and 0.063 at class level. The difference in ICC size between these levels is expected as there is a greater chance for pupils to interact within classes than between classes within the same school. Additionally, it has been reported that ICCs are generally larger when the natural cluster size is smaller [101, 261].

The median ICC for feasibility CRTs was noticeably smaller than that of definitive CRTs. A reason for this could be that the schools recruited to feasibility CRTs may be less representative of the wider population of schools that are recruited in definitive trials [3](p180/181). There was little evidence of relationships of the ICC with health outcome area and education level. This contrasts with a suggestion that ICC estimates informing sample sizes should be outcome specific [71]. Also, a previous summary of ICC estimates for educational outcomes from school-based CRTs showed a tendency for ICCs to be larger for lower grades [88], but this finding was less clear for health outcomes in the current review.

For many ICC estimates, it was unclear whether the analysis used to calculate them had adjusted for prognostic factors. This is important as adjusted ICCs may often be smaller than unadjusted ICCs [88]. Additionally, it was often unclear whether the ICC had been adjusted for cluster-level or individual-level predictors. Adjusting for cluster-level predictors may reduce the size of the ICC estimate but adjusting for individual-level predictors may reduce *or* increase the ICC depending on the extent to which the between- and within-cluster components of variance are accounted for by the adjustment.

The analysis method used to estimate the ICC can impact on its size. For example, an ICC estimate from a repeated measures analysis which includes outcome data from across all study waves in a single analysis will not necessarily be the same as the ICC for an outcome at a specific study wave. In studies with repeated measures, the correlation between observations in the same cluster from *different* waves may be smaller than the correlation between observations in the same cluster at the *same* study wave [286, 289]. This study found little evidence that studies using repeated measures analysis resulted in smaller school-level ICCs than those that do not. As for other comparisons in this chapter,

the lack of a statistically significant difference may be due to confounding with other study-specific characteristics

5.7 Strengths and limitations

This study collated and summarised ICCs for different pupil health outcomes from school-based CRTs worldwide and is the first study of its kind. The summary used a systematic searching approach with a predefined strategy to identify CRTs that randomised school-related units. The protocol was publicly available prior to conducting the study (registered with PROSPERO (CRD42021268782)). Screening and data extraction were undertaken by two independent reviewers. The study did not aim to be comprehensive, but, by using a systematic searching method, the study is reproducible.

A potential limitation of the work was the small number of articles with specific characteristics, which limited the ability to detect differences across subgroups in ICC estimates. Two hundred and four-six CRTs was a not sufficiently large sample size to describe the ICC within different combinations of categories of the study design characteristics. For example, only one combination of region and health outcome area provided at least 10 school-level ICC estimates. Also, partially due to the small sample size, countries were grouped into regions which may have obscured some differences that might otherwise have been found between individual countries. Data from a European-based survey have suggested that ICC estimates assumed in sample size calculations for school-based trials should be country-specific and outcome-specific [71]. With the number of school-based CRTs publishing ICC continuing to increase, more ICC estimates will be available to enable the examination of ICC patterns in relation to key study characteristics.

A potential limitation was the decision to use only the MEDLINE database. Although findings from a previous systematic review of similar studies indicated that few additional studies would have been found by searching other databases (specifically, EMBASE, DARE, PsychINFO and ERIC) [117], further articles may have been found by searching the grey literature. Additionally, some older eligible articles may have been missed because their titles and abstracts did not refer to using a cluster design.

It was decided to not extract multiple ICCs from the same study when they were provided although this would have increased the sample size. This was to avoid a scenario where a small number of studies that reported many ICCs had a disproportionate impact on the observed distribution of ICCs. Given that it is expected that ICCs from the same study would be similar to each other, the inclusion of multiple estimates would have added little extra information to the analyses.

5.8 Implications

The study has a number of important implications regarding the planning of future school-based CRTs. First, the study provides 260 ICC estimates from school-related clusters for use in future sample size calculations. Their distribution indicates a range of plausible values to aid researchers when calculating the sample size for future school-based CRTs. Although the summary found little evidence of a relationship between the ICC and most design/analysis characteristics, researchers should still seek to identify ICC estimates from studies that are similar to the study they plan to undertake. This will help to minimise the chance of an inaccurate sample size calculation. Further research is needed to investigate relationships between the ICC and study characteristics as the work in this study was limited by the number of studies included and the fact that relationships may have been obscured by confounding characteristics at the study level.

5.9 Conclusions

This chapter reported 260 ICC estimates for pupil health outcomes from studies spanning different world regions, health outcome areas and educational settings, and summarised their distribution. The findings are an invaluable resource to researchers for calculating sample size for future school-based CRTs. The study found that ICC estimates had a similar distribution to previously published summaries of ICCs from studies based in the United States. Improved reporting of the ICC in CRTs, in line with CONSORT statement for CRTs [57], will increase the pool of data that can be used to explore the distribution of ICC estimates and the factors that influence their size in greater depth.

5.10 Chapter summary

This chapter presented findings from a study collating and summarising estimates of the intra-cluster correlation coefficient for pupil health outcomes from school-based CRTs. Chapter 6 reports on a secondary data analysis using raw data from five UK school-based CRTs to investigate patterns in the size of the ICC of social emotional functioning outcomes.

Page intentionally left blank

Chapter 6: Estimating intra-cluster correlation coefficients and components of variance for social emotional functioning outcomes of pupils in school-based cluster randomised trials

6.1 Summary

This chapter explains the motivation for this secondary data analysis, which estimates intra-cluster correlation coefficients (ICCs) from CRTs that evaluated interventions for improving social emotional functioning outcomes on pupils. The chapter then states the aims and objectives of the study, describes the methods used to analyse the datasets and reports the results of these analyses. It concludes by discussing the findings, implications, strengths and limitations and areas identified for further research.

6.2 Background

Schools are recognised for the role they can play in the promotion of health in children and young people [50, 55, 115]. Recently, there has been an increased focus on improving social emotional functioning in children and young people through intervening in the school setting [290-292]. Social emotional functioning represents the capacity to understand, experience, express, and manage emotions and to develop meaningful relationships with others [124]. It encompasses concepts including, but not restricted to, mental health, behaviour, well-being, emotional challenges, bullying, neurodiversity and self-esteem. As approximately half of adult mental disorders have their onset during adolescence [293], schools provide an ideal setting in which to promote, prevent and intervene to support good social emotional health during the key developmental years of a young person's life.

Given the time that children and adolescents spend in school and the convenience of recruiting pupils and delivering interventions in this setting, the CRT design is increasingly used to evaluate the impact of such interventions on social emotional functioning outcomes on pupils [145, 146, 153, 161, 173, 202, 203, 294]. Authors have highlighted the impact that the school environment can have on social emotional health [295, 296], and interventions have been designed

to improve such outcomes in the school setting. Many of these interventions are, by their nature, delivered at the cluster level. As reported in Chapter 3, 83% (53/64) of school-based CRTs had at least one component of the intervention that had to be administered to entire clusters [117]; this included 13 of the 15 trials that evaluated the impact of social emotional functioning interventions on pupil health outcomes.

Despite the increase in publications in this area, as reported in Chapters 3 and 5 [117, 123], there is a relative lack of information regarding the ICC and components of variance for use in sample size calculations for school-based CRTs with social emotional functioning outcomes. The ICC quantifies the similarity of observations on individuals within the same cluster, and can take values between 0 and 1. The larger the ICC is, the greater the similarity between individuals within clusters, or, equivalently, the greater the difference between individuals in different clusters [10]. Articles have been written specifically to collate the ICCs for outcomes in health areas such as substance use [61, 71, 97-104], nutrition [105-107] and physical activity [61, 107-109]. This is also true for social emotional functioning outcomes [61, 69, 71, 96, 259], but these articles largely report data from studies undertaken in the US. Previous literature has suggested that ICCs for social emotional functioning outcomes range from around 0 to 0.1 [68, 71, 123], but this is too wide a range to inform sample size calculations given how sensitive the design effect (DE) is to the assumed ICC, especially when large numbers of participants are sampled from each cluster. Additionally, the systematic review presented in Chapter 3 found that less than half ($n=29/64$; 45%) of the school-based CRTs in the UK reported the ICC of the primary pupil health outcome [117]. ICCs for social emotional functioning outcomes would be of value to researchers designing future school-based CRTs in this area. ICC estimates from school-based CRTs are, in theory, most relevant than those from surveys as they may be more reflective of outcome variation across the types of schools that are more likely to participate in health-related CRTs [3](p177).

When designing a CRT in school setting, there are several levels at which randomisation can be undertaken, including schools, year groups, teachers and classrooms. Randomising smaller, lower-level cluster types like classes has a greater risk of contamination between trial arms than if entire schools are

randomised as, in the former scenario, pupils can interact between trial arms within the same school. The risk of contamination will depend on the nature of the intervention. Studies that randomise lower level clusters are, however, potentially more efficient as they will typically include a larger number of allocated cluster units, given there are, for example, more classroom units than school units [297]. Researchers should consider the benefits and detriments of randomising at each level. This requires knowledge of the components of variance and the ICC at the levels of clustering at which randomisation might be undertaken in the school setting, and the risk of contamination when allocating at those levels.

Sample size calculations in CRTs usually only explicitly recognise variation in the outcome at the level of randomisation (the cluster) and the level of observation (the individual participant), a simple two-level data structure. In CRTs where school clusters are randomised, participation may be restricted to pupils that are members of lower-level clusters (e.g., year groups, classes) that are sub-sampled to participate in the study [153, 161, 163, 189, 298]. When planning such studies, outcome variation at the randomisation level (school), the intermediate (subsampling) level and the pupil level should be taken into account [8]. The design effect (DE) for a CRT that has a three-level structure, is determined by the relative sizes of the three components of variance and the number of intermediate-level clusters sampled from each school [8, 218]. In studies that randomise schools, the more intermediate-level cluster units sampled from each school the smaller the observed ICC will be at the school level if estimated from a simple two-level model analysis that recognises only schools and pupils as sampling units. This is because the school level outcome means are estimated with greater precision when more intermediate level clusters are included. Therefore, when using the simple DE formula that recognises only one level of clustering, researchers should specify a school-level ICC for a planned study that reflects the number of intermediate level clusters that will be included from each school. Preferably, DE formulae that are appropriate for trials with intermediate levels of clustering should be used, but this requires knowledge of the components of variation (or the ICC) at all levels of the data.

A characteristic of school-based trials of interventions for improving social emotional functioning of pupils is the reporting of outcomes by different sources, specifically by the pupils themselves, parents/carers and teachers. The

components of variance and ICC may depend on the type of person that is reporting the outcome on the pupil and this needs to be considered when specifying an assumed ICC in the sample size calculation for a CRT.

In the previous chapter, factors that potentially influence the size of the ICC for health outcomes were investigated by extracting data from published studies and, using characteristics at the study level to investigate patterns in the ICC. The relationships examined between those characteristics and the ICC were potentially confounded by other design and contextual differences across the studies. The secondary analysis of raw data from CRTs to be undertaken in this chapter, has several advantages over the use of data extracted from published papers. The use of raw datasets from CRTs provides more control over the level of detail reported on the ICCs as the scope of the analysis is not restricted to the information reported in the publications. For example, the components of variance at the school, year group, class and pupil levels are readily calculated. The analysis of raw data also facilitates the use of within-study information to identify the determinants of the ICC, thus avoiding the limitation of study-level confounding in the previous chapter. By using raw data, it is possible, for example, to investigate whether the size of the ICC is stable across the study waves from baseline through to the final follow-up. Finally, there is the opportunity to, comprehensively, report the ICC for all relevant outcomes.

6.3 Aims and objectives

The aim of this study was to use raw data from five UK school-based CRTs to estimate ICCs and components of variance at different levels of clustering for pupil social emotional functioning outcomes.

The objectives were to:

- Collate estimates of components of variance and ICCs (at school, year group and class levels) for pupil social emotional functioning outcomes.
- Compare components of variance and ICCs across different levels of clustering that are relevant to school settings (i.e., school versus year group versus class).
- Compare components of variance and ICCs across different types of reporter for same outcome (i.e., pupil versus parent/carer versus teacher).

- Assess the stability of the ICCs over time (across study waves).
- Compare components of variance and ICCs for the same outcomes across studies.

6.4 Methods

6.4.1 Datasets

The data used in this secondary analysis are from five published UK school-based CRTs that evaluated interventions for improving social emotional functioning outcomes on pupils [145, 161, 202, 203, 298]. Permission to use these data was granted by the principal investigator for each study, while individual participant information and consent permitting such future secondary analyses was covered by the original consent agreements. All cluster-level and individual-level data were anonymised in the original studies. Ethical approval for use of the datasets was granted by the University of Exeter Medical School Research Ethics Committee (Appendix 11).

6.4.2 Description of datasets

Table 6.1 summarises the characteristics of the five UK school-based CRTs, and Table 6.2 provides information regarding the outcomes, outcome measures, reporters and outcome score calculation in each study. Table 6.3 summarises the baseline demographic characteristics of participants in each study. The datasets are described below. Studies are referred to by their study/intervention acronym throughout the chapter.

6.4.2.1 STARS study

Supporting Teachers and childRen in Schools (STARS) [161] was a CRT undertaken in primary schools in the South West of England. The aim of the study was to evaluate whether the Incredible Years® Teacher Classroom Management (TCM) programme [299] improved children's mental health, behaviour and enjoyment of school. Participants were pupils aged 4-9 years (Reception to Year 4). The study used a two-arm, parallel CRT design that recruited three cohorts of schools (clusters) between 2012 and 2014. Schools were randomised to either the TCM programme (intervention) or teaching-as-usual (control) (Table 6.1).

One class was sub-sampled from each recruited school for participation. Eighty (80) schools were randomised and 2075 pupils were recruited to the study: 40 schools (1037 pupils) in the intervention arm and 40 (1038 pupils) in the control arm. The TCM programme was delivered to teachers in the intervention arm in six whole-day sessions, spread over 6 months. Outcome data were collected at baseline (0), 9, 18, and 30 months. Teacher-reported outcomes were provided by the same teacher for all pupils in a given class at a given data collection point. *Social and emotional functioning* was measured using the **Strengths and Difficulty questionnaire (SDQ)** [300], providing a *total difficulties score* and subscales scores for *emotional symptoms, conduct problems, hyperactivity, peer problem* and *prosocial behaviour*. Parent- and teacher-reported versions of the SDQ were administered. *Pupil behaviour* was measured using the **Pupil Behaviour Questionnaire (PBQ)** [301], completed by the class teacher. *School climate* was measured using the pupil-reported '**How I Feel About My School**' (**HIFAMS**) [302] questionnaire (Table 6.2).

6.4.2.2 KiVa study

KiVa [145] was a CRT undertaken in primary schools in Wales. The study evaluated the effectiveness of the 'Kiusaamista Vastaa' (KiVa) programme [303] to prevent and address bullying in schools. Participants were pupils aged 7-11 years (school Years 3 to 6). The study used a two-arm, parallel CRT design with a waitlist (delayed intervention) control arm. Schools (clusters) were randomised to KiVa (intervention) or usual school provision (control) (Table 6.1). Schools were recruited in the middle of the 2012/13 academic year, with outcomes measured at the end of the 2013/14 academic year. Twenty-two (22) schools were randomised with 146 classes and 3214 pupils included in the study: 11 schools (77 classes, 1588 pupils) in the intervention arm and 11 schools (69 classes, 1892 pupils) in the control arm. Outcome data were collected at baseline (0) and 12 months. The outcomes were: *bullying victimisation* and *bullying perpetration* measured by the **Olweus Bully/Victim Questionnaire (OBVQ)** [304] and the **KiVa student online survey** [305], reported by the pupil; and *social and emotional functioning* measured using the **Strengths and Difficulties Questionnaire (SDQ)** [300], completed by the class teacher (Table 6.2).

6.4.2.3 PACES study

PACES [203] was a CRT undertaken in primary schools in the South West of England. The study evaluated the effectiveness of a classroom-based cognitive behaviour therapy (CBT) prevention programme (FRIENDS for life (FRIENDS) [306]) for reducing anxiety symptoms in children. Participants were pupils aged 9-10 years (school Year 5). The study used a three-arm parallel CRT design and took place between September 2011 and July 2012. Schools (clusters) were randomised to either receive school-led FRIENDS (led by teachers or school staff), health-led FRIENDS (led by trained health facilitators), or usual school provision (Table 6.1). Forty-five (45) schools were randomised and, 73 classes and 1448 pupils were included in the study: 14 schools (25 classes, 489 pupils) in the school-led FRIENDS arm; 14 schools (26 classes, pupils 472) in the health-led FRIENDS arm; and 12 schools (22 classes, pupils 401) in the control arm. Outcomes were measured at baseline (0), 6 and 12 months. *Symptoms of anxiety and low mood* were measured by the **Revised Child Anxiety and Depression Scale (RCADS-30)** [307], with a *total anxiety* score and subscale scores for *separation anxiety disorder, social phobia, generalised anxiety disorder, panic disorder, obsessive compulsive disorder, and low mood (major depressive disorder)*. The RCADS measures was reported separately by the pupil and the parent (RCADS-30-P). *Worry* was measured using the **Penn State Worry Questionnaire for Children** [308], reported by the pupil. *Self-worth and acceptance* was measured using the **Rosenberg Self-Esteem Scale** [309], reported by the pupil. *Bullying victimisation* was measured using the **Olweus Bully/Victim Questionnaire (OBVQ)** [304], reported by the pupil. *Life satisfaction* was measured using the **Child Health Utility instrument (CHU9D)** [310], reported by the pupil. *Social and emotional functioning* was measured by the **Strengths and Difficulties Questionnaire (SDQ)** [300], reported separately by the parent and the class teacher (Table 6.2).

6.4.2.4 PROMISE study

PROMISE [202] was a three-arm CRT undertaken in secondary schools in the East Midlands and South West of England. The study evaluated the effectiveness of classroom-based CBT (*The Resourceful Adolescent Programme* [311]) for improving social emotional outcomes on pupils, using for comparison an attention

control arm (Personal, Social, and Health Education (PSHE) delivered by class teacher aided by two facilitators) and a usual school provision control arm. Participants were aged 12-16 years (school Years 8-11). The study used a three-arm parallel CRT design, allocating year groups (clusters) to either CBT intervention, attention control, or usual school provision (Table 6.1). Twenty-eight (28) year groups from 8 schools with 225 classes and 5030 pupils were randomised: 9 year groups (79 classes, 1753 pupils) to CBT, 9 year-groups (73 classes, 1673 pupils) to attention control, and 9 year groups (73 classes, 1604 pupils) to usual school provision. Outcomes were measured at baseline (0), 6 and 12 months as follows: *symptoms of low mood* using the **Short Mood and Feelings questionnaire (SMFQ)** [312], reported by the pupil; *negative thinking* using the *Personal Failure subscale* of the **Children's Automatic Thoughts Scale (CATS)** [313], reported by the pupil; *self-worth and acceptance* using the **Rosenberg Self-Esteem Scale** [309], reported by the pupil; *anxiety* measured by the **Revised Child Anxiety and Depression Scale (RCADS-30)** [307], reported by the pupil; *school connectedness* measured by **Psychological Sense of School Membership (PSSM) scale** [314], reported by the pupil (Table 6.2).

6.4.2.5 MYRIAD study

MYRIAD [298] was a parallel arm CRT undertaken in secondary schools across the UK. The study evaluated the effectiveness of school-based mindfulness training (intervention) for improving student's mental health, compared to teaching-as-usual (control). Participants were pupils aged 11–14 years (school Years 7-9). Schools (clusters) were randomised to the mindfulness training (intervention) arm or the control arm (Table 6.1). School classes within schools were selected to participate, subsampling a sufficient number of classes to recruit the required number of pupils in each school. Eighty-five (85) schools were randomised with 346 classes and 8376 pupils included in the study: 42 schools (169 classes and 4144 pupils) in the intervention arm, and 43 schools (177 classes and 4232 pupils) in the control arm. Baseline data were collected on the three pupil-reported co-primary outcomes (*risk for depression* using the **Centre for Epidemiologic Studies for Depression Scale (CES-D)** [315], *social and emotional behavioural functioning* using the **Strengths and Difficulties Questionnaire (SDQ)** [300], and *well-being* using the **Warwick-Edinburgh Mental Well-being Scale (WEMWBS)** [316]). These and other secondary

outcomes were administered at 12, 19 and 24 months. The secondary outcomes were: *executive function* measured by the **Behaviour Rating Inventory of Executive Function (BRIEF-2)** [317], reported separately by both the pupil and the class teacher; *anxiety* using the **Revised Child Anxiety and Depression Scale (RCADS-30)** [307], reported by the pupil; *self-harm and suicidal ideation* using **measures devised for study** [298], reported by the pupil; *school climate subscales (school leadership and involvement, respectful climate, peer climate, caring adults)* from the **School Climate and Connectedness Survey (SCCS)** [318], reported by the pupil; *mindfulness skills* using the **Child and Adolescent Mindfulness Measure (CAMM)** [319], reported by the pupil (Table 6.2).

Table 6.1. Characteristics of the school-based cluster randomised trials at randomisation

Author, year (Study acronym)	Education setting; <i>location</i>	Cluster unit allocated	Measurement time points (months)	Number of schools	Number of year groups	Number of classes	Number of pupils
Ford, 2019 [161] (STARS)	Primary schools; <i>South West England</i>	Schools (1 class sampled from each school)	0, 9, 18, 30	80	not applicable	80	2075
Axford, 2020 [145] (KiVa)	Primary schools; <i>Wales</i>	Schools	0, 12	22	not applicable	146	3214
Stallard, 2014 [203] (PACES)	Primary schools; <i>South West England (within 50-miles of the University of Bath)</i>	Schools	0, 6, 12	40	not applicable	73	1448
Stallard, 2012 [202] (PROMISE)	Secondary schools; <i>East Midlands and South West England</i>	Year groups	0, 6, 12	8	28	225	1064
Kuyken, 2022 [298] (MYRIAD)	Secondary schools; <i>England, Northern Ireland, Scotland, Wales</i>	Schools	0, 12, 19, 24	85	not applicable	346	8378

Table 6.2. Description of outcomes, outcome measures and outcome scoring

Author, year (Study acronym)	Outcome	Outcome measure	Type of outcome	Number of items, scoring and scoring range	Outcome reporter(s)
Ford, 2019 [161] (STARS)	Social and emotional functioning	Strengths and Difficulty questionnaire [300] (Total difficulties score, <i>emotional symptoms, conduct problems, hyperactivity, peer problem</i> and <i>prosocial behaviour</i> subscales)	Continuous	<ul style="list-style-type: none"> • 25 ordinal items • Each item scored from 0 to 2 • 20 items are summed to calculate the <i>total difficulties</i> score (excluding the 5 items in the <i>prosocial behaviour</i> subscale) • 5 items are summed to calculate total score for each of 5 subscales • Scoring range for the <i>total difficulties</i> score is 0 to 40 • Scoring range for each of the 5 subscales is 0 to 10 	Class teacher Parent
	School climate	'How I Feel About My School measure' (HIFAMS) [302]	Continuous	<ul style="list-style-type: none"> • 7 ordinal items • Each item scored from 0 to 2 • Total score ranges from 0 to 14 	Pupil
	Pupil behaviour	Pupil Behaviour Questionnaire [301]	Continuous	<ul style="list-style-type: none"> • 6 ordinal items • Each item scored from 0 to 2 • Total score ranges from 0 to 12 	Class teacher
Axford, 2020 [145] (KiVa)	Bullying victimisation and bullying perpetration	Olweus Bully/Victim Questionnaire [304] (Bullying victimisation and bullying perpetration)	Binary ¹	<ul style="list-style-type: none"> • Bullying victimisation was measured using the item: "How often have you been bullied at school in the last couple of months?" • Bullying perpetration was measured using the item: "How often have you bullied others at school in the last few months?" 	Pupil

Author, year (Study acronym)	Outcome	Outcome measure	Type of outcome	Number of items, scoring and scoring range	Outcome reporter(s)
				<ul style="list-style-type: none"> Each item scored from 0 to 4 (0 = "Not at all", 1 = "Only once or twice", 2 = "2-3 times a month", 3 = "About once a week", 4 = "Several times per week") Each item was dichotomised for analysis so that those scoring 2 to 4 were classified as victims/perpetrators and those scoring 0 or 1 as not victims/not perpetrators. 	
	Bullying victimisation and bullying perpetration	KiVa student online survey [305]	Binary	<ul style="list-style-type: none"> Told school about being bullied (Yes/No) Did not tell school about being bullied (Yes/No) Told home about being bullied (Yes/No) 	Pupil
	Social and emotional functioning	Strengths and Difficulty questionnaire [300] (Total difficulties score, <i>emotional symptoms, conduct problems, hyperactivity, peer problem</i> and <i>prosocial behaviour</i> subscales)	Continuous	<ul style="list-style-type: none"> 25 ordinal items Each item scored from 0 to 2 20 items are summed to calculate the <i>total difficulties</i> score (excluding the 5 items in the <i>prosocial behaviour subscale</i>) 5 items are summed to calculate total score for each of 5 subscales Scoring range for the <i>total difficulties</i> score is 0 to 40 Scoring range for each of the 5 subscales is 0 to 10 	Teacher

Author, year (Study acronym)	Outcome	Outcome measure	Type of outcome	Number of items, scoring and scoring range	Outcome reporter(s)
Stallard, 2014 [203] (PACES)	Symptoms of anxiety and low mood	Revised Child Anxiety and Depression Scale (RCADS-30) [307] (Total anxiety scale, <i>separation anxiety disorder (SAD)</i> , <i>social phobia</i> , <i>generalised anxiety disorder (GAD)</i> , <i>panic disorder</i> , <i>obsessive compulsive disorder (OCD)</i> , and <i>low mood (major depressive disorder)</i> subscales)	Continuous	<ul style="list-style-type: none"> • 47 ordinal items • Subscales: SAD - 7 items; Social Phobia - 9 items; GAD - 6 items; Panic Disorder - 9 items; OCD - 6 items; low mood - 10 items. • Each item scored from 0 to 3 • Total anxiety score is the sum of SAD, Social Phobia, GAD, Panic Disorder and OCD subscales. • Scores for total score range from 0 to 111 • Subscales scores range from: SAD - 0 to 21; Social Phobia - to 27; GAD - 0 to 18; Panic Disorder - 0 to 27 ; OCD - 0 to 18 items; low mood - 0 to 30 	Pupil Parent (RCADS-30-P)
	Worry	Penn State Worry Questionnaire for Children [308]	Continuous	<ul style="list-style-type: none"> • 14 ordinal items • Each item scored from 0 to 3 • Scores range from 0 to 42 	Pupil
	Self-worth and acceptance	Rosenberg Self-Esteem Scale [309]	Continuous	<ul style="list-style-type: none"> • 10 ordinal items • Each item scored from 0-3 • Scores range from 0 to 30 	Pupil
	Bullying victimisation	Olweus Bully/Victim Questionnaire [304]	Binary ¹	<ul style="list-style-type: none"> • Bullying victimisation was measured using the item: "How often have you been bullied at school in the last couple of months?" • Each item scored from 0 to 4 (0 = "Not at all", 1 = "Only once or twice", 2 = "2-3 times a month", 3 = "About once a week", 4 = "Several times per week") 	Pupil

Author, year (Study acronym)	Outcome	Outcome measure	Type of outcome	Number of items, scoring and scoring range	Outcome reporter(s)
				<ul style="list-style-type: none"> Dichotomised for analysis (scores greater than or equal to 2) - 'Bullied more than or equal to 2 -3 times per month' 	
	Life satisfaction	Child Health Utility instrument (CHU9D) [310]	Continuous	<ul style="list-style-type: none"> 9 ordinal items Each item scored from 1 to 5 Total scores range from 9 to 45 	Pupil
	Social and emotional functioning	Strengths and Difficulty questionnaire [300] (Total difficulties score, <i>emotional symptoms, conduct problems, hyperactivity, peer problem</i> and <i>prosocial behaviour</i> subscales)	Continuous	<ul style="list-style-type: none"> 25 ordinal items Each item scored from 0 to 2 20 items are summed to calculate the <i>total difficulties</i> score (excluding the 5 items in the <i>prosocial behaviour subscale</i>) 5 items are summed to calculate total score for each of 5 subscales Scoring range for the <i>total difficulties</i> score is 0 to 40 Scoring range for each of the 5 subscales is 0 to 10 	Teacher Parent
Stallard, 2012 [202] (PROMISE)	Symptoms of low mood	Short Mood and Feelings questionnaire [312]	Continuous	<ul style="list-style-type: none"> 13 ordinal items Each item scored from 0 to 2 Total scores range from 0 to 26 	Pupil
	Negative thinking	Personal Failure subscale of the Children's Automatic Thoughts Scale (CATS) [313]	Continuous	<ul style="list-style-type: none"> 10 ordinal items Each item scored from 0 to 4 Total scores range from 0 to 40 	Pupil
	Self-worth and acceptance	Rosenberg Self-Esteem Scale [309]	Continuous	<ul style="list-style-type: none"> 10 ordinal items Each item scored from 0 to 3 Total scores range from 0 to 30 	Pupil

Author, year (Study acronym)	Outcome	Outcome measure	Type of outcome	Number of items, scoring and scoring range	Outcome reporter(s)
	Anxiety	Revised Child Anxiety and Depression Scale (RCADS-30) [307] (Total anxiety scale, <i>separation anxiety disorder (SAD)</i> , <i>social phobia</i> , <i>generalised anxiety disorder (GAD)</i> , <i>panic disorder</i> , <i>obsessive compulsive disorder (OCD)</i> , and <i>low mood (major depressive disorder)</i> subscales)	Continuous	<ul style="list-style-type: none"> • 47 ordinal items • Subscales: SAD - 7 items; Social Phobia - 9 items; GAD - 6 items; Panic Disorder - 9 items; OCD - 6 items; low mood - 10 items. • Each item scored from 0 to 3 • Total anxiety score is the sum of SAD, Social Phobia, GAD, Panic Disorder and OCD subscales. • Scores for total score range from 0 to 111 • Subscales scores range from: SAD - 0 to 21; Social Phobia - to 27; GAD - 0 to 18; Panic Disorder - 0 to 27 ; OCD - 0 to 18 items; low mood - 0 to 30 	Pupil
	School connectedness	Psychological Sense of School Membership (PSSM) scale [314]	Continuous	<ul style="list-style-type: none"> • 18 ordinal items • Each item scored from 1 to 5 • Total scores range from 18 to 90 	Pupil
Kuyken, 2022 [298] (MYRIAD)	Risk for depression	Centre for Epidemiologic Studies for Depression Scale (CES-D) [315]	Continuous	<ul style="list-style-type: none"> • 20 ordinal items • Each item scored from 0 to 3 • Total scores range from 0 to 60 	Pupil
	Social and emotional functioning	Strengths and Difficulty questionnaire [300] (Total difficulties score, <i>emotional symptoms</i> , <i>conduct problems</i> , <i>hyperactivity</i> , <i>peer problem</i> and <i>prosocial behaviour</i> subscales)	Continuous	<ul style="list-style-type: none"> • 25 ordinal items • Each item scored from 0 to 2 • 20 items are summed to calculate the <i>total difficulties</i> score (excluding the 5 items in the <i>prosocial behaviour subscale</i>) • 5 items are summed to calculate total score for each of 5 subscales 	Pupil Teacher

Author, year (Study acronym)	Outcome	Outcome measure	Type of outcome	Number of items, scoring and scoring range	Outcome reporter(s)
				<ul style="list-style-type: none"> Scoring range for the <i>total difficulties</i> score is 0 to 40 Scoring range for each of the 5 subscales is 0 to 10 	
	Well-being	Warwick-Edinburgh Mental Well-being Scale (WEMWBS) [316]	Continuous	<ul style="list-style-type: none"> 14 ordinal items Each item scored from 1 to 5 Total scores range from 14 to 70 	Pupil
	Executive function	Behaviour Rating Inventory of Executive Function (BRIEF-2) [317]	Continuous	<ul style="list-style-type: none"> 55 ordinal items for pupil 63 ordinal items for teachers Each item scored from 1 to 3 Scores ranged from 55-165 for the pupil version and from 63-189 for the teacher version 	Pupil Teacher
	Anxiety	Revised Child Anxiety and Depression Scale (RCADS-30) [307] (Total anxiety scale, <i>separation anxiety disorder (SAD)</i> , <i>social phobia</i> , <i>generalised anxiety disorder (GAD)</i> , <i>panic disorder</i> , <i>obsessive compulsive disorder (OCD)</i> , and <i>low mood (major depressive disorder)</i> subscales)	Continuous	<ul style="list-style-type: none"> 47 ordinal items Subscales: SAD - 7 items; Social Phobia - 9 items; GAD - 6 items; Panic Disorder - 9 items; OCD - 6 items; low mood - 10 items. Each item scored from 0 to 3 Total anxiety score is the sum of SAD, Social Phobia, GAD, Panic Disorder and OCD subscales. Scores for total score range from 0 to 111 Subscales scores range from: SAD - 0 to 21; Social Phobia - to 27; GAD - 0 to 18; Panic Disorder - 0 to 27 ; OCD - 0 to 18 items; low mood - 0 to 30 	Pupil

Author, year (Study acronym)	Outcome	Outcome measure	Type of outcome	Number of items, scoring and scoring range	Outcome reporter(s)
	Self-harm and suicidal ideation	Measures devised for study [298]	Binary	<ul style="list-style-type: none"> Self-harm: 'Have you deliberately harmed yourself?' Response set: "Yes" or "No" Suicide ideation: 'Do you feel like your life is not worth living?' Response set: "Yes" or "No" 	Pupil
	School climate	<i>School climate subscale (School leadership and involvement, respectful climate, peer climate, caring adults)</i> School Climate and Connectedness Survey (SCCS) [318]	Continuous	<ul style="list-style-type: none"> 4 sub-sections make up school climate subscale Each sub-section has 5 ordinal items (20 items in total) Each item scored from 1 to 5 Scores for each sub-section range from 5 to 25 Total scores for the school climate subscale range from 20 to 100 	Pupil
	Mindfulness skills	Child and Adolescent Mindfulness Measure (CAMM) [319]	Continuous	<ul style="list-style-type: none"> 10 ordinal items Each item scored from 0 to 4 Total scores range from 0 to 40 	Pupil

¹ The measure is continuous but dichotomised for analysis in the study

Table 6.3. Demographic characteristics of participants (N indicates sample size)

STARS study

Characteristic	N	Intervention	N	Control
Female, n (%)	1037	483 (46.6)	1038	491 (47.3)
Age in years, mean (SD)	1037	6.2 (1.4)	1038	6.4 (1.3)
White, n (%)	721	689 (95.6)	701	663 (94.6)
Eligible for free school meals (Yes), n (%)	595	70 (11.76)	502	64 (12.75)

KiVa study

Characteristic	N	Intervention	N	Control
Female, n (%)	1578	717 (45.4)	1636	684 (41.8)
Age in years, mean (SD)	1578	8.8 (1.1)	1636	8.9 (1.2)
White, n (%)	1578	1176 (74.5)	1636	1018 (62.2)
Eligible for free school meals (Yes), n (%)	1578	237 (15.0)	1636	220 (13.4)

PACES study

Characteristic	N	Health-led FRIENDS	N	School-led FRIENDS	N	Control
Female, n (%)	489	234 (47.9)	472	235 (49.8)	401	231 (57.6)
White ¹ , n (%)	489	455 (94.2)	472	439 (95.2)	401	359 (92.1)

¹British white

PROMISE study

Characteristic	N	Classroom based CBT	N	Attention control	N	Control
Female, n (%)	1753	873 (50)	1673	824 (49)	1604	770 (48)
Age in years, mean (SD)	1753	14.1 (1.1)	1673	14.0 (1.0)	1604	13.9 (1.2)
White, n (%)	1753	1372 (87)	1673	1271 (84)	1604	1275 (86)

MYRIAD study

Characteristic	N	Intervention	N	Control
Female, n (%)	4232	2350 (56.5)	4144	2159 (53.1)
Age in years, mean (SD)	4232	12.2 (0.6)	4144	12.2 (0.6)
White, n (%)	4232	3237 (78.1)	4144	2965 (73.2)

6.4.3 Data analysis

Data analysis was undertaken using Stata 17 software [227]. Mixed effects (“multilevel”) linear regression models were fitted to each outcome to estimate the variance components and the ICCs.

A 2-level mixed effects model was fitted to estimate the ICCs for the STARS study that had a single level of clustering at the school level:

$$Y_{il} = \alpha + s_i + e_{il}$$

- Y_{il} is the outcome for the l^{th} individual in the i^{th} school (cluster)
- α is the constant
- s_i is the random effect of the i^{th} school, assumed to be Normally distributed with 0 mean and constant variance σ_s^2
- e_{il} is the residual effect of the l^{th} individual in the i^{th} school assumed to be Normally distributed with 0 mean and constant variance σ_e^2

The school-level ICC (ρ_s) is calculated from the between-cluster (σ_s^2) (school) and within-cluster (σ_e^2) components of variances using:

$$\rho_s = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2}$$

Three-level mixed effects models were fitted to estimate the ICCs for the KiVa, PACES and MYRIAD studies that had two levels of clustering (school and class):

$$Y_{ikl} = \alpha + s_i + c_{ik} + e_{ikl}$$

- Y_{ikl} is the outcome for the l^{th} individual in the k^{th} class, in the i^{th} school (cluster)
- α is the constant
- s_i is the random effect of the i^{th} school, assumed to be Normally distributed with 0 mean and constant variance σ_s^2
- c_{ik} is the random effect of the k^{th} class in the i^{th} school, assumed to be Normally distributed with 0 mean and constant variance σ_c^2
- e_{ikl} is the residual effect of the l^{th} individual in the k^{th} class, in the i^{th} school, assumed to be Normally distributed with 0 mean and constant variance σ_e^2

The school-level ICC (ρ_s) is calculated from the variance components as:

$$\rho_s = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_c^2 + \sigma_e^2}$$

and the class-level ICC (ρ_c) is calculated as:

$$\rho_c = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_e^2}$$

This definition of the class level ICC is appropriate to use when designing cluster randomised trials where allocation of classroom clusters is stratified by school membership.

Four-level mixed effects models were fitted to estimate the ICCs for the PROMISE study that had three levels of clustering (school, year group and class):

$$Y_{ijkl} = \alpha + s_i + g_{ij} + c_{ijk} + e_{ijkl}$$

- Y_{ijkl} is the outcome for the l^{th} individual in the k^{th} class, in the j^{th} year group, in the i^{th} school (cluster)
- α is the constant
- s_i is the random effect of the i^{th} school, assumed to be Normally distributed with 0 mean and constant variance σ_s^2
- g_{ij} is the random effect of the j^{th} year group in the i^{th} school, assumed to be Normally distributed with 0 mean and constant variance σ_g^2
- c_{ijk} is the random effect of the k^{th} class, in the j^{th} year group in the i^{th} school, assumed to be Normally distributed with 0 mean and constant variance σ_c^2
- e_{ijkl} is the residual effect of the l^{th} individual in the k^{th} class, in the j^{th} year group in the i^{th} school, assumed to be Normally distributed with 0 mean and constant variance σ_e^2

The school-level ICC (ρ_s) is calculated from the variance components as:

$$\rho_s = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_g^2 + \sigma_c^2 + \sigma_e^2}$$

the year group-level ICC (ρ_g) is calculated as:

$$\rho_g = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_c^2 + \sigma_e^2}$$

and the class-level ICC is calculated as:

$$\rho_c = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_e^2}$$

The definitions of the year group-level and class-level ICCs are appropriate to use when designing cluster randomised trials where allocation of clusters at those levels is stratified by higher level clusters.

ICC estimates at the baseline time point were obtained by fitting “null” or “empty” models that had no predictor variables. ICC estimates at follow-up were adjusted for trial arm status by adding the variable as a predictor (fixed effect) to the above models.

All components of variance and ICC estimates were rounded to three decimal places. Ninety five percent confidence intervals were calculated for school-level ICC estimates and are reported in Appendix 12.

6.5 Results

6.5.1 STARS study

ICC estimates for outcomes from the STARS study are reported for the Strengths and Difficulties Questionnaire (SDQ) (Table 6.4), the Pupil Behaviour Questionnaire (PBQ) (Table 6.5) and the “How I Feel About My School” (HIFAMS) measure (Table 6.5). There was no marked pattern of change in the ICC estimates for all three outcomes over time.

The ICC estimates for the SDQ (Table 6.4) were markedly larger when the outcome was reported by the class teacher than the parent. For example, the ICC estimates for the SDQ total difficulties score reported by the class teacher ranged from 0.120 to 0.180 across the 4 study waves, while the corresponding estimates reported by parents ranged from 0.026 to 0.046. Teacher ICCs were markedly smaller for *conduct problems* and *hyperactivity* subscales and larger for the *prosocial behaviour* subscales. Parent-reported ICCs were very small (near 0) for the SDQ *prosocial behaviour* subscale compared with the other subscales. There was no clear pattern of change in the size of the ICCs over time for teacher- and parent-reported SDQ outcomes. ICC estimates were generally imprecise for teacher and parent-reported outcomes, the wide confidence intervals (Appendix 12) indicating palpable uncertainty about the correct value to assume when planning a future study given that even small differences in the ICC can have a large impact on the design effect and, therefore, the required sample size.

The ICCs for the teacher-reported Pupil Behaviour Questionnaire (PBQ) at the baseline and 9-month study waves (Table 6.5) were similar to those for the teacher-reported SDQ *conduct problems* subscale (Table 6.4). Both measures quantify the teacher’s view of the pupil’s conduct.

The ICC estimates for the ‘How I Feel About My School’ measure (HIFAMS) increased over time, from 0.052 at baseline to 0.111 at 30 months (Table 6.5).

Table 6.4. STARS study intra-cluster correlation coefficients (ICCs) for the Strengths and Difficulties Questionnaire (SDQ) outcomes at different time points

Outcome	Measurement time (months) ¹	Teacher report				Parent report			
		N	σ_s^2	σ_e^2	ρ_s	N	σ_s^2	σ_e^2	ρ_s
Total difficulties score	0	2074	4.118	30.181	0.120	1466	0.915	34.799	0.026
	9	2001	6.114	27.896	0.180	1285	1.855	39.342	0.046
	18	1848	7.812	35.842	0.179	1225	1.238	38.425	0.031
	30	1756	4.894	35.502	0.121	1125	1.512	43.246	0.034
Emotion symptoms subscale	0	2074	0.421	3.754	0.101	1467	0.098	3.864	0.025
	9	2001	0.854	3.370	0.202	1286	0.147	4.645	0.031
	18	1848	0.853	3.921	0.179	1227	0.071	4.828	0.014
	30	1756	0.393	3.952	0.090	1126	0.109	5.569	0.019
Conduct problems subscale	0	2074	0.144	2.190	0.062	1467	0.035	2.610	0.013
	9	2001	0.237	2.324	0.092	1287	0.046	2.810	0.016
	18	1848	0.359	2.705	0.117	1228	0.074	2.436	0.030
	30	1756	0.291	2.505	0.104	1127	0.004	2.967	0.001
Hyperactivity subscale	0	2074	0.509	9.028	0.053	1466	0.024	6.742	0.004
	9	2001	0.787	7.937	0.090	1287	0.088	7.002	0.012
	18	1848	0.826	8.302	0.091	1227	0.061	6.576	0.009
	30	1756	0.601	7.787	0.072	1127	0.070	6.777	0.010
Peer problems subscale	0	2074	0.391	2.180	0.152	1466	0.056	2.584	0.021
	9	2001	0.288	2.13	0.119	1286	0.145	2.840	0.049
	18	1848	0.368	2.434	0.131	1227	0.081	2.912	0.027
	30	1756	0.271	2.498	0.098	1126	0.140	2.952	0.045
Prosocial behaviour subscale	0	2074	1.404	4.600	0.234	1467	0	2.982	0
	9	2001	1.320	3.946	0.251	1287	0	2.929	0

Outcome	Measurement time (months) ¹	Teacher report			Parent report				
		N	σ_s^2	σ_e^2	ρ_s	N	σ_s^2	σ_e^2	ρ_s
	18	1848	1.135	4.420	0.204	1228	0.021	2.786	0.007
	30	1756	0.839	4.289	0.164	1127	0	1.888	0

¹ Time points at 9, 18, 30 months adjusted for trial arm status

σ_s^2 : School-level component of variance

σ_e^2 : Pupil-level component of variance

ρ_s : School-level ICC

Table 6.5. STARS study intra-cluster correlation coefficients (ICCs) for the teacher-reported Pupil Behaviour Questionnaire and the pupil-reported 'How I Feel About My School' measure and at different time points

Outcome measure	Reporter	Measurement time (months) ¹	N	σ_s^2	σ_e^2	ρ_s
Pupil Behaviour Questionnaire	Teacher	0	2053	0.373	5.472	0.064
		9	1986	0.507	5.401	0.086
		18	1886	0.545	6.095	0.082
		30	1760	0.499	5.688	0.081
'How I Feel About My School' measure	Pupil	0	2074	0.302	5.450	0.052
		9	2001	0.466	5.549	0.077
		18	1848	0.728	6.153	0.106
		30	1756	0.850	6.829	0.111

¹ Time points at 9, 18, 30 months adjusted for trial arm status

σ_s^2 : School-level component of variance

σ_e^2 : Pupil-level component of variance

ρ_s : School-level ICC

6.5.2 KiVa study

ICC estimates for the KiVa study are reported for the teacher-reported Strengths and Difficulties Questionnaire (SDQ) measure in Table 6.6 and the pupil-reported bullying/victimisation outcomes in Table 6.7.

In general, for the teacher-reported SDQ outcomes, variance components were smaller at the school level than at the class level of clustering (Table 6.6). Only for the SDQ *conduct* subscale were the school- and class-level components of variance of similar size. ICC estimates were also generally smaller at the school-level than the class-level for teacher-reported SDQ outcomes. The school-level ICC at baseline was largest for the *conduct* subscale (0.042), and at 12-month follow-up it was largest for the *emotional* subscale (0.092). At the class level, ICCs were largest for *prosocial behaviour* (0.206 and 0.148) and *emotional* (0.156 and 0.103) subscales.

Similar to the SDQ, the school-level variance components were smaller than at the class level for pupil-reported *bullying perpetration* and *bullying victimisation* (Table 6.7). School-level ICC estimates (ranging from 0.01 to 0.019) were also smaller than class-level estimates (0.019 to 0.036).

Table 6.6. KiVa study intra-cluster correlation coefficients (ICCs) for teacher-reported Strengths and Difficulties Questionnaire (SDQ) outcomes at different time points

Outcome	Measurement time (months) ¹	N	σ_s^2	σ_c^2	σ_e^2	ρ_s	ρ_c
Total difficulties score	0	2832	1.487	5.688	33.117	0.037	0.147
	12	2652	3.220	4.790	35.245	0.075	0.120
Emotional symptoms subscale	0	2832	0.144	0.655	3.549	0.033	0.156
	12	2652	0.403	0.411	3.574	0.092	0.103
Conduct problems subscale	0	2832	0.123	0.171	2.614	0.042	0.061
	12	2652	0.178	0.152	2.876	0.055	0.050
Hyperactivity subscale	0	2832	0.045	0.728	7.736	0.005	0.086
	12	2652	0.252	0.715	7.431	0.030	0.088
Peer problems subscale	0	2832	0.073	0.287	2.506	0.025	0.103
	12	2652	0.115	0.217	2.489	0.041	0.080
Prosocial behaviour subscale	0	2832	0.057	1.123	4.330	0.010	0.206
	12	2652	0.085	0.718	4.148	0.017	0.148

¹ Time point at 12 months adjusted for trial arm status

σ_s^2 : School-level component of variance

σ_c^2 : Class-level component of variance

σ_e^2 : Pupil-level component of variance

ρ_s : School-level ICC

ρ_c : Class-level ICC

Table 6.7. KiVa study intra-cluster correlation coefficients (ICCs) for pupil-reported Olweus Bully/Victim Questionnaire (OBVQ) and bullying outcomes (KiVa questionnaire) at different time points

Outcome	Measurement time (months) ¹	N	σ_s^2	σ_c^2	σ_e^2	ρ_s	ρ_c
Bullying victimisation (OBVQ)	0	2876	0.002	0.006	0.183	0.012	0.034
	12	2581	0.003	0.005	0.134	0.019	0.036
Bullying perpetration (OBVQ)	0	2876	0.001	0.002	0.076	0.010	0.031
	12	2581	0.001	0.002	0.076	0.010	0.031
Told school about being bullied (KiVa questionnaire)	0	2876	0.001	0.002	0.108	0.013	0.019
	12	2581	0.001	0.002	0.073	0.009	0.032
Did not tell school about being bullied (KiVa questionnaire)	0	2876	0.002	0.006	0.161	0.010	0.036
	12	2581	0.002	0.003	0.115	0.018	0.029
Told home about being bullied (KiVa questionnaire)	0	2876	0.001	0.004	0.133	0.006	0.032
	12	2581	0.002	0.002	0.096	0.017	0.024

¹ Time point at 12 months adjusted for trial arm status

σ_s^2 : School-level component of variance

σ_c^2 : Class-level component of variance

σ_e^2 : Pupil-level component of variance

ρ_s : School-level ICC

ρ_c : Class-level ICC

6.5.3 PACES study

ICC estimates are reported for the parent- and pupil-reported Revised Child Anxiety and Depression Scale (RCADS-30) outcomes in Table 6.8. ICC estimates for the parent-reported Strengths and Difficulties Questionnaire (SDQ) are reported in Table 6.9. The ICCs for pupil-reported bullying victimisation, worry, self-esteem and life satisfaction are reported in Table 6.10.

Revised Child Anxiety and Depression Scale (RCADS-30)

Parent report

For the parent-reported RCADS-30 outcomes, the school-level and class-level variance components were often very small (sometimes zero) for many subscales, especially at the class level (Table 6.8). Generally, ICC estimates at both school and class level were small (all under 0.05), but sometimes larger at the school level than the class level. ICC estimates were zero at the class level for *Generalised Anxiety Disorder (GAD)*, *panic disorder* and *obsessive-compulsive disorder* across all time points.

Pupil report

In comparison, pupil-reported RCADS-30 outcomes had smaller school-level variance components than at the class level (Table 6.8). Similarly, to parent-reported RCADS-30, the ICC estimates at both school and class level were small (all under 0.05). However, for pupil-reported outcome, ICC estimates at the class level were always larger than at the school level. The largest school-level ICC estimate was for *separation anxiety disorder subscale* at 12-month follow-up (0.026).

Strengths and Difficulties Questionnaire (SDQ)

For the parent-reported SDQ outcomes, there was no particular pattern in the size of school-level and class-level variance components (Table 6.9). For many of these outcomes the variance components at these levels was 0. In tandem with this, ICC estimates were often 0 at both school and class level. The largest ICC was 0.059 for the *peer problems* subscale at 6-month follow-up at the school-level. Across the three time points, the largest school-level ICCs were for the *peer problems* subscale (0.017 to 0.059), and the smallest were for the *prosocial*

behaviour subscale (0 at all three time points). The largest class-level ICC estimate (0.045) was for the *conduct problems* subscale at baseline. The class-level ICCs for the *prosocial behaviour* subscale were zero across the three time points.

Other pupil-reported outcomes

Across all pupil-reported outcomes, ICC estimates were generally larger at the class-level compared to the school-level (Table 6.10). For pupil-reported bullying victimisation, components of variance at both the school and class-level were small (between 0.001 and 0.008 at both levels). ICC estimates for pupil-reported bullying victimisation ranged from 0.005 to 0.051. For the pupil-reported worry outcome, ICC estimates at school-level for the baseline and 6-months timepoints were both 0. ICCs at both the class and school level were no larger than 0.02. For pupil-reported self-esteem, ICC estimates had no clear pattern at both the school and class-levels. For pupil-reported total life satisfaction, the school and class-level ICC estimates were similar at 12-month timepoints (0.027 and 0.028, respectively).

Table 6.8. PACES study intra-cluster correlation coefficients (ICCs) for the Revised Child Anxiety and Depression Scale (RCADS-30) at different time points

Outcome	Measurement time (months) ¹	Parent report						Pupil report					
		N	σ_s^2	σ_c^2	σ_e^2	ρ_s	ρ_c	N	σ_s^2	σ_c^2	σ_e^2	ρ_s	ρ_c
Total anxiety score	0	482	0	0	79.054	0	0	1281	0	9.662	227.314	0	0.041
	6	426	0	0	71.077	0	0	1274	0.544	9.336	228.518	0.002	0.039
	12	406	0.986	0.458	61.231	0.016	0.007	1203	2.417	7.032	223.379	0.010	0.031
Low mood subscale	0	560	0.042	0	2.401	0.017	0	1332	0.052	0.094	6.461	0.008	0.014
	6	477	0	0.018	2.472	0	0.007	1305	0	0.188	6.414	0	0.028
	12	445	0	0.024	2.280	0	0.010	1250	0.089	0.191	6.410	0.013	0.029
Separation Anxiety Disorder subscale	0	519	0	0.029	5.990	0	0.005	1330	0.106	0.292	10.577	0.010	0.027
	6	448	0.103	0	4.778	0.021	0	1308	0.217	0.354	8.946	0.023	0.038
	12	432	0.068	0.189	4.077	0.016	0.044	1247	0.235	0.263	8.582	0.026	0.030
Social phobia subscale	0	558	0.055	0	7.502	0.007	0	1328	0	0.346	10.271	0	0.033
	6	479	0	0.033	7.111	0	0.005	1307	0.151	0.248	10.988	0.014	0.022
	12	441	0.150	0	6.303	0.023	0	1244	0.071	0.298	11.018	0.006	0.026
Generalised Anxiety Disorder subscale	0	557	0	0	5.836	0	0	1328	0	0.496	13.593	0	0.035
	6	477	0.052	0	4.470	0.011	0	1305	0	0.521	12.956	0	0.039
	12	444	0.092	0	4.159	0.022	0	1242	0.022	0.489	12.710	0.002	0.037
Panic disorder subscale	0	550	0.005	0	1.377	0.004	0	1328	0	0.073	8.435	0	0.009
	6	473	0	0	1.337	0	0	1305	0	0.164	8.568	0	0.019
	12	443	0.007	0	0.898	0.007	0	1247	0.049	0.064	7.485	0.006	0.008
Obsessive-compulsive Disorder subscale	0	559	0.008	0	2.095	0	0	1325	0	0.416	9.945	0	0.040
	6	478	0	0	2.098	0.004	0	1307	0	0.279	10.315	0	0.026
	12	444	0.011	0	1.891	0.006	0	1245	0	0.223	9.805	0	0.022

¹ Time points at 6 and 12 months adjusted for trial arm status

σ_s^2 : School-level component of variance
 σ_c^2 : Class-level component of variance
 σ_e^2 : Pupil-level component of variance
 ρ_s : School-level ICC
 ρ_c : Class-level ICC

Table 6.9. PACES study intra-cluster correlation coefficients (ICCs) for the parent-reported Strengths and Difficulties Questionnaire (SDQ) outcomes at different time points

Outcome	Measurement time (months) ¹	N	σ_s^2	σ_c^2	σ_e^2	ρ_s	ρ_c
Total difficulties score	0	547	0.090	0.138	39.269	0.002	0.003
	6	460	0	1.248	34.869	0	0.035
	12	425	1.743	0	32.800	0.050	0
Emotional symptoms subscale	0	566	0	0	5.511	0	0
	6	475	0	0.001	4.213	0	<0.001
	12	439	0.119	0	3.663	0.032	0
Conduct problems subscale	0	563	0	0.139	2.949	0	0.045
	6	473	0	0.060	2.522	0	0.023
	12	441	0.015	0	2.330	0.006	0
Hyperactivity subscale	0	566	0	0	6.420	0	0
	6	475	0	0.153	5.260	0	0.028
	12	437	0.051	0	4.856	0.010	0
Peer problems subscale	0	561	0.093	0.028	3.189	0.028	0.037
	6	475	0.212	0	3.459	0.059	0
	12	438	0.058	0.039	3.280	0.017	0.012
Prosocial behaviour subscale	0	561	0	0	3.339	0	0
	6	471	0	0	3.296	0	0
	12	440	0	0	2.794	0	0

¹ Time points at 6 and 12 months adjusted for trial arm status

σ_s^2 : School-level component of variance

σ_c^2 : Class-level component of variance

σ_e^2 : Pupil-level component of variance

ρ_s : School-level ICC

ρ_c : Class-level ICC

Table 6.10. PACES study intra-cluster correlation coefficients (ICCs) for pupil-reported outcomes at different time points

Outcome	Measurement time (months) ¹	N	σ_s^2	σ_c^2	σ_e^2	ρ_s	ρ_c
Bullying victimisation (Olweus Bully/Victim Questionnaire)	0	1338	0.003	0.004	0.196	0.015	0.018
	6	1316	0.006	0.002	0.187	0.031	0.011
	12	1254	0.001	0.008	0.154	0.005	0.051
Worry (Penn Worry Scale)	0	1310	0	0.360	67.391	0	0.005
	6	1298	0	1.000	67.009	0	0.015
	12	1230	0.694	1.314	65.922	0.010	0.020
Self-esteem (Rosenberg Self-Esteem Scale)	0	1295	0	1.334	29.467	0	0.043
	6	1285	0.834	0.489	34.333	0.023	0.014
	12	1224	0.431	1.645	34.080	0.012	0.046
Total life satisfaction (CHU9D)	0	1333	0.328	0.821	37.111	0.009	0.022
	6	1302	0.135	1.443	42.155	0.003	0.033
	12	1241	1.114	1.108	38.341	0.027	0.028

¹ Time points at 6 and 12 months adjusted for trial arm status

σ_s^2 : School-level component of variance

σ_c^2 : Class-level component of variance

σ_e^2 : Pupil-level component of variance

ρ_s : School-level ICC

ρ_c : Class-level ICC

6.5.4. PROMISE study

The ICC estimates for the pupil-reported Revised Child Anxiety and Depression Scale (RCADS-30) are reported in Table 6.11. ICCs for all other pupil-reported outcomes from the PROMISE study are presented in Table 6.12.

Revised Child Anxiety and Depression Scale (RCADS-30)

For the RCADS-30 outcomes, the school-level variance components were generally smaller than at the year group level, which, in turn, were generally smaller than at the class level (Table 6.11). School-level ICCs were generally below 0.01, the exception being for *separation anxiety disorder* at baseline (0.014). Class-level ICCs ranged between 0.012 and 0.034 across all subscales. The year group-level ICCs were generally smaller than the class-level ICCs, but there were exceptions. The year group-level ICC estimates for the *social phobia subscale* were over twice the size of the ICC estimates at the class level (for example, 0.063 at year group level compared with 0.017 at class level for the baseline assessment).

Other pupil-reported outcomes

ICC estimates for self-esteem were notably smaller at the school level compared to the year group and class levels (Table 6.12). For pupil-reported personal failure using Children's Automatic Thoughts Scale (CATS), ICCs were largest at the class level across all timepoints. There was no clear pattern in the size of ICC estimates for school connectedness. ICCs were generally smallest at the school level compared to the year-group and class levels for the short mood and feelings questionnaire (SMFQ). For outcomes reported using the SMFQ, ICC estimates for a given level were similar to those seen for RCADS-30 *depression subscale* (SMFQ ICCs ranged from 0.001 to 0.10 and for RCADS-30 *depression subscale* ICCs ranged from 0 to 0.10) (Table 6.12).

Table 6.11. PROMISE study intra-cluster correlation coefficients (ICCs) for the pupil-reported Revised Child Anxiety and Depression Scale (RCADS-30) at different time points

Outcome	Measurement time (months) ¹	N	σ_s^2	σ_g^2	σ_c^2	σ_e^2	ρ_s	ρ_g	ρ_c
Total anxiety score	0	4588	0.760	2.350	3.467	95.071	0.007	0.023	0.035
	6	4395	0	3.093	2.905	105.624	0	0.028	0.027
	12	3948	0.720	3.219	3.303	108.924	0.006	0.028	0.029
Low mood subscale	0	4607	0.073	0.116	0.195	6.680	0.010	0.017	0.028
	9	4416	0	0.151	0.155	7.284	0	0.020	0.021
	12	3954	0.039	0.154	0.209	7.493	0.005	0.020	0.027
Panic disorder subscale	0	4612	0.055	0.066	0.198	5.568	0.009	0.011	0.034
	6	4422	0	0.082	0.093	6.727	0	0.012	0.014
	12	3957	0.029	0.074	0.126	6.172	0.005	0.012	0.020
Social phobia subscale	0	4612	0	0.569	0.143	8.326	0	0.063	0.017
	6	4420	0.015	0.657	0.219	8.675	0.002	0.069	0.025
	12	3956	0.061	0.521	0.116	9.348	0.006	0.052	0.012
Generalised Anxiety Disorder subscale	0	4616	0	0.115	0.160	6.992	0	0.016	0.022
	6	4427	0.029	0.054	0.219	7.635	0.004	0.007	0.028
	12	3958	0.061	0.107	0.158	7.693	0.008	0.013	0.020
Separation Anxiety Disorder subscale	0	4616	0.041	0.005	0.057	2.806	0.014	0.002	0.020
	6	4426	0.009	0.013	0.071	3.384	0.002	0.004	0.020
	12	3958	0.023	0.032	0.075	3.332	0.007	0.009	0.022

¹ Time points at 6 and 12 months adjusted for trial arm status

σ_s^2 : School-level component of variance

σ_g^2 : Year group-level component of variance

σ_c^2 : Class-level component of variance

σ_e^2 : Pupil-level component of variance

ρ_s : School-level ICC

ρ_g : Year group-level ICC

ρ_c : Class-level ICC

Table 6.12. PROMISE study intra-cluster correlation coefficients (ICCs) for pupil-reported outcomes at different time points

Outcome	Measurement time (months) ¹	N	σ_s^2	σ_g^2	σ_c^2	σ_e^2	ρ_s	ρ_g	ρ_c
Self-esteem (Rosenberg self-esteem scale)	0	4576	0.121	0.645	0.760	26.899	0.004	0.023	0.027
	6	4392	0	0.533	0.452	30.528	0	0.017	0.015
	12	3944	0	0.488	0.353	30.946	0	0.015	0.011
Personal failure (Children's Automatic Thoughts Scale (CATS))	0	4596	0.420	0.419	1.127	46.767	0.009	0.009	0.024
	6	4401	0.015	0.448	0.647	53.109	<0.001	0.008	0.012
	12	3945	0.035	0.573	1.227	48.746	0.001	0.011	0.025
School connectedness (Psychological Sense of School Membership scale) (PSSM scales)	0	4567	0.293	0.578	0.654	37.968	0.007	0.015	0.017
	6	4367	0.699	0.489	0.682	41.364	0.016	0.011	0.016
	12	3913	0.709	0.531	0.807	42.220	0.016	0.012	0.019
Short moods and feelings questionnaire (SMFQ)	0	4784	0.238	0.320	0.740	22.149	0.010	0.014	0.032
	6	4480	0.021	0.566	0.523	25.374	0.001	0.021	0.020
	12	4140	0.119	0.379	0.683	24.618	0.005	0.015	0.027

¹ Time points at 6 and 12 months adjusted for trial arm status

σ_s^2 : School-level component of variance

σ_g^2 : Year group-level component of variance

σ_c^2 : Class-level component of variance

σ_e^2 : Pupil-level component of variance

ρ_s : School-level ICC

ρ_g : Year group-level ICC

ρ_c : Class-level ICC

6.5.4 MYRIAD study

The ICC estimates for pupil- and teacher-reported Strengths and Difficulties Questionnaire (SDQ) are reported in Table 6.13. ICC estimates for the pupil- and teacher-reported Behaviour Rating Inventory of Executive Function, Second Edition (BRIEF-2) are reported in Table 6.14. ICCs for pupil-reported Revised Child Anxiety and Depression Scale (RCADS-30) outcomes are reported in Table 6.15. Table 6.16 reports ICC estimates of pupil-reported Centre for Epidemiologic Studies for Depression Scale, Warwick-Edinburgh Mental Well-being Scale, Child and Adolescent Mindfulness Measure, suicide ideation and self-harm outcomes. Lastly, ICC estimates for pupil-reported school climate and connectedness survey (SCCS) outcomes are reported in Table 6.17.

Ninety-five percent confidence intervals were narrow for all school-level ICC estimates demonstrating good precision (Appendix 12).

Strengths and Difficulties Questionnaire (SDQ)

For the pupil-reported SDQ outcomes, the components of variance and the ICC estimates were generally larger at school than at class level (Table 6.13). The ICCs were generally similar across subscales (ranging from 0.011 to 0.022 at the school level and from <0.001 to 0.021 at the class level). School-level ICCs were mostly larger than class-level estimates.

The ICC estimates for the teacher-reported SDQ outcomes were considerably larger than for pupil-reported SDQ outcomes (Table 6.13). For example, class-level ICCs for teacher-reported SDQ outcomes ranged from 0.077 to 0.197 compared to <0.001 to 0.021 for pupil-reported SDQ outcomes across the same time points. This is expected as there is only one teacher reporting on all pupils in their class and teachers will differ in their general tendency to give higher or lower scores (Table 6.13). Class-level variation partly reflects variation across teachers. School-level variance components were smaller than at the class level for all teacher-reported SDQ subscales.

Behaviour Rating Inventory of Executive Function, Second Edition (BRIEF-2)

For pupil-reported executive functioning, quantified by the BRIEF-2, ICC estimates were markedly larger than other pupil-reported measures (ICCs ranged from 0.058 to 0.090 at the school level and ranged from 0.042 to 0.103 at the

class level) (Table 6.14). Large ICC estimates were also noted for the teacher-reported BRIEF-2 (larger than the pupil-reported BRIEF-2), but the difference between BRIEF-2 and other teacher-reported outcomes was less marked.

Pupil-reported Revised Child Anxiety and Depression Scale (RCADS-30) outcomes

For pupil-reported RCADS-30 outcomes, ICC estimates were larger at the school-level than the class-level (Table 6.15). School-level ICC estimates ranged from 0.016 to 0.04, while class-level ICCs ranged from 0 to 0.02. The largest school- and class-level ICCs were for the *Social Anxiety* subscale.

Pupil-reported Centre for Epidemiologic Studies for Depression Scale (CES-D), Warwick-Edinburgh Mental Well-being Scale (WEMWBS), Child and Adolescent Mindfulness Measure (CAMM), suicide ideation and self-harm outcomes

School- and class-level ICC estimates for pupil-reported CES-D and WEMWBS were of similar size to those for the pupil-reported SDQ and RCADS-30 (Table 6.16). At the school level, ICCs for CES-D ranged from 0.016 to 0.023, for WEMWBS ranged from 0.015 to 0.019, for RCADS-30 ranged from 0.016 to 0.040, and for SDQ ranged from 0.011 to 0.022.

ICCs at both the school and class level were of similar magnitude, particularly for suicide ideation and self-harm. School-level ICC estimates ranged from 0.019 to 0.024 for CAMM, 0.011 to 0.013 for suicide ideation, and 0.005 to 0.011 for self-harm.

Pupil-reported school climate and connectedness survey (SCCS)

ICCs for the pupil-reported SCCS outcomes were generally larger than those for other pupil-reported outcomes, particularly at the school level (Table 6.17). For example, school level ICCs for SCCS ranged from 0.029 to 0.064 compared with 0.011 to 0.022 for the SDQ outcomes across the same time points.

Table 6.13. MYRIAD study intra-cluster correlation coefficients (ICCs) for pupil- and teacher-reported Strengths and Difficulties Questionnaire outcomes at different time points

Outcome	Measurement time (months) ¹	Pupil report						Teacher report					
		N	σ_s^2	σ_c^2	σ_e^2	ρ_s	ρ_c	N	σ_s^2	σ_c^2	σ_e^2	ρ_s	ρ_c
Total difficulties score	0	8252	1.051	0.666	40.010	0.025	0.016						
	12	8042	0.941	0.492	42.545	0.021	0.011	5873	1.820	5.561	28.343	0.051	0.164
	19	7542	0.933	0.933	45.043	0.020	0.021	5522	1.918	6.610	26.967	0.054	0.197
	24	7225	0.792	0.792	45.289	0.017	0.008	4477	2.523	5.195	26.077	0.075	0.166
Emotional symptoms subscale	0	8254	0.117	0.050	6.378	0.018	0.008						
	12	8042	0.156	0.003	6.937	0.022	<0.001	5873	0.234	0.486	3.247	0.059	0.130
	19	7542	0.164	0.094	7.213	0.022	0.013	5522	0.156	0.605	2.882	0.043	0.173
	24	7226	0.146	0.013	7.162	0.020	0.002	4477	0.181	0.523	2.818	0.051	0.157
Conduct difficulties subscale	0	8253	0.078	0.052	3.456	0.022	0.015						
	12	8042	0.061	0.029	3.543	0.017	0.008	5873	0.072	0.257	2.152	0.029	0.107
	19	7542	0.057	0.070	3.872	0.014	0.018	5522	0.032	0.321	2.249	0.043	0.125
	24	7226	0.044	0.062	3.850	0.011	0.016	4477	0.071	0.173	2.073	0.031	0.077
Hyperactivity subscale	0	8253	0.128	0.044	5.795	0.021	0.008						
	12	8042	0.087	0.049	6.174	0.014	0.008	5873	0.163	0.725	6.105	0.023	0.106
	19	7542	0.100	0.092	6.395	0.015	0.014	5522	0.341	0.727	6.132	0.047	0.106
	24	7225	0.089	0.047	6.479	0.013	0.007	4477	0.374	0.630	5.749	0.055	0.099
Peer problems subscale	0	8253	0.050	0.026	3.304	0.015	0.008						
	12	8042	0.060	0.024	3.435	0.017	0.007	5873	0.087	0.375	2.614	0.028	0.125
	19	7542	0.047	0.046	3.618	0.013	0.013	5522	0.087	0.453	2.334	0.020	0.163
	24	7225	0.056	0.011	3.563	0.015	0.003	4477	0.121	0.411	2.277	0.043	0.153
Prosocial behaviour subscale	0	8254	0.041	0.068	3.158	0.012	0.021						
	12	8042	0.067	0.045	3.341	0.019	0.013	5873	0.176	1.200	5.280	0.026	0.185
	19	7542	0.076	0.047	3.667	0.020	0.013	5522	0.266	1.177	5.358	0.039	0.180

Outcome	Measurement time (months) ¹	Pupil report						Teacher report					
		N	σ_s^2	σ_c^2	σ_e^2	ρ_s	ρ_c	N	σ_s^2	σ_c^2	σ_e^2	ρ_s	ρ_c
	24	7226	0.087	0.034	3.901	0.022	0.009	4477	0.619	0.933	5.147	0.092	0.154

¹ Time points at 12, 19 and 24 months adjusted for trial arm status

σ_s^2 : School-level component of variance

σ_c^2 : Class-level component of variance

σ_e^2 : Pupil-level component of variance

ρ_s : School-level ICC

ρ_c : Class-level ICC

Table 6.14. MYRIAD study intra-cluster correlation coefficients (ICCs) for pupil- and teacher-reported Behaviour Rating Inventory of Executive Function, Second Edition (BRIEF-2) outcomes at different time points

Outcome	Measurement time (months) ¹	Pupil report						Teacher report					
		N	σ_s^2	σ_c^2	σ_e^2	ρ_s	ρ_c	N	σ_s^2	σ_c^2	σ_e^2	ρ_s	ρ_c
BRIEF-2	12	7121	0.025	0.015	0.234	0.090	0.062	5898	26.419	97.887	491.598	0.043	0.166
	19	7022	0.018	0.027	0.235	0.065	0.103	5534	61.127	107.378	456.745	0.098	0.190
	24	6878	0.010	0.007	0.153	0.058	0.042	4479	57.848	85.448	426.158	0.102	0.167

¹ Time points at 12, 19 and 24 months adjusted for trial arm status

σ_s^2 : School-level component of variance

σ_c^2 : Class-level component of variance

σ_e^2 : Pupil-level component of variance

ρ_s : School-level ICC

ρ_c : Class-level ICC

Table 6.15. MYRIAD study intra-cluster correlation coefficients (ICCs) for the pupil-reported Revised Child Anxiety and Depression Scale (RCADS-30) outcomes at different time points

Outcome	Measurement time (months) ¹	N	σ_s^2	σ_c^2	σ_e^2	ρ_s	ρ_c
Total anxiety score	12	7585	12.347	0	387.102	0.031	0
	19	7175	12.837	5.484	433.151	0.028	0.013
	24	6987	13.237	1.623	449.604	0.028	0.004
Separation Anxiety Disorder subscale	12	7599	0.239	0.037	11.072	0.021	0.003
	19	7184	0.233	0.105	12.119	0.019	0.009
	24	6996	0.205	0	12.283	0.016	0
Generalised Anxiety Disorder subscale	12	7619	0.539	0.083	17.545	0.030	0.005
	19	7196	0.526	0.193	18.723	0.027	0.010
	24	7002	0.501	0.179	18.758	0.026	0.009
Panic Disorder subscale	12	7587	0.571	0	29.551	0.019	0
	19	7176	0.728	0.280	34.844	0.020	0.008
	24	6989	0.858	0.169	35.821	0.023	0.005
Social Anxiety subscale	12	7603	1.504	0.240	39.661	0.036	0.006
	19	7186	1.800	0.854	42.565	0.040	0.020
	24	6998	1.530	0.470	44.571	0.033	0.010
Obsessive Compulsive Disorder subscale	12	7606	0.219	N/A	12.752	0.017	N/A
	19	7191	0.251	0.115	13.872	0.018	0.008
	24	7001	0.246	0.017	13.993	0.017	0.001

¹ Time points at 12, 19 and 24 months adjusted for trial arm status

σ_s^2 : School-level component of variance

σ_c^2 : Class-level component of variance

σ_e^2 : Pupil-level component of variance

ρ_s : School-level ICC

ρ_c : Class-level ICC

Table 6.16. MYRIAD study intra-cluster correlation coefficients (ICCs) for the pupil-reported Centre for Epidemiologic Studies for Depression Scale, Warwick-Edinburgh Mental Well-being Scale, Child and Adolescent Mindfulness Measure, suicide ideation and self-harm outcomes at different time points

Outcome	Measurement time (months) ¹	N	σ_s^2	σ_c^2	σ_e^2	ρ_s	ρ_c
Centre for Epidemiologic Studies for Depression Scale (CES-D)	0	8370	1.518	0.932	95.276	0.016	0.010
	12	8054	2.862	1.173	118.292	0.023	0.010
	19	7561	2.570	2.662	131.886	0.019	0.020
	24	7238	2.660	2.084	136.147	0.019	0.015
Warwick-Edinburgh Mental Well-being Scale (WEMWBS)	0	8333	1.454	1.917	91.341	0.015	0.021
	12	8058	1.559	1.535	78.882	0.019	0.019
	19	7572	1.549	1.640	86.517	0.017	0.019
	24	7244	1.541	1.364	93.362	0.016	0.014
Child and Adolescent Mindfulness Measure (CAMM)	12	7924	1.175	0.314	60.508	0.019	0.005
	19	7472	1.626	1.064	65.942	0.024	0.016
	24	7171	1.483	0.678	71.924	0.020	0.009
Suicide ideation	12	6698	0.002	0.002	0.151	0.011	0.011
	19	6497	0.002	0.002	0.170	0.013	0.013
	24	6322	0.002	0.002	0.176	0.012	0.010
Self-harm	12	7232	<0.001	<0.001	0.075	0.006	0.005
	19	6820	0.001	0.001	0.093	0.011	0.011
	24	6598	0.001	0.002	0.101	0.005	0.017

¹ Time points at 12, 19 and 24 months adjusted for trial arm status

σ_s^2 : School-level component of variance

σ_c^2 : Class-level component of variance

σ_e^2 : Pupil-level component of variance

ρ_s : School-level ICC

ρ_c : Class-level ICC

Table 6.17. MYRIAD study intra-cluster correlation coefficients (ICCs) for the pupil-reported School Climate and Connectedness Survey (SCCS) outcomes at different time points

Outcome	Measurement time (months) ¹	N	σ_s^2	σ_c^2	σ_e^2	ρ_s	ρ_c
Total score	12	7805	0.021	0.017	0.458	0.042	0.036
	19	7332	0.020	0.013	0.490	0.039	0.026
	24	7087	0.016	0.010	0.495	0.032	0.019
School leadership and student involvement subscale	12	7843	0.036	0.030	0.677	0.048	0.042
	19	7355	0.053	0.021	0.759	0.064	0.026
	24	7117	0.051	0.021	0.772	0.060	0.027
Respectful climate subscale	12	7838	0.034	0.021	0.625	0.050	0.033
	19	7346	0.032	0.015	0.664	0.045	0.022
	24	7109	0.022	0.012	0.677	0.031	0.017
Peer climate subscale	12	7826	0.038	0.013	0.576	0.060	0.023
	19	7343	0.038	0.012	0.601	0.059	0.019
	24	7104	0.031	0.010	0.589	0.049	0.016
Caring adults subscale	12	7812	0.027	0.012	0.722	0.035	0.016
	19	7337	0.027	0.017	0.803	0.032	0.021
	24	7094	0.026	0.004	0.844	0.029	0.005

¹ Time points at 12, 19 and 24 months adjusted for trial arm status

σ_s^2 : School-level component of variance

σ_c^2 : Class-level component of variance

σ_e^2 : Pupil-level component of variance

ρ_s : School-level ICC

ρ_c : Class-level ICC

6.5.6 Comparison across studies

Teacher-reported Strengths and Difficulties Questionnaire (SDQ)

ICC estimates for the teacher-reported SDQ outcomes were obtained from the STARS, KiVa and MYRIAD datasets. It is difficult to draw comparisons across STARS and the other two studies as the school-level ICC for STARS combines school-level and class-level variation as only one class was sub-sampled from each school. When comparing KiVa (primary schools) with MYRIAD (secondary schools), the median school- and class-level ICC estimates were smaller in KiVa (0.00315 and 0.103, respectively) than MYRIAD (0.043 and 0.1535, respectively). In both studies, the smallest ICCs were for the *hyperactivity* and *conduct problems* subscales and ICCs were generally largest for the *prosocial behaviour* subscale.

Parent-reported Strengths and Difficulties Questionnaire (SDQ)

School-level ICC estimates were obtained from STARS and PACES (both studies recruited primary schools in South West England) for the parent-reported SDQ, although it should be noted that these two studies have a different cluster structure (i.e., in STARS the school-level ICC estimate incorporates school-level and class-level variation as only one class was sampled in each school). Almost all ICC estimates were smaller in PACES compared with STARS. For example, in PACES, only the *peer problems subscale* had an ICC estimate greater than 0 at baseline, compared with STARS where the ICC estimate at baseline ranged from 0 to 0.025. The smallest ICCs for both studies were for the *prosocial behaviour subscale*.

Pupil-reported bullying victimisation (Olweus Bully/Victim Questionnaire)

The KiVa and PACES studies (both undertaken in primary schools) measured bullying victimisation, reported by pupils. School-, class- and pupil-level components of variance were similar in PACES and KiVa. For example, at baseline, the variance component was 0.003 at the school level, 0.004 at the class level and 0.196 at the pupil level in PACES compared with 0.002 at the school level, 0.006 at the class level and 0.183 at the pupil level in KiVa).

Pupil-reported school climate and connectedness

Despite the studies using different measures, STARS (primary schools), MYRIAD (secondary schools) and PROMISE (secondary schools) all measured school climate and connectedness. Again, it is difficult to use STARS as a comparison as the school-level ICC estimate combines school-level and class-level variation. At the school level, ICC estimates were smaller in PROMISE (range 0.011 to 0.015) than MYRIAD (range 0.029 to 0.064), although it should be noted that PROMISE measured school connectedness while MYRIAD measured school climate subscale. The difference between studies might be partially explained by the fact that MYRIAD recruited schools from all over the United Kingdom, in contrast to PROMISE which recruited schools from just two regions in England (East Midlands and South West of England). Greater variation across schools might, therefore, be expected in the MYRIAD study.

Pupil-reported self-esteem (Rosenberg self-esteem scale)

The PACES (primary) and PROMISE (secondary) studies both used the Rosenberg self-esteem scale. Although, generally, ICCs were similar between the studies, the class-level ICC estimates were slightly larger in PACES than in PROMISE at the same time points. For example, at baseline, the class-level ICC was 0.043 in PACES compared with 0.027 in PROMISE.

Pupil-reported Revised Child Anxiety and Depression Scale (RCADS-30)

PACES (primary schools), PROMISE (secondary schools) and MYRIAD (secondary schools) all administered the pupil-reported RCADS-30. The median school-level ICC was smaller in PACES (0.002) and PROMISE (0.005) than MYRIAD (0.0245). The median class-level ICC was smallest in MYRIAD (0.006) compared with PACES (0.029) and PROMISE (0.022). In PACES and PROMISE the school-level variance component was smaller than the class-level component, but the opposite was seen in MYRIAD.

6.6 Discussion

This secondary data analysis estimated and explored patterns in the components of variance and ICCs from five UK school-based CRTs of interventions for improving social emotional functioning outcomes on pupils. One of the features explored was the difference in ICC estimates across reporters. Commonly, class-level ICCs for outcomes where the same teacher reported for all pupils in a given class, were larger than when the same outcomes were reported by parents or pupils. This was also the case at the school-level where ICCs reported by teachers were larger than those reported by parents. Similar findings were also seen in another primary school-based CRT in Northern Ireland where one class participated from each school and the class teacher reported outcomes for all children in their class [153], with ICCs for teacher-reported SDQ outcomes being of a similar size to the STARS study. These findings suggest that if the same person reports for all members of the cluster then ICCs are generally larger than if the individual pupil is reporting or if there are multiple outcome reporters in the cluster. Also, the within-cluster components of variance were smaller for teacher-reported outcomes compared with pupil- or parent-reported outcomes, but the between components were larger, suggesting there is less variation within clusters for teacher-reported outcomes. This may be because the same teacher will have a tendency to give similar ratings for all children in their given class. The variation may also be explained by the nature of the relationship with the pupil which differs clearly for parents versus teachers [320], which may cause teachers to rate pupils differently to parents. For some outcomes, the sample sizes varied between outcome reporters. For example, the sample size in PACES was larger for pupil report than parent report, due to missing data on the latter. This may account for the difference in the size of the ICCs between pupil- and parent-reported outcomes. Furthermore, even for the same outcome, different reporters will have different perspectives of how they subjectively rate an particular outcome, therefore the ICC would be expected to be different; Previous work has shown that there is substantial variation across outcome reporters [321].

In order to demonstrate how the ICC from different outcome reporters would affect the sample size estimate for a CRT, an example is provided. Using the targetted mean cluster size (19 pupils) from the sample size calculation in the

STARS study [161], and the ICC estimates for teacher- and parent-reported outcomes at baseline for the Total Difficulties Score from the SDQ from STARS (0.120 and 0.026, respectively), would result in a design effect of 3.16 for the teacher-reported ICC, and 1.47 for the parent-reported ICC. This would result in over twice as many individuals being required for a study where the ICC for teacher-reported SDQ was used compared to if the parent-reported SDQ was used.

Notable patterns were observed for the Strengths and Difficulties Questionnaire (SDQ) across studies. Teacher-reported ICC estimates for the SDQ at both the school and class levels were similar between KiVa and MYRIAD. ICCs were generally lower for the teacher-reported SDQ *conduct* and *hyperactivity* subscales compared to the other subscales. The Pupil Behaviour Questionnaire (PBQ) [301], which obtains teacher report on pupil conduct, had a similar ICC to the SDQ *conduct subscale*. The lower ICC for teacher report of pupil conduct may be due to there being less variation across teachers regarding their awareness of *conduct* and *hyperactivity* problems than there is for their awareness of *emotional difficulties*, *peer problems* and *prosocial behaviour* [322]. All teachers may be aware and concerned with challenging behaviours within schools and rate the presence of these behaviours more similarly than the other SDQ subscales. Schools have behavioural policies which may provide more “guidance” on how to handle conduct problems compared with other aspects of behaviour [323]. For studies that had teacher-reported SDQ outcomes, ICCs were generally largest for the *prosocial behaviour* subscale. This observation may be due to greater variation in the ability of teachers to recognise prosocial behaviour, and the fact that many of the aspects of it (e.g., helping, sharing, consoling and comforting) are exhibited more frequently outside of the classroom/learning environment where the teacher may not observe them [320]. Behaviours that are more difficult to observe and measure may be more susceptible to variation across teachers resulting in larger ICC estimates for these outcomes [320].

The ICC estimates varied greatly across studies, similar to the results from previous summarises [69, 123, 259], further highlighting the heterogeneity in the ICC from school-based CRTs. There were some consistencies in the size of ICCs with other studies that have collated ICCs. For example, a study in the US by

Dong and colleagues found that school-level ICCs for teacher-reported for prosocial behaviour measured by the Teacher Observation of Classroom Adaption - Checklist (TOCA-C) ranged from 0.29 to 0.54, while all other teacher-reported social emotional functioning outcomes had school-level ICCs ranging from 0.03 to 0.23 [69]. The findings are consistent with the observations in the current secondary data analysis in this chapter that suggest that ICCs for teacher-reported prosocial behaviour outcomes are generally larger than other teacher-reported social emotional functioning outcomes.

Hale and colleagues, using three large UK datasets, found that school-level ICCs for pupil-reported bullying victimisation ranged from 0.01 to 0.09 [259], while bullying victimisation outcomes from KiVa and PACES ranged from 0.005 to 0.031. Furthermore, the same study by Hale and colleagues found school-level ICCs for the pupil-reported SDQ ranged from 0.01 to 0.04 [259]. Data from the MYRIAD study suggested that school-level ICCs for the pupil-reported SDQ ranged from 0.011 to 0.022.

Shackleton and colleagues reporting survey findings from 21 European countries found that school-level ICCs for pupil-reported depressive mood ranged from 0.01 to 0.07 [71]. The school-level ICCs for pupil-reported RCADS-30 (measures symptoms of anxiety and low mood) in MYRIAD, PACES and PROMISE ranged from 0 to 0.04. School-level ICCs were larger for pupil-reported self-esteem which ranged from 0.01 to 0.08 in the paper by Shackleton and colleagues [71] compared with PACES and PROMISE where school-level ICCs ranged from 0 to 0.023.

Findings from the results of the study presented in Chapter 5 also suggest differences in the ICC across studies, with estimates for social emotional functioning ranging from 0.018 to 0.217 [123]. Of the studies undertaken in the United Kingdom [117], school-level ICC estimates for social emotional functioning outcomes ranged from 0.012 to 0.121, which are again of a similar magnitude to that seen in this chapter.

Few outcomes showed stability or consistent patterns in the change in ICC estimates over time. This may partially be a result of the change in sample size due to drop-out. However, there were exceptions. For example, in the STARS study, the ICC for pupil-rated 'How I Feel About My School' (HIFAMS) increased

over time (0.052 at baseline to 0.111 at 30-months). In KiVa, all school-level ICCs for the teacher-rated SDQ increased between baseline and 12 months (for example, the ICC for the total difficulties score increased from 0.037 to 0.075). ICC estimates may increase or decrease over time as the intervention may increase or reduce variability in the outcome across clusters.

Generally, components of variance and ICCs were larger at lower levels of clustering. For example, in the PROMISE study the ICCs at class level were larger than the ICCs at year group level, which in turn were larger than the ICCs at school level. Previous reports also suggest that the ICC will generally be larger when the natural cluster size is smaller [101, 261]. A notable exception in the current secondary analysis was the finding in the MYRIAD study that school-level ICCs were often larger than class-level ICCs for pupil-reported mental health outcomes (SDQ and RCADS). Not only did MYRIAD recruit the largest number of schools, but it was the only study that recruited schools from across the UK and included both state-maintained and independent school types; these aspects may explain why the school-level ICC was greater. On the other hand, the finding in MYRIAD may also be partially explained by a dominant school-level culture [324]. This stronger influence of school-level culture over class-level culture may be a particular feature in secondary schools as the pupils mix more across different classes. MYRIAD was undertaken in secondary schools while KiVa and PACES took place in the primary school setting, although PROMISE, which was undertaken in secondary schools, did not mirror the MYRIAD findings. In UK secondary schools, pupils may interact less within the same class as the membership of classes depends on academic attainment and subject choices. In contrast, for most primary schools, all lessons are delivered to classes of fixed membership. A higher school-level ICC for a given outcome might indicate that there is greater potential for a school-level intervention based on existing knowledge to improve the outcome [71]; the idea being that there is variation in the extent to which schools implement policies and existing guidance that are known to be good practice.

The findings in this study illustrate the fact that the number of classes subsampled in each school in a CRT can have an impact on the school-level ICC, if intermediate-level clustering is not incorporated in the analysis model. In STARS, where only one class was sampled from each school, the school-level ICC

estimates reported by the class teacher were large compared to the KiVa and MYRIAD studies. Large teacher-reported ICCs were also noted in another school-based CRT measuring social emotional functioning outcomes on pupils that sampled one class per school [153]. This may be explained by the fact that, in studies like STARS, the variation at school level cannot be separated from the variation at the class level. Comparison between STARS and KiVa, the latter study including pupils from more than one class from each school enabling the school- and class-level variance components to be separately estimated, revealed that the school-level ICCs for the teacher-reported SDQ in KiVa were markedly smaller than those in STARS which were of a similar size to the class-level ICCs from KiVa. On this basis, the school-level ICCs from the STARS study are best interpreted as class-level ICCs. This work highlights the importance of knowing the sampling approach used at each level of clustering in the study from which an ICC is calculated.

Despite using different measures, STARS, MYRIAD and PROMISE all included pupil-reported school climate and connectedness. Particularly in MYRIAD, ICC estimates for school climate outcomes were markedly larger than those for pupil-reported mental health outcomes. School-level ICCs for pupil-reported school climate ranged 0.029 to 0.064, compared with school-level ICCs for all other outcomes ranging from 0.005 to 0.04. A previous study by Bradshaw and colleagues in the US also noted that school-level ICCs ranged from 0.04 to 0.1 for the school climate outcomes [325]. The ICCs in MYRIAD for both pupil-reported and teacher-reported executive functioning (using the BRIEF-2) were also noticeably larger (ranging from 0.058 to 0.09 for pupil-reported executive functioning) than for reports of pupil mental health. Compared to mental health outcomes, school climate, school connectedness and executive functioning might be considered to be more directly impacted by the school environment [296], which may explain why the ICCs for such outcomes are larger than mental health outcomes.

6.7 Strengths and limitations

There are several strengths to the secondary data analysis. The ICCs were estimated using data from CRTs and, therefore, might be generalisable to future studies as the schools and participants are more likely to be representative of

those that take part in trials [3](p177). The studies included in the analysis were also undertaken relatively recently which means ICC estimates are likely to be relevant. A range of social emotional functioning outcomes were analysed meaning the findings from this work will be useful to many future trials. Furthermore, the settings of the studies span different UK regions and both primary and secondary school settings. Three of the five studies, however, drew their samples from the South West of England only, which may have resulted in underestimates of the school-level ICC relative to that in the UK as a whole. Conversely, they may provide relevant ICC estimates for planning trials that are to be undertaken in a single region or area of the country.

A key limitation is that it was only possible to include in this secondary analysis the five studies for which access to the data were available. Knowledge from other school-based CRTs with social emotional functioning outcomes [146, 150, 153, 157, 173, 198, 200, 294] would further enrich knowledge of the patterns in the ICC. Future research should aim to expand and replicate this work with additional datasets.

Another limitation is that all datasets were from UK studies which, whilst providing focussed and rich data in a specific setting, potentially limits the applicability of the findings internationally, especially to countries with different educational systems. However, the work in Chapter 5 revealed little evidence of marked differences in the size of the ICCs across world regions. The majority of ICCs fell within the range of ICCs (0 to 0.217) collated for social emotional functioning outcomes in Chapter 5 [123].

6.8 Implications

There are several implications from this study. First, researchers need to consider the design features and contexts of the studies from which they obtain ICC estimates for use in the sample size calculations for new studies. This chapter has provided empirical evidence of how sub-sampling one class from each school cluster can markedly increase the size of the observed school-level ICCs. If an ICC from this scenario was used to calculate the sample size for a study that includes a greater number of classes per school cluster, this may result in an overestimate of the number of pupils required. Researchers should consider this aspect when using previous estimates of the school-level ICC to calculate sample

size for future CRTs, but should also, when reporting their own ICC estimates, provide information on the size of the components of variance at all levels and describe the sampling approach used to recruit clusters at all levels.

Second, the individual reporting the outcome (i.e., pupil, parent or teacher) may have a marked impact on the size of the ICC. Particularly if all the pupil outcomes in any given class are reported by the same teacher, this by itself may lead to a larger ICCs as teachers will vary in the tendency to give low or high scores. This needs to be taken into consideration when specifying the ICC for sample size calculations in school-based CRTs.

6.9 Conclusions

The findings indicate that researchers need to take into consideration multiple factors when deciding on the ICC to assume in their sample size calculation for school-based CRTs measuring social emotional functioning outcomes on pupils. Generally, the class-level component of variance and ICC were larger than at the school-level. Class-level ICCs for teacher-reported outcomes where the same teacher reports on all pupils from a given class were larger than ICCs when the same outcomes were reported by the parents or pupils. The ICC values fluctuated across the timepoints for each study, with few outcomes showing consistent patterns in change in ICCs over time. ICCs for school climate, connectedness and executive functioning were generally larger than for mental health outcomes. More work needs to be undertaken to see if these patterns are consistent for other social emotional functioning outcomes and in educational settings in other countries.

6.10 Chapter summary

This chapter presented findings from a secondary data analysis calculating and investigating patterns in components of variance and ICCs from five UK school-based CRTs evaluating interventions for improving the social emotional functioning of pupils. The chapter adds to the knowledge gained about patterns in ICCs presented in Chapter 5. Chapter 7 discusses the findings and implications of this thesis as a whole and outlines areas of potential future work.

Page intentionally left blank

Chapter 7: Discussion

This closing chapter discusses the findings from the original research studies presented in Chapters 3 to 6, outlining how they add to the knowledge base of school-based cluster randomised trials (CRTs) of interventions for improving health outcomes on school pupils. The strengths and limitations are then discussed, followed by the implications of the findings for the design of future studies. Finally, the chapter outlines potential areas for future research in the field.

7.1 Chapter summary and contribution to knowledge

The original research in this thesis is comprised of systematic reviews of definitive (Chapter 3) and feasibility (Chapter 4) school-based CRTs measuring pupil health outcomes in the UK, a summary of ICC estimates from school-based CRTs worldwide (Chapter 5), and a secondary analysis of raw data from five UK school-based CRTs to estimate the ICCs and components of variance for social emotional functioning outcomes (Chapter 6). This section summarises the findings and discusses the contributions to knowledge made by the four studies (Chapters 3 to 6) and the thesis as a whole. This will focus specifically on two main areas: First, knowledge of the methodological characteristics of school-based CRTs measuring health outcomes on pupils and, second, knowledge of the ICC in such studies.

7.1.1 Summary and contribution to knowledge of methodological characteristics and design features of school-based cluster randomised trials measuring pupil health outcomes

Findings from the studies in Chapters 3 and 4 of the thesis contribute knowledge to the literature on methodological characteristics and challenges of undertaking school-based CRTs in the UK. The finding that the rate of publication of school-based CRTs evaluating interventions for improving health outcomes on pupils in the UK is increasing was also observed worldwide, when collating estimates of the ICC (Chapter 5). This builds on knowledge from the 2011 systematic review examining CRTs in children [53], which found an increase in the use of the CRT design in schools. Given the growing use of CRTs for evaluating the effect of

interventions on pupil health outcomes, describing the current methodological practices of school-based CRTs, as done in Chapters 3 and 4, is highly beneficial to future researchers as it highlights areas where improvements can be made.

The findings from the systematic reviews highlight room for improvement in the quality of reporting of CRT-specific elements in published articles. In the systematic review of definitive CRTs, 78% of studies reported accounted for clustering in their sample size calculation. This is an improvement from the previous review in 2011 by Walleser and colleagues of CRTs in children [53] which found that 59% of included studies reported how clustering was accounted for in the sample size calculation. The improvement may reflect the impact of the CONSORT-CRT extension published in 2012 [57]. Despite this, other elements were poorly reported. Of the studies included in the systematic reviews reported in Chapters 3 and 4, only 27% and 21%, respectively, provided the rationale for the use of the CRT design. This was slightly lower than the 32% of studies that provided the rationale in the Walleser review [53]. Only 3 of the 24 feasibility studies reported undertaking a formal sample size calculation, with only one of these allowing for clustering. Although there is no comparable review for school-based feasibility CRTs, a systematic review of feasibility CRTs in primary care found that only 17% allowed for clustering in their sample size calculation [221]. Low quality of reporting quality has been noted for CRTs [214, 221].

The findings in this thesis emphasise the need for better reporting of ICCs in order to inform sample size calculations for future school-based CRTs with pupil health outcomes. This was illustrated in the systematic review of definitive CRTs in the UK, where the assumed ICC for the primary outcome was often markedly different from the ICC estimated from the study data. This is consistent with findings in the wider CRT literature [209]. This may indicate lack of availability of relevant and precise estimates of the ICC at the time of sample size calculation. In the systematic review of definitive CRTs, 45% of studies reported the ICC for the primary outcome, with the figure rising to 55% (18/33) for studies published after the 2012 CONSORT-CRT extension [57]. The need for better reporting of the ICC has been stated by previous systematic reviews of CRTs, in children [53] and in primary care settings [112]. A review of CRTs in primary care found that only 44% of studies reported the ICC for the primary outcome [112]. Similarly, Walleser and colleagues' systematic review of CRTs in children found that only

37% reported this information [53]. This demonstrates that there may still not be sufficient information to inform sample size calculations, in school-based CRTs and in other settings.

Eighty percent of definitive and 54% of feasibility school-based CRTs in the UK used some form of restricted randomisation to balance cluster-level characteristics across the trial arms. The use of restricted randomisation is in line with recommendations [4, 6, 213]. Of the definitive CRTs that used restricted allocation in Chapter 3, 61% balanced the randomisation on the percentage of children who are eligible for free school meals, which is readily available information provided by the UK Department for Education [215]. Only 18% of studies that used restricted randomisation justified their choice of balancing factors. The best candidates for balancing randomisation are school-level characteristics that are predictive of the study outcomes, account for between-cluster variation, or influence effectiveness of the intervention [3, 216].

In the systematic review of definitive CRTs in the UK, only 6% of studies reported allowing for loss to follow-up of clusters in their sample size calculation. This is less than a previous review in primary care that found 38% of studies reported allowing for loss to follow-up of entire clusters in their sample size calculation [209]. It is surprising that so few allowed for loss of clusters in their sample size calculation given that almost half (48%) of studies in Chapter 3 reported losing at least one cluster to follow-up. The problem of missing data resulting from entire school drop-out was also discussed by authors of studies included in the systematic review [162, 199, 204]. The findings from Chapter 3 were better than two systematic reviews examining all types of CRTs with human participants which found that 31% [83] and 18% [210] of included studies reported having whole clusters missing from the primary analysis, respectively.

Not only is recruiting a sufficient number of schools and pupils important, but so is obtaining a representative sample as this enhances wider applicability of the findings and improves inclusiveness. Nearly two-thirds of definitive CRTs recruited schools from only one geographic region in the UK. Furthermore, information regarding the schools that declined to participate was often missing, and it was difficult to assess how representative the study schools were of the general population. Few feasibility studies described the baseline characteristics

of their clusters, a problem noted previously [221] as feasibility CRTs typically include a small number of clusters.

The findings from this thesis further illustrate challenges with recruiting pupils to school-based CRTs. The results from both Chapters 3 and 4 showed that only 46% of feasibility CRTs and 77% of definitive CRTs in the UK achieved their recruitment target of pupils. Facilitators and barriers to the recruitment and retention of pupils to school-based CRTs have been discussed in detail in the literature [81, 256, 257], noting issues such as a lack of time [81], a lack of incentivisation [256], and incompatibility of the intervention with the needs of pupils or parents, or with the school's ethos [81]. These are issues reiterated by authors of studies included in the systematic review reported in Chapter 3 [148, 157, 182, 183, 196]. Furthermore, many of the interventions assessed by school-based CRTs required a teacher or member of school staff to deliver them, and/or to report the outcomes. This requires extra resources, time and cost to schools and their staff, which may prevent certain types of schools from participating. Furthermore, pressures resulting from arranging examinations may be a reason why no sixth forms and colleges were included in the study populations of the trials summarised in Chapters 3 and 4. Also, pupils in year groups 11 (16 years old) and 13 (18 years old) may be more challenging to follow-up over longer durations as they leave the education system. The exclusion of these pupils from the study population may undermine representativeness. For example, in the context of social emotional functioning outcomes, this is a key age for the development of social skills [326]. Additionally, a large proportion of pupils attend further education in the UK; According to the 2021/22 school census, 415,185 pupils in England were enrolled at sixth forms [327]. School-based CRTs could potentially be missing out of information from these pupils.

Methodological characteristics and challenges specific to school-based feasibility CRTs were described in the thesis. The work in Chapter 4 found that the median sample size of feasibility studies in the UK (8 clusters) is large enough to estimate pupil-level feasibility parameters (e.g., percentage followed up) with reasonable precision. This was demonstrated using the formula by Eldridge and colleagues [44] for calculating the sample size required in feasibility CRTs to estimate percentages of individual-level characteristics with a confidence interval of specified width whilst allowing for clustering. Using the average sample size from

the review (i.e., 8 schools and 240 pupils), and assuming an ICC of 0.05, a study of this size would be large enough to estimate percentages for pupil-level characteristics with a margin of error no greater than 10 percentage points based on a 95% CI. This was despite only 3 (13%) studies included in the review performing a formal sample size calculation. The median number of clusters recruited was similar to a previous review in primary care [221]. The thesis found that few feasibility CRTs explored challenges specific to CRT design. Only one of the 24 feasibility CRTs in Chapter 4 used the study to assess whether a cluster design was needed, which was similar to a previous review of feasibility CRTs measuring health outcomes that found only 2 out of 18 CRTs tested the feasibility of the CRT design [221]. Additionally, in the systematic review of feasibility CRTs, none of the studies assessed which type of cluster (e.g., school versus classes) was best to randomise, and none of the studies that randomised clusters before recruiting pupils explored the possibility of recruitment bias as an objective.

7.1.2 Summary and contribution to knowledge of the ICC for pupil health outcomes from school-based cluster randomised trials

The thesis contributes knowledge important to the understanding of ICC estimates for pupil health outcomes from school-based CRTs. First, as discussed in the previous section, the findings from the systematic reviews in Chapters 3 and 4 illustrated the need for better reporting of ICC estimates for pupil health outcomes from school-based CRTs. In order to address the relative lack of ICC estimates for pupil-health outcomes from school-based CRTs, Chapter 5 provided a summary of ICC estimates for pupil outcomes across a range of health areas and across world regions. Generally, other summarises of ICCs focussed on one health area, and most of these used data from studies in the United States [69, 96-103, 105-109]. The median (IQR; range) ICC was 0.031 (0.011 to 0.08; 0 to 0.47) at the school level (N=210) and 0.063 (0.024 to 0.1; -0.009 to 0.262) at the class level (N=46). The findings suggest that ICCs are generally larger for class than school clusters. Higher ICCs at the class level than school level were also observed in Chapter 6 for social emotional functioning outcomes. This is in keeping with previous findings where authors have reported that ICC estimates are generally larger when the natural cluster size is smaller [101, 260, 261]. Two-thirds of school-level ICCs were no greater than 0.05 and three-quarters were

under 0.08. The distribution of school-level ICCs worldwide was comparable to earlier summaries of school-based ICCs for pupil health outcomes [69, 96-103, 105-109].

The findings from Chapter 5 demonstrated that the ICC varies greatly across school-based CRTs, similar to the results from previous summaries [69, 259]. The school-level ICC estimates are well described by the beta and exponential distributions. The distribution of ICCs in primary care health studies has also previously been noted to follow a beta distribution [26, 287]. Additionally, the work in Chapter 5 compared the distribution of ICCs across categories defined by world region, health outcome area and educational level. The ICC distribution was similar to previous papers reporting summaries of the ICC, the majority of which were undertaken in the US [69, 96-103, 105-109].

The work in Chapter 5 found few relationships between the size of the ICC and study design features. Firstly, there was little evidence of a relationship between the ICC and both health outcome area and world region, respectively. This was in contrast to the findings from a previous paper that suggests that estimates of ICCs from school-based CRTs are both outcome- and country-specific [71]. Further to this, the work in Chapter 5 found no relationship between the ICC and educational level, again in contrast to previous work that showed that the ICC for educational outcomes tend to be larger at lower attainment grades [88]. The findings in Chapter 5, however, did indicate that ICCs are larger in definitive trials than feasibility studies. This might be expected because schools recruited to feasibility studies may be more restricted and less representative of the wider types of schools that are recruited in larger definitive CRTs [3](p180/181).

The secondary analysis of raw data from school-based CRTs with social emotional functioning outcomes in the UK in Chapter 6, revealed that school- and class-level ICCs were larger if the same teacher reported outcomes for all pupils in a given class compared to if the parents or pupils reported the same outcome. This is to be expected as there will be more correlation within clusters as teachers will have an underlying tendency to provide high or low ratings for all pupils [321]. Similar findings were also seen in another primary school-based CRT in Northern Ireland where one class was sampled from each school and the same class teacher reported outcomes for all children [153].

The thesis found that subsampling a small number of classes from each school can result in an inflated school-level ICC when the intermediate-level clustering is not formally incorporated in the analysis model. In Chapter 6, school-level ICCs were larger for the STARS study [161] that sampled one class from each school compared with others that included pupils from multiple classes from each school. This was particularly true when the same class teacher reported the outcomes for all pupils within the class. In the case of the STARS study, the variation at school level cannot be separated from the variation at the class level. Larger teacher-reported ICCs were seen in another school-based CRT by Connolly and colleagues that measured social emotional functioning outcomes on pupils and subsampled one class per school [153].

The findings from Chapter 6 highlighted differences in the size of ICCs for teacher-reported measures across specific social emotional functioning outcomes. This was particularly notable for the Strength and Difficulties Questionnaire (SDQ) *conduct* and *hyperactivity* subscales, where teacher-reported school- and class-level ICCs were generally smaller than for the other subscales. In contrast, ICCs were generally largest for the *prosocial behaviour* subscale. The findings are consistent with another study that found that the teacher-reported ICCs for concentration problems and disruptive behaviours in pupils were generally smaller than for prosocial behaviour [69]. Hyperactivity and conduct are externalising behaviours that are observable while prosocial and emotional require the assessment of intent and internal phenomena, respectively [328]. Teacher assessments on hyperactivity and conduct might, therefore, be expected to show less variation across teachers.

The secondary analyses of data from school-based CRTs indicated that ICCs for school climate and connectedness and executive functioning were markedly larger than ICCs for pupil mental health outcomes. In the MYRIAD study [298], for example, school-level ICCs for pupil-reported school climate ranged from 0.029 to 0.064, compared with school-level ICCs for all other pupil-reported outcomes, which ranged from 0.005 to 0.04. A study by Bradshaw and colleagues in the US also noted that school-level ICCs ranged from 0.04 to 0.1 for the school climate outcomes [325]. In the MYRIAD study, the ICC for pupil-reported executive functioning (using the BRIEF-2) (which ranged from 0.058 to 0.09) were also noticeably larger than other pupil-reported outcomes.

A list of key findings from the work undertaken in this thesis is presented in Table 7.1.

Table 7.1. Key findings from this thesis

Key findings
<p>The rate of publication of school-based CRTs evaluating interventions for improving health outcomes in pupils is increasing.</p> <ul style="list-style-type: none"> • Both UK and worldwide
<p>Quality of reporting of CRT-specific elements in published articles</p> <ul style="list-style-type: none"> • Of the studies included in the systematic reviews presented in Chapters 3 and 4, only 27% and 21% provided rationale for the use of the CRT design, respectively. • In the case of feasibility CRTs, only 3 of the 24 (13%) studies reported undertaking a formal sample size calculation, with only one of these allowing for clustering.
<p>Reporting of ICCs</p> <ul style="list-style-type: none"> • The assumed ICC was often markedly different from the ICC estimated from the primary outcome. • In Chapter 3, 45% of studies reported the ICC for the primary outcome.
<p>80% of definitive and 54% of feasibility school-based CRTs used some form of restricted randomisation to balance specific cluster-level characteristics across the trial arms.</p>
<p>In Chapter 3, only four (6%) definitive CRTs reported allowing for loss to follow-up of clusters in their sample size calculation.</p>
<p>48% of studies in Chapter 3 reported losing least one cluster to follow-up.</p>
<p>In Chapter 3, 63% of studies recruited schools from only one geographic region.</p>
<p>46% of feasibility CRTs and 77% of definitive CRTs achieved their recruitment target of pupils.</p>
<p>In Chapter 4, the average sample size of feasibility studies included would be large enough to estimate pupil-level feasibility parameters (e.g., percentage followed up) with reasonable precision, despite only 3 (13%) studies performing a formal sample size calculation.</p>

<p>Few feasibility CRTs explored challenges specific to CRT design. Only one (4%) study assessed whether a cluster design was needed, no studies assessed which type of cluster (e.g., school versus classes) was best to randomise, and none of the studies that randomised clusters before recruiting pupils explored the possibility of recruitment bias as an objective.</p>
<p>Chapter 5 provided a summary of 260 ICC estimates for pupil outcomes across a range of health areas and across world regions. The median (IQR; range) ICC was 0.031 (0.011 to 0.08; 0 to 0.47) at the school level (N=210) and 0.063 (0.024 to 0.1; -0.009 to 0.262) at the class level (N=46). Two-thirds of school-level ICCs were no greater than 0.05 and three-quarters were under 0.08.</p>
<p>ICC estimates are generally larger for class than school clusters.</p>
<p>The distribution of school-level ICCs worldwide was comparable to earlier summaries of school-based ICCs for pupil health outcomes.</p> <ul style="list-style-type: none"> • The ICC distribution was similar to studies in the US.
<p>The size of the ICC varies greatly across studies.</p>
<p>School-level ICC estimates are well described by the beta and exponential distributions.</p>
<p>There was little evidence of relationships of the ICC with health outcome area and educational level.</p>
<p>The findings in Chapter 5 suggested that ICCs are larger in definitive trials than feasibility studies.</p>
<p>School and class-level ICCs were larger if the same teacher reported outcomes for all pupils in a given class compared to if the parents or pupils reported the same outcome.</p>
<p>Chapter 6 showed that school-level ICCs were larger for the STARS study that sampled one class from each school compared with others that did not subsample classes within schools.</p>
<p>Size of teacher-rated ICCs for the Strength and Difficulties Questionnaire (SDQ) <i>conduct</i> and <i>hyperactivity</i> subscales were generally smaller than the other subscales. ICCs were generally largest for the <i>prosocial behaviour</i> subscale.</p>
<p>ICCs for school climate and connectedness and executive functioning were markedly larger than ICCs for pupil mental health outcomes.</p>

7.2 Strengths and limitations

The work presented in this thesis benefits from using different methodological approaches. The strengths and limitations have previously been discussed in detail in the respective chapters. This section summarises the key points and then reflects on the fundamental strengths and limitations of the thesis as a whole.

The systematic review studies in Chapters 3 and 4 used the best approach for synthesising empirical evidence and summarising the characteristics of published school-based CRTs in the UK. The methods for both systematic reviews were developed with the advice of information specialists and were registered on PROSPERO [329]. The search strategy was developed using an existing strategy that had been shown to provide good precision for identifying CRTs [127]. A strength of the systematic reviews was their broad scope with evidence synthesised across different school settings, educational levels, health areas and outcomes.

The systematic reviews were limited by time and resource constraints and the search was limited to school-based CRTs in the UK. However, the reviews were thus able to focus on schools within a single system. Despite being focussed on the UK, the findings of the reviews will still be of global interest. Other high and upper/middle income countries such as Australia have a similar school system to the UK, and some of the findings may have applicability to those settings. Some methodological challenges in the design of cluster randomised trials will be similar across different settings.

The systematic reviews only searched the MEDLINE database. MEDLINE was chosen because the research question for both systematic reviews was to describe the characteristics of trials that evaluated the impact of health interventions on health outcomes. Extensive scoping revealed that the size of the literature was considerable, and a pragmatic decision was made to examine MEDLINE exclusively in order to align with available resources. Scoping searches of other databases only resulted in one additional eligible definitive CRT article; therefore, it was considered inefficient use of time to search other databases.

The collation of ICC estimates from school-based CRTs worldwide presented in Chapter 5 shared much of the same methodology with Chapters 3 and 4 in terms of search strategy and identifying school-based CRTs, and similarly shared the same strengths and limitations. The work in Chapter 5 did not, however, limit the search to the UK and instead identified school-based CRTs worldwide which is a considerable strength. By including CRTs from around the world, this study was able to explore the difference between world regions and incorporate knowledge from different school systems into the thesis.

A key strength of the thesis is that it draws on data from both published sources and raw datasets. Published data are easily accessible and allowed the thesis to include data from a range of CRTs in different contexts, making the findings more generalisable. The use of raw datasets in Chapter 6 allowed for a more indepth exploration of patterns in the ICC.

A limitation of the systematic review of feasibility CRTs in the UK was that the search strategy only aimed to identify randomised feasibility studies, potentially overlooking other useful studies that may be undertaken to assess whether a definitive CRT can be undertaken. Non-randomised comparative studies [330] and single-arm feasibility studies [331] may provide useful knowledge for planning a school-based CRT. Only randomised feasibility studies were included as the systematic review aimed to summarise studies whose feasibility objectives potentially covered the full range of uncertainties relevant to planning definitive trials, including testing the randomisation process and testing the willingness of participants to be randomised. However, with time and resources, other types of feasibility study would have been included in the systematic review and provided further insight for researchers in the area.

The thesis was limited by available time and resources. This meant that only one database was searched for potentially eligible CRTs for the work undertaken in Chapters 3, 4 and 5, although this was not thought to reduce the sensitivity of the search significantly. Further to this, mainly due to time constraints, only one ICC estimate from each CRT was extracted in the work detailed in Chapter 5. Summarising every ICC in every paper would have been comprehensive, but multiple ICCs from the same studies would likely be correlated with each other and would have added relatively little additional information to those analyses.

A limitation of the thesis was that the data collected from published research articles for the studies reported in Chapters 3, 4 and 5, were often poorly reported, resulting in missing data. This also meant that judgements were often made regarding extracted data (outlined in Sections 3.4.5, 4.4.5, 5.4.5) and classification of data may be considered crude for some variables. The work in the thesis was also limited by the number of eligible studies that were available. For example, in Chapter 5 the small number of articles with specific study-level characteristics may have limited the ability to detect differences in the ICC across subgroups defined by different combinations of study design characteristics (e.g., health area and world region). Partly for this reason, in Chapters 3, 4 and 5, the aspects of social emotional functioning (which included mental health, neurodiversity, behaviour, bullying, self-esteem, school connectedness and executive functioning) and countries had to be grouped into broader categories, which may have obscured some differences. The findings in Chapter 6, based on secondary analyses of raw data from school-based CRTs, showed that the size of pupil-reported ICCs for behavioural and executive functioning outcomes may be larger than for anxiety and depressive symptoms (in the STARS [161], PROMISE [202] and MYRIAD [298] studies).

The secondary analyses in Chapter 6 were limited in that data from only five CRTs could be accessed. The included datasets do, however, provide data from different education levels, countries and outcomes. All five CRTs were undertaken in the UK. While this may limit the generalisability of the findings to other educational systems, the results are relevant to the UK. The limit in available datasets also meant that the study focussed on social emotional functioning outcomes. However, given the lack of information regarding ICCs for social emotional functioning outcomes and the marked increase in the rate of publication of definitive CRTs measuring such outcomes on school pupils, this work will be useful to many researchers. Additionally, many outcomes were also not included in the datasets such as body image measures [157] and The Reported and Intended Behaviour Scale (RIBS) [150, 332].

A limitation of the systematic review of feasibility CRTs in the UK was that the search strategy only aimed to identify randomised feasibility studies, potentially overlooking other useful studies that may be undertaken to assess whether a definitive CRT can be undertaken. Non-randomised comparative studies [330]

and single-arm feasibility studies [331] may provide useful knowledge for planning a school-based CRT. Only randomised feasibility studies were included as the systematic review aimed to summarise studies whose feasibility objectives potentially covered the full range of uncertainties relevant to planning definitive trials, including testing the randomisation process and testing the willingness of participants to be randomised. However, with time and resources, other types of feasibility study would have been included in the systematic review and provided further insight for researchers in the area.

A potential limitation of the thesis is the use of the term 'social emotional functioning' to define all health outcomes associated with mental health, behaviour and wellbeing. This term was selected as it represents the capacity to understand, experience, express, and manage emotions and to develop meaningful relationships with others [124]. There may be a lack of clarity about what health conditions should be included under this definition. Additionally, different people may have different ideas about what this term means and how it may be interpreted.

7.3 Implications

The work discussed in this thesis has several important implications for future research. First, the increasing number of published school-based CRTs provides researchers with an increasing pool of information to aid the design and conduct of future studies. The work in the thesis summarising the current methodological practices of school-based CRTs measuring health outcome on pupils has helped to identify areas where improvements can be made. A limitation of current school-based CRTs is the poor reporting of CRT-specific elements. This includes the need to justify using the CRT design, allowing for potential loss to follow-up of clusters when calculating sample size, and reporting the ICC for the study outcomes. These findings are important as they highlight areas where reporting and execution of these studies need to be improved and should be made in line with recommendations and reporting guidelines such as the CONSORT-CRT statement [57]. Better reporting will help improve the quality of published school-based CRTs and the pool of knowledge for future researchers to use in designing CRTs.

There are challenges with obtaining representative samples of schools to school-based CRTs. The work in Chapter 3 found that recruitment is often limited to one or two geographical regions and one school type. There was a lack of information to assess the representativeness of schools included in the studies in the systematic reviews compared with the general population. Additionally, few feasibility studies described the baseline characteristics of the clusters, adding to the difficulties in describing the type of schools that took part in such studies. A representative sample of schools is important in school-based CRTs as the findings of the study will then be more applicable and generalisable to a wider range of schools in the study population. Researchers should aim to improve inclusivity and recruit subgroups of pupils and schools that are known to be less likely to be involved in research. Underlining these challenges will encourage researchers to improve the diversity of their recruitment at both the cluster and individual level. Researchers should also be encouraged to publish the information needed to be able to assess the representativeness of their sample.

Although studies often use restricted randomisation to balance cluster-level factors, few provided a rationale for their choice of balancing factors. Chapter 3 summarised the common cluster-level factors that were used to balance randomisation in school-based CRTs that used restricted randomisation. This will help to inform future work identifying the best cluster-level characteristics on which to balance the randomisation.

Furthermore, providing a summary of relevant parameter values, such as the number of clusters, number of classes per school, and number of pupils per school, will help to inform simulation studies for evaluating the statistical properties of methods used to analyse data from school-based CRTs.

There has been a lack of knowledge of suitable values to assume for the ICC when undertaking sample size calculations for school-based CRTs with pupil health outcomes. This was illustrated by the marked difference between the ICC assumed in the sample size calculation and the ICC estimated from the resulting study data in the systematic review of school-based CRTs. Additionally, the review generally found poor reporting of ICC estimates from such studies. These findings are important because poorly specified values of the ICC at sample size calculation lead to poor estimates of the required number of schools and pupils for the trials. This knowledge will encourage researchers to publish ICC estimates

from their studies to help inform sample size calculations for subsequent school-based CRTs.

The thesis collated over 200 estimates of the ICC from school-based CRTs. These will help to convey the uncertainty regarding the value of the ICC that should be assumed in a sample size calculation; this uncertainty should be acknowledged when undertaking the sample size calculation [333]. This is important as ICC estimates from a single study may have little generalisability [108], and the description of the distribution of ICC estimates provided in this thesis indicates plausible values with which to model the sensitivity of sample size calculations [3, 26, 286]. This information can also be used to construct informative priors to incorporate uncertainty regarding the ICC when undertaking sample size calculations and analyses of the intervention effect that use the Bayesian framework [121, 288].

The thesis provides further information regarding the need to consider design features and contextual factors when interpreting ICC estimates. Sampling one class from each school was shown to lead to markedly larger school-level ICC estimates. This is important, as an ICC from this scenario would not be relevant for calculating the sample size for a trial that subsamples multiple classes or includes all eligible classes from each school, and vice-versa. ICCs also vary between different outcome reporters; notably ICCs for teacher-reported outcomes can be relatively large when the same teacher reports on all pupils within a given class. Researchers should consider these aspects when using ICC estimates from previous CRTs as they can have a marked impact on the extent to which the sample size is inflated to allow for clustering. When reporting findings from their own CRTs, researchers should provide relevant contextual information with their ICC estimates. This would include: the sampling approach used to recruit clusters at all levels; the size of the components of variance at all levels; whether the ICC was adjusted for prognostic factors and, if so, which ones; whether the ICC was estimated from a repeated measures analysis.

A list of key recommendations resulting from the findings of the thesis are listed in Table 7.2.

Table 7.2. Key recommendations for future research based on the findings from the thesis

Recommendations arising from thesis findings	
1	Provide justification for the ICC value assumed in the sample size calculation. If the ICC is from a previous school-based CRT, this should be referenced.
2	Researchers should consider the potential impact that subsampling of classes within schools can have on the size of school-level ICCs, when using ICC estimates from previous studies to inform the sample size calculation in a new study. When reporting the results of school-based CRTs, information on subsampling of classes should be provided to help contextualise the ICC estimates from the study.
3	Researchers should consider the potential impact the outcome reporter (i.e., pupil, teacher, parent) can have on the size of the ICC for social emotional functioning outcomes on pupils, when undertaking a sample size calculation for a planned school-based CRT.
4	Researchers should provide justification for their choice of factors used to balance or restrict the randomisation in school-based CRTs.
5	Reports of school-based CRTs with health outcomes should include estimates of the components of variance and the ICC/CV at all levels of clustering for all outcomes.
6	When reporting estimates of the ICC, researchers should provide details of the analysis used for the calculation, including: which analysis method was used (e.g., mixed effects model); which prognostic factors were adjusted for; whether the ICC was estimated from a repeated measures analysis; whether the ICC was calculated based on change scores in the outcome.

7.4 Future research

Although, the study presented in Chapter 5 collated 260 ICC estimates of pupil health outcomes from school-based CRTs, more estimates are needed. The database of ICCs hosted by the University of Aberdeen does not provide ICCs for school-related clusters and pupil health outcomes [73]. The development of a repository of ICCs with contextual details for a range of pupil health outcomes from school-based CRTs and survey data would provide researchers with a wealth of information to inform the design of future trials. Further research is also needed to examine whether the size of the ICC differs across subgroups defined by study design characteristics and this requires a larger sample of ICC estimates than were available for the analyses in this thesis.

Four-fifths of definitive UK-based CRTs included in the systematic review described in Chapter 3 used some form of restricted randomisation to balance cluster-level characteristics across trial arms. When analysing outcomes from CRTs, adjusting for the characteristics used to balance the allocation provides greater precision for estimating the intervention effect if those characteristics are predictive (i.e., prognostic) of the trial outcomes [3](p76-78). Balancing the randomisation on factors that have no association with the outcome makes the process more complicated with no improvement in efficiency [3](p76-78). It has been suggested that randomisation of clusters may be balanced on cluster size, geographical location, characteristics of the cluster population (e.g., a socio-demographic measure) and a cluster-level summary of the baseline measure of the outcome [3](p76-78). Chapter 3 found that the most common factors used to balance the randomisation in school-based CRTs were deprivation, cluster size, geographical location, pupil ethnicity, co-educational status, and school performance [117]. There is little evidence, however, on which factors *should* be balanced on this setting to improve the efficiency of the study. In the systematic review of definitive school-based CRTs in the UK, only 18% of studies provided justification for the choice of balancing factors. Mixed effects (“multilevel”) regression models could be fitted to secondary data from previous school-based CRTs to examine the relationships between candidate cluster-level characteristics and pupil health outcomes and identify the ones that account for the intra-cluster correlation. Such factors would potentially be useful as balancing

factors in the randomisation and as adjustment factors in the analysis of the intervention effect.

The findings in Chapter 6 also need to be explored using larger data sets, other outcome measures and for other health areas to understand if the patterns observed for the ICCs of social emotional functioning outcomes are similar to those of ICCs from other school-based CRTs measuring health outcomes on pupils.

7.5 Closing remarks



The thesis provides a wealth of information relating to ICCs for use in sample size calculations for future school-based CRTs. The thesis has outlined the complexities and nuanced nature of selecting an ICC for use in a sample size calculation; it is not as simple as just selecting an ICC for the outcome you wish to measure. Researchers must also take into account contextual factors, such as: subsampling of classes/year groups within schools; who reported the outcome (e.g., teacher or pupil); and aspects of the study methodology (e.g., method of analysis). Researchers may wish to use distributions of ICCs to inform the use of Bayesian methods to help calculate sample size for CRTs. The findings from this thesis, including information about the numbers of school, number of classes per school, number of pupils per school and ICCs, can also be used in simulation studies to evaluate the properties of statistical methods in the context of CRTs. There is still much more to learn regarding the factors that influence the size of ICCs from school-based CRTs, as highlighted by the areas suggested for future work. This research should be communicated to researchers, policy makers and the public in order to help them better understand the importance of factors influencing the size of ICCs and the need for accurate sample size calculation.

Page intentionally left blank

Appendices

Appendix 1 - Characteristics and practices of school-based cluster randomised controlled trials for improving health outcomes on pupils in the UK: a systematic review protocol

BMJ Open Characteristics and practices of school-based cluster randomised controlled trials for improving health outcomes in pupils in the UK: a systematic review protocol

Kitty Parker ¹, Michael P Nunns,² ZhiMin Xiao,³ Tamsin Ford,⁴ Obioha C Ukoumunne ¹

To cite: Parker K, Nunns MP, Xiao Z, *et al.* Characteristics and practices of school-based cluster randomised controlled trials for improving health outcomes in pupils in the UK: a systematic review protocol. *BMJ Open* 2021;11:e044143. doi:10.1136/bmjopen-2020-044143

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2020-044143>).

Received 26 August 2020
Revised 07 December 2020
Accepted 03 February 2021



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY. Published by BMJ.

¹NIHR ARC South West Peninsula (PenARC), University of Exeter, Exeter, UK

²College of Medicine and Health, University of Exeter, Exeter, UK

³Graduate School of Education, University of Exeter, Exeter, UK

⁴Department of Psychiatry, University of Cambridge, Cambridge, UK

Correspondence to
Kitty Parker;
kp477@exeter.ac.uk

ABSTRACT

Introduction Cluster randomised trials (CRTs) are studies in which groups (clusters) of participants rather than the individuals themselves are randomised to trial arms. CRTs are becoming increasingly relevant for evaluating interventions delivered in school settings for improving the health of children. Schools are a convenient setting for health interventions targeted at children and the CRT design respects the clustered structure in schools (ie, pupils within classrooms/teachers within schools). Some of the methodological challenges of CRTs, such as ethical considerations for enrolment of children into trials and how best to handle the analysis of data from participants (pupils) that change clusters (schools), may be more salient for the school setting. A better understanding of the characteristics and methodological considerations of school-based CRTs of health interventions would inform the design of future similar studies. To our knowledge, this is the only systematic review to focus specifically on the characteristics and methodological practices of CRTs delivered in schools to evaluate interventions for improving health outcomes in pupils in the UK.

Methods and analysis We will search for CRTs published from inception to 30 June 2020 inclusively indexed in MEDLINE (Ovid). We will identify relevant articles through title and abstract screening, and subsequent full-text screening for eligibility against predefined inclusion criteria. Disagreements will be resolved through discussion. Two independent reviewers will extract data for each study using a prepiloted data extraction form. Findings will be summarised using descriptive statistics and graphs.

Ethics and dissemination This methodological systematic review does not require ethical approval as only secondary data extracted from papers will be analysed and the data are not linked to individual participants. After completion of the systematic review, the data will be analysed, and the findings disseminated through peer-reviewed publications and scientific meetings.

PROSPERO registration number CRD42020201792.

Strengths and limitations of this study

- To our knowledge, this is the first systematic review to describe the characteristics and methodological practices of school-based cluster randomised trials (CRTs) of health interventions in the UK.
- The review has a defined search strategy that is tailored to identifying school-based CRTs, eligibility criteria, and prepiloted screening and data extraction strategies to minimise inaccuracies.
- Two independent reviewers will perform screening and data extraction, with any uncertainty resolved by consulting a third reviewer.
- The review will focus on studies conducted in the UK in order to align with available resources and create a relevant and focused review.
- There is the possibility that we are missing the opportunity to learn from studies in countries that have a similar education system to the UK.

INTRODUCTION

Cluster randomised trials (CRTs), also known as group randomised or place randomised trials, are studies in which groups (clusters) of participants (eg, general practices, organisations, areas, etc) are randomly allocated to the trial arms, rather than the individual participants on whom outcomes are measured.¹ These studies are in contrast to the more traditional individually randomised trials, where the participants themselves are randomised. The CRT design is commonly used in healthcare research when interventions must be delivered at the cluster level and to minimise contamination of the trial arms that might otherwise occur when individuals are randomised.²

A characteristic feature of CRTs is that observations on participants who are in the same cluster are usually more similar than



observations on participants who are from different clusters.² For example, patients registered with the same general practice are more likely to have similar health outcomes than those registered with different practices.³ This similarity, or lack of statistical independence, between observations from the same cluster means that the usual procedures for calculating sample size and analysing data in individually randomised trials should not be used in CRTs.¹ The use of standard sample size methods are likely to result in studies that lack power to detect the specified intervention effect and the use of standard analytical methods may produce results that exaggerate evidence for the true effect of the intervention. Therefore, alternative methods should be used when conducting CRTs.

The intracluster correlation coefficient (ICC), denoted ρ , quantifies the similarity of observations of individuals within the same cluster. The ICC is the proportion of the total variability in the trial outcome that is between clusters (σ_b^2) as opposed to between individuals within clusters (σ_w^2):⁴

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

ρ can take values between 0 and 1. The larger ρ is, the greater the similarity between individuals within clusters, or equivalently the greater variability between clusters.

Information about ρ for the primary outcome is invaluable when designing a CRT. It can be estimated from previous studies or feasibility studies, with a similar cluster structure and outcome to the planned trial.^{5,6} Authors of CRTs should report estimates of their ICCs, ideally with confidence intervals (CIs) because they are usually based on studies with relatively few clusters,⁷ to aid the design of future similar studies.

When calculating the sample size in CRTs one needs to determine the total number of clusters that need to be recruited and the number of individuals that need to be recruited from within each cluster. Methods for calculating the required sample size need to take the ICC into account. When the number of participants in each cluster is fixed and known in advance, the total number of individuals required in a CRT is calculated by inflating the number of individuals that would be required in an individually randomised trial by the *design effect* (DE), which is function of ρ :

$$DE = 1 + (n - 1)\rho$$

where n is the number of participants providing outcome data in each cluster (cluster size).⁸ Having calculated the total number of participants required, this is divided by the cluster size to obtain the total number of clusters that is required. For scenarios where the total number of clusters available for a trial is known, an alternative calculation based on the same approach is used to calculate the number of participants that need to be recruited from each cluster.³

When estimating the intervention effect in CRTs, analytical methods that take account of ρ should be

used,¹ otherwise CIs will be too narrow and p values will be too small, resulting in an exaggeration of the amount of evidence for a true intervention effect.⁹ Furthermore, the degrees of freedom (df) used for calculating the CI and p value for the intervention effect should take account of the number of clusters.^{10,11} A distinction can be made between statistical analyses that are carried out at the cluster level and those that are carried out at the individual level.⁸ For cluster-level analyses, the outcome is summarised for each cluster, for example, by calculating the mean for continuous outcomes or percentages for binary outcomes across individuals in the cluster. Standard analytical methods are then used to compare the outcome between the trial arms using the cluster-level summary statistics as the observations. This method of analysis is valid because the cluster is both the unit of randomisation and the unit of analysis.² Alternatively, analyses of individual-level data involve the application of statistical methods that allow for the within cluster correlation.² This approach is exemplified by methods such as mixed effects ('multilevel') models and marginal models estimated using generalised estimating equations.

CRTs are increasingly used to evaluate interventions for improving health outcomes in children.^{12,13} Because of the amount of time children spend in school, it provides a natural setting in which interventions for preventing health problems can be delivered, participating children can be recruited, and health outcomes measured.¹³ At a policy level, there is increasing awareness of the potential for using the school setting to deliver, non-pharmacological, complex, prevention public health interventions.^{12,14,15} Cluster randomisation is a more natural approach than individual randomisation in the school-based setting. It is often difficult to randomise individuals as pupils belong to predefined clusters (eg, class, year group, school), and contamination between the trial arms can result as pupils interact within clusters. The CRT design respects the clustered structure in schools.

School-based CRTs share the same challenges of trials where other types of cluster are randomised. Within-cluster correlation is expected in school-based CRTs for a number of reasons: parents choose the schools their children attend and this may be related to factors associated with pupil outcomes; the school environment and culture will have a common influence on the pupils; pupils interact within schools and this can result in similar behaviours and outcomes.^{12,13,16} Other recognised challenges may be even more salient for trials that randomise schools. There are additional ethical considerations for enrolment of children into trials to ensure pupils remain protected as research subjects.¹² Consent needs to be sought from several key agents including parents, pupils, head teachers and teachers. Consent for the school to be allocated the intervention is usually provided by the head teacher, but there may be interventions delivered to entire classes that some parents do not want their children to receive (eg, aspects of sex education programmes that are not part of the standard curriculum). Retention

of recruited pupils is an issue for trials that have a long follow-up duration and there is the need to consider how best to handle the analysis of data from pupils that change schools (clusters) during the course of the study.¹⁷

Several books have been published regarding CRT methodology.^{1 2 18-20} In addition, there have been a number of reviews of the conduct and reporting quality of CRTs.^{12 21-26} One systematic review examined the characteristics and quality of reporting of CRTs worldwide involving children¹² and highlighted the specific difficulties of conducting such studies; nearly three-quarters of the included studies randomised schools as the clusters. That review and our initial scoping research suggests a sharp increase in the number of these studies. No systematic review has specifically focused on the characteristics of CRTs of health interventions delivered in the school-based setting in the UK. Such a review would: provide a pool of relevant knowledge for researchers planning future similar trials in the UK; highlight good practices and common methodological challenges; obtain useful trial-based data on the intraclass (intraschool) correlation coefficient; provide relevant parameter values for simulation-based studies that use synthetic data to assess the statistical properties of methods used to analyse data from school-based CRTs.

This review aims to summarise the characteristics of, and methodological practices in, school-based CRTs with pupil health outcomes in the UK. The review examines several areas, including: participant characteristics; intervention type; recruitment, sampling and allocation methods; consent and ethical approval procedures; retention and analysis methods. The main outcome is a description of the methodological characteristics of school-based CRTs in the UK with a health outcome. Knowledge of the study characteristics and practices of researchers will greatly aid the design of CRTs in the school health setting.

METHODS

The systematic review will describe the characteristics of CRTs with health outcomes in the school setting. This section contains a description of the methodological strategy based on guidelines from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement.²⁷

Search strategy

Peer reviewed articles written in English, published from inception to 30 June 2020 inclusively indexed in MEDLINE (Ovid) will be the source of data for this systematic review. The search strategy was developed following an initial scoping of the focus area and in consultation with an information specialist. The search strategy combined free text and index terms for the concepts study type (CRTs) and *schools* (box 1). The study type concept was developed based on a sensitive MEDLINE search strategy for identification of CRTs developed by Taljaard *et al.*²⁸ Cluster design-related terms, 'cluster*',

Box 1 Search strategy implemented in this systematic review

Search strategy

Terms for Randomised Controlled trials:

1. random:.mp.
2. trial.ab,kw,ti.

Cluster design-related terms:

3. "cluster*".ab, kw, ti.
4. "group*".ab, kw, ti.
5. "communit*".ab, kw, ti.

6. 3 OR 4 OR 5

School MeSH term:

7. exp Schools/

Highest precision:

8. 1 AND 2 AND 6 AND 7

9. 8 limited to English language

'group*' and 'communit*' were combined with the terms 'random' and 'trial', along with 'Schools'. The search was then limited to English language as our resources make it unfeasible to translate papers (online supplemental table S1).

Eligibility criteria

Eligible papers will be those reporting the results from school-based CRTs of health-related interventions in the UK for which there is a primary health outcome (including physical and mental health outcomes, health attitudes and well-being) measured on pupils.

The review will include participants who are children of school age in education in the UK. Participants are pupils in preschool, primary school and secondary school settings. The types of eligible clusters include schools themselves, year groups, classrooms, teachers or any other relevant school-related unit. Any health-related intervention(s) will be considered. There must be a control/usual care comparison group within the published article. The primary outcome must be health related and measured on pupils.

All types of CRT design are eligible, including parallel group, crossover and stepped wedge trials. Only definitive CRTs will be included. Only studies published in English will be included and we anticipate that all studies carried out in the UK will be published in English. If more than one publication of the primary outcome result for an eligible CRT is identified, a key study report (index paper) will be designated and used for data extraction.

Papers which do not report the main findings (primary outcome) will be excluded along with feasibility/pilot studies, protocol/design articles, process evaluations, economic evaluations/cost-effectiveness studies, statistical analysis plans, commentaries and papers reporting only findings from mediation/mechanism analyses. Studies for which the primary outcome is not health based (eg, education attainment) will be excluded.



Screening and selection

All potentially eligible studies will undergo a two-stage screening process.

Stage 1: The titles and abstracts of the studies will be retrieved from MEDLINE (Ovid) and downloaded into Endnote (X9).²⁹ Any duplicate citations will be removed and remaining citations will be dual screened (KP and OU) for eligibility against the inclusion criteria above. Disagreements will be resolved through discussion.

Stage 2: Full-text articles will be obtained for all papers that are potentially eligible following title and abstract screening. The reviewers (KP and OU) will evaluate articles based on the inclusion criteria using a pre-piloted coding method. Any discrepancies which cannot be resolved through discussion will be sent to a third reviewer (ZMX) for a decision.

Reviewers will keep a record of any studies excluded at each step. Results will be reported using a PRISMA flow diagram.

Risk of bias (quality) assessment

A risk of bias assessment is not necessary for this methodological review as we are not interested in the specific estimates of intervention effect in the included studies. We seek only to describe characteristics of the studies. Some of the information extracted from the papers is indicative of quality in CRTs and this will be summarised as part of our review.

Data extraction

A data extraction form will be piloted on a random sample of 10 included papers. Any modifications to the form will be made following the pilot. KP will extract data from all eligible papers, while OU will check extraction. Any uncertainty will be resolved by consulting a third reviewer (ZMX). Information will be recorded using a data extraction form in Microsoft Excel.

The following data will be extracted from included articles: characteristics of the participating schools and pupils; intervention type and mode of delivery; health condition/aspect targeted by the intervention; justification for using cluster trial design; unit of randomisation (ie, type of cluster); school-level (or other cluster-level) characteristics used to balance the randomisation; allocation ratio; length of follow-up; number of follow-ups; target sample size (ie, number of schools and pupils); assumptions underlying sample size (eg, ICC, anticipated loss to follow-up); committee that provided ethical approval; activities covered by the consent agreements; primary outcome; reporter of primary outcome (eg, teacher, parent, pupil); method of data collection; achieved sample size; number of schools (clusters) and pupils that were lost to follow-up; analysis method used to estimate intervention effect; baseline factors that were adjusted for in the analysis; value of the ICC in the primary analysis model; methodological challenges that were highlighted by the authors.

Missing information that is not available in the included papers will be obtained from corresponding protocol papers and other sibling publications for the studies. Authors may be contacted for missing or incomplete information and given 2 weeks to respond.

Data analysis

No formal sample size in terms of the number of required eligible papers has been calculated because we are seeking to obtain all school-based CRTs in the UK to date published in MEDLINE (Ovid). Meta-analysis is not appropriate as the review is focused on summarising methodological characteristics. Study characteristics will be summarised using means and standard deviations (or medians and IQRs) for continuous variables, and numbers and percentages for categorical variables. Appropriate graphs (eg, histograms, line graphs, scatterplots) will also be used to summarise specific features of the data. Challenges reported by authors will be summarised narratively.³⁰ Statistical analysis will be performed using Stata V.16.³¹

Patient and public involvement

There has been no contribution from patients or the public to the design of this systematic review protocol.

DISCUSSION

To our knowledge, this is the first systematic review to describe the characteristics and methodological practices of school-based CRTs of health interventions in the UK. We have a defined search strategy that is tailored to identifying school-based CRTs, selection criteria and a pre-piloted extraction strategy. Pilot testing, and subsequent screening and data extraction will be conducted by two independent reviewers, with disagreements resolved by consulting a third reviewer. In doing this we hope to minimise inaccuracy. Additionally, the review aims to cover a range of CRTs conducted in schools for a variety of different health conditions/areas.

Identifying CRTs is challenging as many papers do not explicitly use the word 'cluster' in the title or abstract. We have included terms in our search such as 'group' randomised and 'community' randomised to try and improve the sensitivity, thus widening the search so not to miss any eligible papers. We have also used the exploded Medical Subject Headings (MeSH) term, 'exp School/', in the hope of identifying publications that may state schools or classes as their unit of randomisation.

This review will summarise data using descriptive statistics. Meta-analysis is not used here to pool ICC estimates. Our initial scoping of the literature indicated that most papers do not report the SE of the ICC, which is required for pooling the estimates. Furthermore, the studies to be included in the review will be methodologically and clinically diverse (eg, different outcomes and health conditions). There is, therefore, no true single underlying ICC; rather there is a range of true ICCs specific to different

scenarios. A single pooled ICC from a meta-analysis would not be meaningful and would obscure nuances about how its size depends on the context of the study. It is more useful to summarise the variability in the estimated ICCs as this provides a range of values within which to assess the sensitivity of the sample size calculation to uncertainty about the true value of the ICC in a given scenario.³²

We have conducted extensive scoping searches in order to best identify the studies of interest. A limitation of the review is that we will limit our search to the MEDLINE (Ovid) database, thus, potentially missing out on articles published in other journals (eg, mental health interventions published in PsycINFO). MEDLINE was used because our research question is to describe the characteristics of trials that evaluate the impact of health interventions on health outcomes. The database includes journals of interest for both physical health and mental health. We have also chosen not to examine grey literature therefore potentially missing out on studies with greater methodological challenges. Also there will be no forward and backward citation searching, but we do have a clearly defined population of papers. Feasibility studies have been excluded, but there are different learning issues from such studies that will be the subject of a separate review. These decisions have been taken to enable the review to be more focused and time-effective.

The review is focused on health-based CRTs in schools. There is a wider literature of other types of intervention (particularly in educational research) that have been evaluated in this setting using the CRT design, but, given the limited resources and the large number of potentially eligible studies identified during scoping, it was considered more relevant and efficient to restrict the review to studies in the health area.³³

Another limitation of this review is the difficulty in identifying CRTs as many papers do not explicitly use the word 'cluster' in the title or abstract. Therefore, we have included terms in our search such as 'group' and 'community' to try and improve the sensitivity, thus widening the search so not to miss any eligible papers. We have also used the exploded MeSH term, 'exp School/', in the hope of identifying publications that may state schools or classes as their unit of randomisation.

A pragmatic decision has been made to focus on UK studies in order to align with available resources and create a relevant and focused review. There is the possibility that we are missing out on the opportunity to learn from studies in countries that have a similar education system to the UK. Our scoping searches established that there is a considerably large number of eligible papers and we restricted the study eligibility to the UK. Ideally, we would include papers globally, but this is not practical. Despite being focused on the UK, the findings of this review will be of wider interest as many methodological challenges in the design of CRTs will be similar across some countries.

Because of the amount of time children spend in school, it provides a natural setting in which interventions

for preventing health problems and improving health outcomes in children can be delivered and evaluated.¹³ Cluster randomised controlled trials in the school-based setting are particularly relevant for non-pharmacological interventions, such as social programmes aimed at improving public health¹³ and the use of this study design is increasing.¹² Through summarising the methodological aspects of health-related cluster randomised controlled trials conducted in a schools, this review will provide methodology-related knowledge specific to these trials which will help researchers plan future similar studies effectively in the UK and elsewhere.

ETHICS AND DISSEMINATION

This methodological systematic review does not require ethical approval as only secondary data extracted from papers will be analysed and the data are not linked to individual participants. After completion of the systematic review, the data will be analysed, and the findings disseminated through peer-reviewed publications and scientific meetings.

Acknowledgements KP was supported by a PhD studentship funded by the National Institute for Health Research Applied Research Collaboration South West Peninsula (NIHR ARC South West Peninsula). OU was supported by the NIHR ARC South West Peninsula.

Contributors KP and OU conceptualised the study. KP drafted the manuscript and incorporated comments from authors for successive drafts. MPN, ZMX and TF contributed to the design and content. All authors read and approved the final manuscript. KP is the guarantor of the review.

Funding This work was supported by the National Institute for Health Research (NIHR) Applied Research Collaboration South West Peninsula (NIHR ARC South West Peninsula). Our funded group National Institute for Health Research Applied Research Collaboration South West Peninsula does not have a grant number because we are part of the NIHR infrastructure.

Disclaimer The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>.

ORCID iDs

Kitty Parker <http://orcid.org/0000-0003-2319-8227>

Obioha C Ukoumunne <http://orcid.org/0000-0002-0551-9157>

REFERENCES

- 1 Eldridge S, Kerry S. *A practical guide to cluster randomised trials in health services research*. Chichester, West Sussex: John Wiley & Sons, 2012.
- 2 Donner A, Klar N. *Design and analysis of cluster randomization trials in health research*. Wiley, 2000.
- 3 Campbell MJ. Cluster randomized trials in general (family) practice research. *Stat Methods Med Res* 2000;9:81–94.
- 4 Eldridge SM, Ukoumunne OC, Carlin JB. The intra-cluster correlation coefficient in cluster randomized trials: a review of definitions. *Int Stat Rev* 2009;77:378–94.
- 5 Kerry SM, Bland JM. The intraclass correlation coefficient in cluster randomisation. *BMJ* 1998;316:1455–60.
- 6 Eldridge SM, Ukoumunne OC, Carlin JB. The Intra-Cluster correlation coefficient in cluster randomized trials: a review of definitions. *Int Stat Rev* 2009;77:378–94.
- 7 Ukoumunne OC. A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. *Stat Med* 2002;21:3757–74.
- 8 Eldridge S, Kerry S. *A practical guide to cluster randomised trials in health services research*. Wiley, 2012.
- 9 Bland JM, Kerry SM, notes S. Trials randomised in clusters. *BMJ* 1997;315:600.
- 10 Kahan BC, Forbes G, Ali Y, et al. Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. *Trials* 2016;17:438.
- 11 Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol* 1978;108:100–2.
- 12 Walleser S, Hill SR, Bero LA. Characteristics and quality of reporting of cluster randomized trials in children: reporting needs improvement. *J Clin Epidemiol* 2011;64:1331–40.
- 13 Goesling B. A practical guide to cluster randomized trials in school health research. *J Sch Health* 2019;89:916–25.
- 14 Crocetti MT, Amin DD, Scherer R. Assessment of risk of bias among pediatric randomized controlled trials. *Pediatrics* 2010;126:298–305.
- 15 Thomson D, Hartling L, Cohen E, et al. Controlled trials in children: quantity, methodological quality and descriptive characteristics of pediatric controlled trials published 1948–2006. *PLoS One* 2010;5:e13106.
- 16 Shackleton N, Hale D, Bonell C, et al. Intraclass correlation values for adolescent health outcomes in secondary schools in 21 European countries. *SSM Popul Health* 2016;2:217–25.
- 17 Schweig JD, Pane JF, McCaffrey DF. Switching cluster membership in cluster randomized control trials: implications for design and analysis. *Psychol Methods* 2020;25:516–34.
- 18 Campbell M, Walters S, Design Hto. *Analyse and report cluster randomised trials in medicine and health related research*. Hoboken, NJ: John Wiley & Sons, 2014.
- 19 Hayes R, Moulton L. *Cluster randomised trials*. Boca Raton, FL: Chapman and Hall/CRC Press, 2009.
- 20 Murray D. *Design and analysis of Group-Randomized trials*. New York: Oxford University Press, 1998.
- 21 Eldridge SM, Ashby D, Feder GS, et al. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clin Trials* 2004;1:80–90.
- 22 Fiero MH, Huang S, Oren E, et al. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials* 2016;17:72.
- 23 Ivers NM, Taljaard M, Dixon S, et al. Impact of consort extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000–8. *BMJ* 2011;343:d5886.
- 24 Donner A, Brown KS, Brasher P. A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979–1989. *Int J Epidemiol* 1990;19:795–800.
- 25 Simpson JM, Klar N, Donnor A. Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. *Am J Public Health* 1995;85:1378–83.
- 26 Varnell SP, Murray DM, Janega JB, et al. Design and analysis of group-randomized trials: a review of recent practices. *Am J Public Health* 2004;94:393–9.
- 27 Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339:b2535.
- 28 Taljaard M, McGowan J, Grimshaw JM, et al. Electronic search strategies to identify reports of cluster randomized trials in MEDLINE: low precision will improve with adherence to reporting standards. *BMC Med Res Methodol* 2010;10:15.
- 29 EndNote [program]. *Endnote X9 version*. Philadelphia, PA: Clarivate, 2013.
- 30 Popay J, Roberts H, Sowden A. Guidance on the conduct of narrative synthesis in systematic reviews: a product from the ESRC methods programme 2006.
- 31 Stata16 [program]. *Release 16 version*. College Station, TX: StataCorp LLC, 2019.
- 32 Turner RM, Thompson SG, Spiegelhalter DJ. Prior distributions for the intraclass correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clin Trials* 2005;2:108–18.
- 33 Torgerson D. *Designing randomised trials in health education and the social sciences: an introduction*. Springer, 2008.

Appendix 2 – Characteristics and practices of school-based cluster randomised controlled trials for improving health outcomes on pupils in the United Kingdom: a methodological systematic review

Parker et al. *BMC Med Res Methodol* (2021) 21:152
<https://doi.org/10.1186/s12874-021-01348-0>

BMC Medical Research
Methodology

RESEARCH

Open Access

Characteristics and practices of school-based cluster randomised controlled trials for improving health outcomes in pupils in the United Kingdom: a methodological systematic review



Kitty Parker^{1*}, Michael Nunns², ZhiMin Xiao³, Tamsin Ford⁴ and Obioha C. Ukoumunne¹

Abstract

Background: Cluster randomised trials (CRTs) are increasingly used to evaluate non-pharmacological interventions for improving child health. Although methodological challenges of CRTs are well documented, the characteristics of school-based CRTs with pupil health outcomes have not been systematically described. Our objective was to describe methodological characteristics of these studies in the United Kingdom (UK).

Methods: MEDLINE was systematically searched from inception to 30th June 2020. Included studies used the CRT design in schools and measured primary outcomes on pupils. Study characteristics were described using descriptive statistics.

Results: Of 3138 articles identified, 64 were included. CRTs with pupil health outcomes have been increasingly used in the UK school setting since the earliest included paper was published in 1993; 37 (58%) studies were published after 2010. Of the 44 studies that reported information, 93% included state-funded schools. Thirty six (56%) were exclusively in primary schools and 24 (38%) exclusively in secondary schools. Schools were randomised in 56 studies, classrooms in 6 studies, and year groups in 2 studies. Eighty percent of studies used restricted randomisation to balance cluster-level characteristics between trial arms, but few provided justification for their choice of balancing factors. Interventions covered 11 different health areas; 53 (83%) included components that were necessarily administered to entire clusters. The median (interquartile range) number of clusters and pupils recruited was 31.5 (21 to 50) and 1308 (604 to 3201), respectively. In half the studies, at least one cluster dropped out. Only 26 (41%) studies reported the intra-cluster correlation coefficient (ICC) of the primary outcome from the analysis; this was often markedly different to the assumed ICC in the sample size calculation. The median (range) ICC for school clusters was 0.028 (0.0005 to 0.21).

Conclusions: The increasing pool of school-based CRTs examining pupil health outcomes provides methodological knowledge and highlights design challenges. Data from these studies should be used to identify the best school-level

*Correspondence: kp477@exeter.ac.uk

¹ NIHR Applied Research Collaboration South West Peninsula (PenARC), University of Exeter, Room 2.16, South Cloisters, St Luke's Campus, 79 Heavitree Rd, Exeter EX1 2LU, UK

Full list of author information is available at the end of the article



© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

characteristics for balancing the randomisation. Better information on the ICC of pupil health outcomes is required to aid the planning of future CRTs. Improved reporting of the recruitment process will help to identify barriers to obtaining representative samples of schools.

Keywords: Child and adolescent health, Cluster randomised trials, Public health, Randomised trials, Research methods, Schools, Systematic review

Background

Cluster randomised trials (CRTs) are studies in which groups, or clusters, of individuals are allocated to trial arms rather than the individuals themselves [1]. The clusters may be geographic areas, health organisations or social units. CRTs are used when the intervention is delivered to the entire cluster or there is a chance of contamination between trial arms if individuals are randomised [2].

CRTs can be more complex to design and analyse than individually randomised controlled trials. The most documented methodological consideration for CRTs is that observations on participants from the same cluster are more likely to be similar to each other than those on participants from different clusters [2]. This similarity is quantified by the intra-cluster correlation coefficient (ICC), defined as the proportion of the total variability in the trial outcome that is between clusters as opposed to between individuals within clusters [3]. The statistical dependence between observations within clusters needs to be taken account of when calculating the sample size and analysing data in CRTs [1]. The use of standard methods may result in the sample size being too small to detect the intervention effect, and analysis results that exaggerate the evidence for a true intervention effect. Estimates of the ICC or coefficient of variation of clusters for the outcome from previous studies are required to calculate the design effect, the factor by which the number of individuals that would be required in an individually randomised trial needs to be inflated to account for within-cluster correlation in the sample size calculation. In addition, when calculating the sample size in CRTs, a degrees of freedom correction should be incorporated to take account of the uncertainty with which variability in the outcome across clusters is estimated in the analysis [4], and a further inflation of the sample size should be considered to allow for loss of efficiency that results from recruiting unequal numbers of participants from the clusters [5]. When estimating the intervention effect from the resulting trial data the main analytical approaches are to either apply standard statistical methods to summary statistics that represent the cluster response (cluster-level analyses) or use methods at the individual participant level that account for within-cluster correlation in the model or by weighting the analysis. Another important

methodological consideration in CRTs is the potential for recruitment bias that might occur in studies where the participating individuals are recruited after the clusters are randomised. Finally, when using meta-analysis to pool findings from studies that use the CRT design, there is the need to consider how best to incorporate estimated effects from studies that did not allow for clustering in the analysis, and consider the extent to which differences in the types of clusters that were randomised are a source of heterogeneity. These considerations are detailed in several textbooks [1, 2, 6–8].

CRTs are increasingly used to evaluate non-pharmacological interventions for improving child health outcomes [9–11]. Although the use of CRTs to evaluate the effectiveness of interventions for improving educational outcomes is long established [12, 13], their use to evaluate health interventions in schools is more recent [10]. Schools provide a natural environment to recruit, deliver public health interventions to and measure outcomes on children, due to the amount of time they spend there [10]. Cluster randomisation is consistent with the natural clustering found within school settings (i.e., classrooms within year groups within schools). School-based CRTs share common challenges with other settings, but specific considerations may be more challenging when schools are randomised, for example, consent procedures [10, 14].

In 2011, a methodological systematic review on the characteristics and quality of reporting of CRTs involving children reported a marked increase in such studies [9]; three quarters of the included studies randomised schools. To date, no systematic review has focussed specifically on the characteristics of school-based CRTs for improving pupil health outcomes. Such a review would help identify common methodological challenges, obtain estimates of parameters (e.g., the ICC) that are of use to researchers planning similar trials and inform the design of simulation studies that use synthetic data to evaluate the properties of statistical methods applied in the context of school-based CRTs with health outcomes.

The aim of this methodological systematic review is to describe the characteristics and practices of school-based CRTs for improving health outcomes in pupils in the United Kingdom (UK).

Methods

This is a systematic review of school-based CRTs with pupil health outcomes that were conducted in the UK. The review was focussed on the UK to align with constraints on available resources and collect richer data on CRT methodology in a single education system.

Data sources and search methods

The systematic review was registered with PROSPERO (CRD42020201792) and the protocol has been published [15]. After extensive scoping of the subject area, a pragmatic decision was made to search MEDLINE (through Ovid) in order to make the review more time-efficient and align with available resources. MEDLINE was exclusively searched from inception to 30th June 2020 for peer-reviewed articles of school-based CRTs. The search strategy (Table 1) was developed in consultation with information specialists, based on a sensitive MEDLINE search strategy for identifying CRTs [16]. Cluster design-related terms ‘cluster*’, ‘group*’ and ‘communit*’ were combined with the terms ‘random’ and ‘trial’, along with the ‘Schools’ Medical Subject Heading (MeSH) term. The search was limited to English language.

Inclusion and exclusion criteria

The systematic review included school-based definitive CRTs of the effectiveness of an intervention versus a comparison group that evaluated health outcomes on pupils. The population of interest was children in full-time education in the UK. Studies that took place outside the UK were excluded. The pragmatic decision was made to limit the population to educational settings within the UK as it made the review more focussed and applicable to a specific setting. Eligible studies included pupils in pre-school, primary school and secondary school. The types

of eligible clusters included schools themselves, year groups, classes, teachers or any other relevant school-related unit. All school types were eligible, including special schools. Any health-related intervention(s) and control groups were considered. The primary outcome had to be related to pupils’ health. Studies for which the primary outcome was not health-based (e.g., academic attainment) were excluded. All types of CRT design were eligible including parallel group, factorial, crossover and stepped wedge studies.

If more than one publication of the primary outcome result for an eligible CRT was identified, a key study (index) report was designated and used for data extraction. Papers that did not report the primary outcome were excluded along with pilot/feasibility studies, protocol/design articles, process evaluations, economic evaluations/cost-effectiveness studies, statistical analysis plans, commentaries and mediation/mechanism analyses.

Sifting and validation

Two reviewers (KP and OU) independently screened the titles and abstracts of all references (downloaded into Endnote [17]) for eligibility against the inclusion criteria. Any studies for which the reviewers were uncertain of for inclusion were taken to full text screening. Full-text articles were evaluated by the same reviewers based on the inclusion criteria using a pre-piloted coding method. Any discrepancies which could not be resolved through discussion were sent to a third reviewer (ZMX) for a decision.

Data extraction and analysis

For each eligible study, data were extracted using a pre-piloted form in Microsoft Excel. Data were extracted by two reviewers (KP and OU), and any discrepancies that could not be resolved through discussion were sent to a third reviewer (ZMX) for a final decision. Missing information that was not available in the index papers was sought from corresponding protocol papers and other “sibling” publications.

The items of information extracted are listed as follows:

Publication details: year of publication and journal name.

Setting characteristics: country/region, school level and type of school.

Intervention: health area and intervention type.

Primary outcome: name, health area, reporter of outcome and method of data collection.

Study design and analysis methods: unit of randomisation (i.e., type of cluster), justification for using the cluster trial design, method used to sample schools, method used to balance the randomisation, length

Table 1 Systematic review search strategy

Search strategy
Terms for randomised controlled trials:
1. random.mp.
2. trialab, kw, ti.
Cluster design-related terms:
3. "cluster*"ab, kw, ti.
4. "group*"ab, kw, ti.
5. "communit*"ab, kw, ti.
6. 3 OR 4 OR 5
School MeSH term:
7. exp Schools/
Highest precision:
8. 1 AND 2 AND 6 AND 7
9. 8 limited to English language

and number of follow-ups, design of follow-up (cohort versus repeated cross-sectional design) and method used to account for clustering in the analysis.

Sample size calculation: target sample size (i.e., number of clusters and pupils) and assumptions underlying the sample size calculation (e.g., assumed ICC, percentage loss to follow-up).

Ethics and consent procedures: activities covered by the consent agreements and use of “opt-out” consent.

Other study characteristics of methodological interest: number of clusters and pupils that were recruited and lost to follow-up, estimate of the ICC of the primary outcome.

Study characteristics were described using medians, interquartile ranges (IQRs) and ranges for continuous variables, and numbers and percentages for categorical variables, using Stata software [18]. Formal quality assessment was not performed as it was not an objective of this review to estimate intervention effects in the included studies. Some information relevant to the quality of CRTs was, however, extracted and summarised as part of the review.

Results

Search results

After deduplication, 3103 articles were identified through MEDLINE, 159 were full-text screened and 64 were included in the review [19–82]. Of 95 excluded studies, 88 did not meet the inclusion criteria, and 7 studies met inclusion criteria but were subsequently excluded because they were sibling reports of an index paper. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram is in Fig. 1.

Study characteristics

The included papers were published in 36 different journals, including: *British Medical Journal* (n=9 papers); *BMC Public Health* (n=4); *International Journal of Behavioural Nutrition and Physical Activity* (n=4); *Archives of Disease in Childhood* (n=3); *BMJ Open* (n=3); *Journal of Epidemiology and Community Health* (n=3); *Public Health Nutrition* (n=3); and *The Lancet* (n=3). The CRT design has been increasingly used in the UK school setting to evaluate health interventions for pupils since the first paper was published in 1993 (Fig. 2). Twenty three papers were published between 2001 and 2010, compared to 37 between January 2011 and June 2020.

Table 2 summarises the characteristics of included studies.

Setting

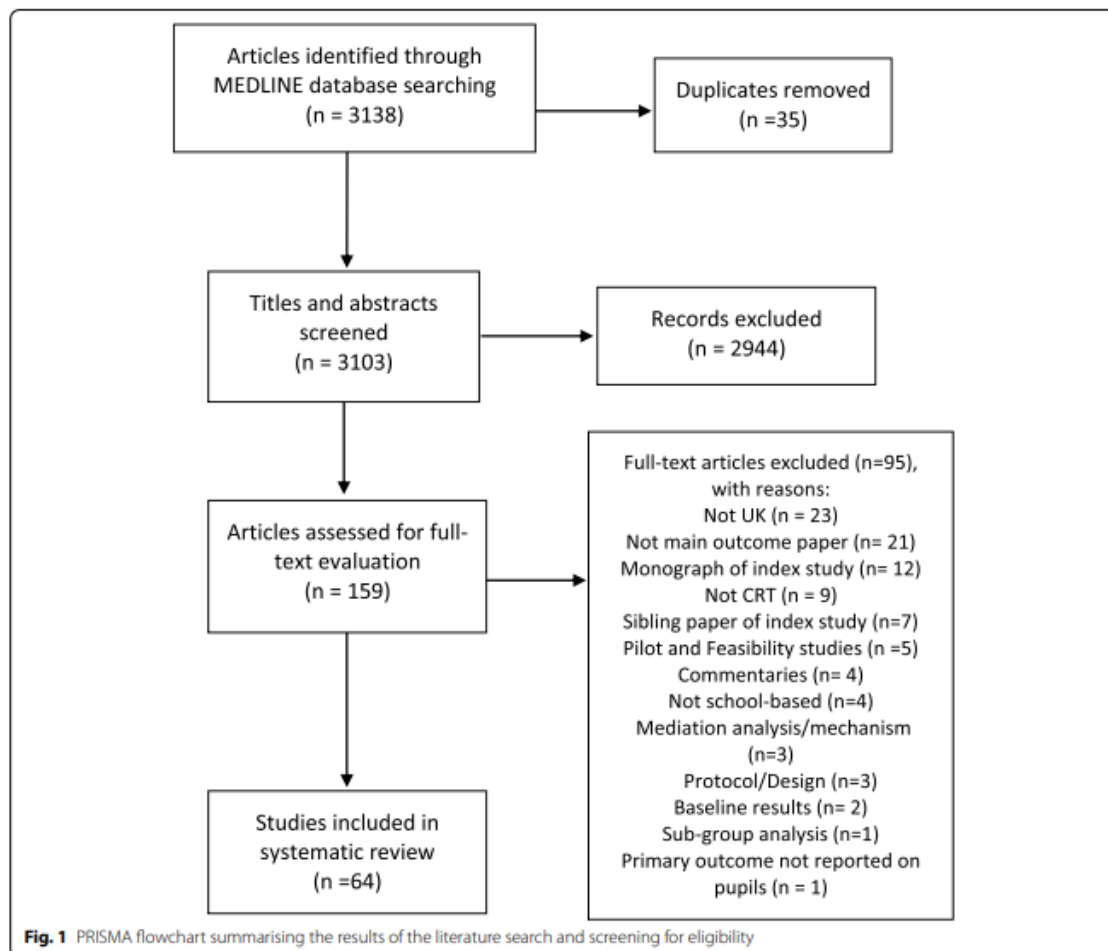
Almost three quarters of the studies were conducted exclusively in England (n=47; 73%); most studies (50 of the 52 studies that provided the data) took place in one or two geographic regions (e.g., West Midlands). Just over half the studies (56%) were based exclusively in primary schools (age 5–11 years), and 38% were exclusively in secondary schools (age 11–16 years). Of the 44 studies that reported information on the types [83] of schools recruited, 93% included state-funded schools.

Intervention type

Eighteen (28%) studies evaluated interventions that targeted nutrition, 15 (23%) physical activity, 15 (23%) socioemotional function and its influences, 7 (11%) dental health, 5 (8%) smoking and 5 (8%) injury, amongst others. Physical health interventions are increasingly prominent (13 published since 2011 in contrast to just 2 prior to then). Of the 15 studies targeting socioemotional function and its influences, 13 were published since 2011, highlighting increasing use of the CRT design in this area. Of the 7 CRTs related to dental health, the most recent one was published in 2011. The vast majority of interventions were in primary prevention (94%).

In 53 (83%) studies, the intervention had at least one component that necessarily had to be administered to entire clusters (“cluster–cluster” interventions [1]). Such components often included educational lessons (e.g., classroom-based lessons [23], physical activity [43] and gardening [25]). Other less common components included breakfast clubs [46, 73], funding/resources [37], change in school policy [50] and advertisements [40]. Eleven (17%) studies had intervention components that directly targeted individual pupils (“individual-cluster” interventions [1]), such as the use of fluoride varnish [72]. Thirty three (52%) studies had “professional-cluster” interventions [1]: in 30 (47%) studies the teacher was either trained in or provided with guidance to deliver components of the intervention, in 3 studies pupils were trained to deliver peer-led intervention components [21, 26, 42], and in 1 study the school nurse was trained [66]. Half the studies (n=32) had “external-cluster” interventions [1] where people external to the school delivered intervention components (e.g., researchers [23], trained facilitators [53], dental professionals [51], dance instructors [41] and student volunteers [47]).

Two studies [53, 78] had 2 control groups (one “usual care” and one active) and 16 (25%) used a delayed intervention (waitlist) design.



Primary outcome

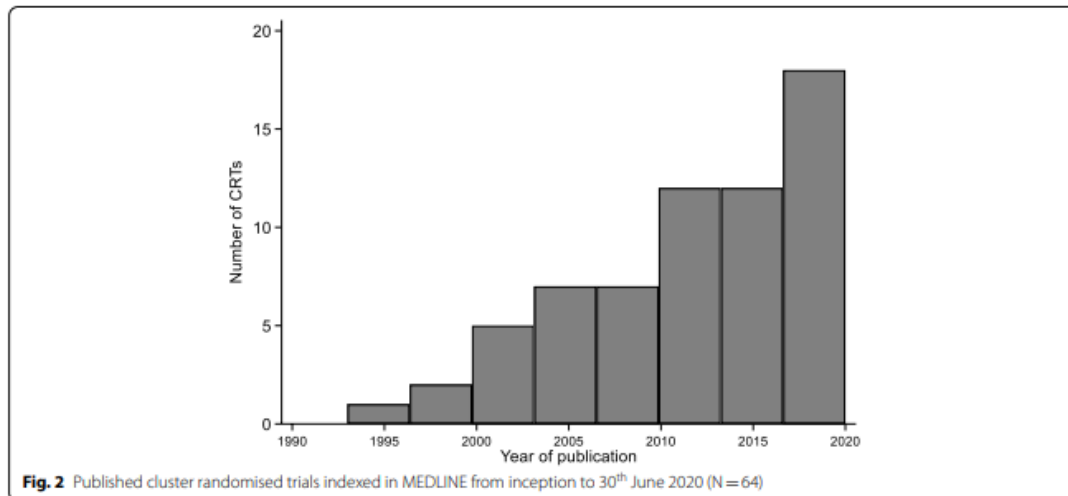
Health areas assessed by the primary outcomes are summarised in Table 2. In 53% of the studies pupils reported the primary outcome, with researchers reporting primary outcomes in 20%, teachers in 8%, and parents in 8%. In 28% of the studies the primary outcome reporter was blind to allocation status (some authors specifically commented on the challenges of blinding trial arm status [33, 36, 56, 60]), and 22% measured the outcome using an objective method.

Study design and analysis methods

Explicit justification for use of the CRT design was only provided in 17 (27%) studies; the most common reason was to avoid contamination (13 studies altogether). Most studies (n=56; 88%) randomised school clusters, while classes and year groups were allocated in 6 (9%)

and 2 (3%) studies, respectively. Two authors said that in order to maintain power, classes were randomised instead of schools and that this may have led to contamination between the intervention and control arms [22, 28]. Nearly all studies used a parallel group design (n=61; 95%); the remaining 3 used a factorial design [21, 37, 39]. Of the 46 studies with sufficient information to establish the approach used to sample schools, 33 initially invited all potentially eligible schools to participate, 5 used random sampling, 4 used purposive sampling, 3 used convenience sampling, and 1 used a mixed random/convenience sampling approach.

Eighty percent of studies reported using a restricted allocation method to balance cluster-level characteristics between the trial arms. Most commonly a measure of socio-economic status (SES) was balanced on (48%), with a third of studies (21/64) specifically balancing



the allocation on the percentage of pupils eligible for free school meals. Other commonly-used balancing factors are described in Table 3. Few studies gave justification for their choice of balancing factors.

One of the challenges of CRTs is to avoid recruitment bias that might occur if participants are recruited after the clusters are randomised [88, 89]. One third (33%) of studies avoided this by recruiting pupils before the clusters were randomised; furthermore, 25% collected baseline data before randomisation. This information, however, was unclear in many studies (41% and 33%, respectively). Generally, insufficient information was provided on whether recruitment bias was avoided in studies where pupils were recruited after randomisation of clusters. A notable exception was one study [57] where recruitment bias was avoided because allocation was not revealed to the schools until after recruitment and baseline assessment.

Nearly all studies used the cohort design as their method of follow-up ($n = 62$, 97%), where the same pupils provided data at each study wave. One study used a repeated cross-sectional design where different pupils provided data at each wave [46], and one used an a priori mixed design incorporating elements of the cohort and repeated cross-sectional designs, with only a subset of participating pupils providing data at each wave [49].

Seventy two percent of studies analysed their data using individual-level methods that allow for clustering, 16% used cluster-level analysis methods, and 12% did not allow for clustering in their analysis.

Sample size calculation

Seventy eight percent of studies accounted for clustering in their sample size calculation and 72% reported the ICC or coefficient of variation [90] that was assumed for the outcome. None of the studies made a degrees of freedom correction to the sample size calculation. Only two studies [57, 63] allowed for unequal cluster sizes in their sample size calculation, and only one of these [57] specified the anticipated variation in the number of pupils across clusters. The median (range) assumed ICC for school clusters was 0.05 (0.005 to 0.175) based on the 37 studies that provided these data. Of the 3 studies that specified the coefficient of variation of the outcome, 2 assumed it to be 0.2 [42, 60] and 1 assumed it to be 0.25 [19]. The median (range) assumed design effect was 2.21 (1.22 to 8.11). The median targeted sample size was 30 and 964 clusters and pupils, respectively. Most studies (94%) did not state whether their sample size calculation allowed for loss to follow-up of clusters.

Ethics and consent procedures

Information regarding consent procedures was not well reported and consent for the participation of the cluster was often implied rather than explicitly detailed. In 63% of studies it was stated that both parents/guardians and pupils provided consent or assent for study participation. Forty five percent of studies reported that opt-out consent [14] from either the parent/guardian and/or the pupil was used for participation.

Table 2 Characteristics of included studies (N = 64)

Characteristic	N	Statistics
Setting		
Country	64	
England, n (%)		47 (73)
Scotland, n (%)		5 (8)
Wales, n (%)		3 (5)
Northern Ireland, n (%)		3 (5)
More than one country ^a , n (%)		6 (9)
Number of regions from which schools were drawn ^b	64	
One		40 (62)
Two		10 (16)
Three		1 (2)
Four		1 (2)
Unclear		12 (19)
School level	64	
Preschool only, n (%)		2 (3)
Primary only, n (%)		36 (56)
Secondary only, n (%)		24 (38)
Primary and Secondary, n (%)		2 (3)
School types that were included [83] ^c	44	
State, n (%)		41 (93)
Independent, n (%)		6 (14)
Academies, n (%)		2 (5)
Grammar, n (%)		2 (5)
Special, n (%)		2 (5)
Voluntary aided, n (%)		2 (5)
Foundation, n (%)		1 (2)
Faith, n (%)		1 (2)
Intervention		
Health area of intervention ^d	64	
Nutrition, n (%)		18 (28)
Physical activity, n (%)		15 (23)
Socioemotional function and its influences ^e , n (%)		15 (23)
Dental health, n (%)		7 (11)
Smoking, n (%)		5 (8)
Injury, n (%)		5 (8)
Sexual health, n (%)		3 (5)
Alcohol misuse, n (%)		2 (3)
Cancer, n (%)		1 (2)
Communication skills (for children with autism), n (%)		1 (2)
Health attitudes (breast feeding), n (%)		1 (2)
Level of prevention	64	
Primary prevention, n (%)		60 (94)
Secondary prevention, n (%)		4 (6)
Type of intervention [1] ^f	64	
Individual-cluster, n (%)		11 (17)
Professional-cluster, n (%)		33 (52)
External-cluster, n (%)		32 (50)
Cluster-cluster, n (%)		53 (83)
Multifaceted, n (%)		53 (83)

Table 2 (continued)

Characteristic	N	Statistics
Primary outcome		
Primary outcome health area	64	
Socioemotional function and its influences ⁹ , n (%)		15 (23)
Nutrition, n (%)		10 (16)
Dental health, n (%)		7 (11)
Physical activity, n (%)		7 (11)
Obesity, n (%)		7 (11)
Smoking, n (%)		5 (8)
Injury, n (%)		3 (5)
Sexual health, n (%)		2 (3)
Obstetrics, n (%)		2 (3)
Alcohol misuse, n (%)		2 (3)
Cancer, n (%)		1 (2)
Communication skills (for children with autism), n (%)		1 (2)
Gross motor skills, n (%)		1 (2)
Safety, n (%)		1 (2)
Main reporter of primary outcome	64	
Pupil, n (%)		34 (53)
Researcher, n (%)		12 (19)
Dentist, n (%)		6 (9)
Teacher, n (%)		5 (8)
Parent, n (%)		4 (6)
Routine data, n (%)		2 (3)
Researcher and parent, n (%)		1 (2)
Primary outcome reporter blind to allocation status	64	
Yes, n (%)		18 (28)
No, n (%)		46 (72)
Primary outcome measurement was objective	64	
Yes, n (%)		14 (22)
No, n (%)		50 (78)
Study design and analysis methods		
Justification provided for randomising clusters	64	
Yes, n (%)		17 (27)
No, n (%)		47 (73)
Reason for randomising clusters	17	
To avoid contamination, n (%)		9 (53)
Intervention was delivered at the cluster level, n (%)		4 (24)
To avoid contamination and for logistical reasons, n (%)		2 (12)
To avoid contamination and avoid "selection bias", n (%)		1 (6)
To avoid contamination and because intervention was delivered at the cluster level, n (%)		1 (6)
Unit of randomisation	64	
Schools, n (%)		56 (88)
Classes, n (%)		6 (9)
Year groups, n (%)		2 (3)
Number of trial arms	64	
Two, n (%)		55 (86)
Three, n (%)		5 (8)
Four, n (%)		4 (6)
Study design	64	

Table 2 (continued)

Characteristic	N	Statistics
Parallel group, n (%)		61 (95)
Factorial, n (%)		3 (5)
Method used to sample schools	64	
All potentially eligible schools invited, n (%)		33 (52)
Random sample, n (%)		5 (8)
Purposive sample, n (%)		4 (6)
Convenience sample, n (%)		3 (5)
Mixed random/convenience sample, n (%)		1 (2)
Unclear, n (%)		18 (28)
Type of randomisation	64	
Completely randomised, n (%)		13 (20)
Stratified, n (%)		29 (45)
Matched, n (%)		8 (13)
Minimisation, n (%)		8 (13)
Constrained [84, 85], n (%)		6 (9)
Type of follow-up	64	
Cohort, n (%)		62 (97)
Repeated cross-sectional, n (%)		1 (2)
Mixed, n (%)		1 (2)
Number of follow-ups	64	
1, n (%)		32 (50)
2, n (%)		21 (33)
3, n (%)		6 (9)
4, n (%)		5 (8)
Length of follow-up	64	
Up to 6 months, n (%)		22 (34)
7 to 12 months, n (%)		19 (30)
13 to 18 months, n (%)		6 (9)
19 to 24 months, n (%)		8 (13)
25 to 36 months, n (%)		7 (11)
More than 36 months, n (%)		2 (3)
Participants recruited before clusters were randomised	64	
Yes, n (%)		21 (33)
No, n (%)		17 (27)
Unclear, n (%)		26 (41)
Baseline data collected before clusters were randomised	64	
Yes, n (%)		16 (25)
No, n (%)		27 (42)
Unclear, n (%)		21 (33)
Method of analysis	64	
Individual-level analysis that allows for clustering, n (%)		46 (72)
Cluster-level analysis, n (%)		10 (16)
Did not allow for clustering, n (%)		8 (12)
Sample size calculation		
Assumed school-level intra-cluster correlation coefficient of outcome, median (IQR; range)	37	0.05 (0.02 to 0.1; 0.005 to 0.175)
Assumed design effect, median (IQR; range)	36	2.21 (1.98 to 3.53; 1.22 to 8.11)
Study allowed for drop-out at cluster level	64	
Yes, n (%)		4 (6)

Table 2 (continued)

Characteristic	N	Statistics
Not stated, n (%)		60 (94)
Study allowed for drop-out at individual level ^h	62	
Yes, n (%)		18 (29)
Not stated, n (%)		44 (71)
Target number of clusters, median (IQR; range)	46	30 (20 to 40; 4 to 160)
Target number of schools, median (IQR; range)	41	30 (20 to 42; 4 to 160)
Target number of individuals, median (IQR; range) ^j	45	964 (498 to 2000; 90 to 9000)
Ethics and consent procedures		
From whom was consent/assent sought for pupil participation?	64	
Parents and pupils, n (%)		40 (63)
Parents only, n (%)		15 (23)
Pupils only, n (%)		2 (3)
Not stated / Neither parent nor pupil, n (%)		7 (11)
Opt-out consent/assent procedure used for either parent/guardian or pupils	64	
Yes, n (%)		29 (45)
Not stated / No, n (%)		35 (55)
Other study characteristics of methodological interest		
Ethnicity: percentage of pupils that are White, median (IQR; range)	33	76.8 (51.5 to 86.2; 24 to 95.3)
Total number of clusters recruited, median (IQR; range)	62	31.5 (21 to 50; 4 to 486)
Total number of schools recruited, median (IQR; range)	63	29 (15 to 50; 4 to 486)
Total number of pupils recruited, median (IQR; range) ^l	60	1308 (604 to 3201; 17 to 27,435)
Percentage of clusters followed-up for primary outcome, median (IQR; range)	62	100 (92.5 to 100; 60.5 to 100)
Percentage of pupils followed-up for primary outcome, median (IQR; range) ^k	58	79.9 (64.1 to 87.5; 7.7 to 100)
Observed school-level intra-cluster correlation coefficient of primary outcome, median (IQR; range)	23	0.028 (0.017 to 0.12; 0.0005 to 0.21)

^a Studies that included schools from more than one country in the United Kingdom

^b English regions included: South West, South East (including Greater London), East of England, West Midlands, East Midlands, North West, North East, Yorkshire and The Humber, "Southern England", "Central England" and "West of England". Scottish regions included: Glasgow, Inverclyde, Tayside, Grampian, Lanarkshire, Lothian and Fife. Welsh regions included: North Wales, South West Wales and South East Wales. Northern Irish regions included: South Belfast, East Belfast, Ulster, Leinster, Connacht and Munster

^c Some studies included more than one school type. This is the number of studies that included specific types of school. State schools receive funding through their local authority or directly from the government. The most common ones are local authority, foundation and voluntary aided school which are all funded by the local authority. Academies are run by government and not-for-profit trusts, and are independent of local authority. Grammar schools are run by local authorities but intake is based on assessment of the pupils' academic ability. Special schools cater for pupils with special educational needs. Faith schools follow the national curriculum but can decide what they teach in religious studies. Independent schools follow the national curriculum but charge fees for attending pupils

^d Some interventions targeted more than one health area

^e Includes mental health, behaviour, ADHD, wellbeing, quality of life, bullying, social and emotional learning, and self-esteem

^f Intervention type was summarised based on the typology described by Eldridge and colleagues [1]. 'Individual-cluster' interventions include components that are directed at individual participants (e.g. pupils) on whom outcomes are measured. 'Professional-cluster' interventions include components for training professionals in the cluster (e.g. teachers in schools) to deliver the intervention. 'External-cluster' interventions involve additional staff outside the cluster to deliver the intervention (e.g. researchers, trained facilitators). 'Cluster-cluster' interventions include components that necessarily have to be administered to entire clusters (e.g., school policy). 'Multifaceted' interventions include components across more than one of the 'individual-cluster', 'professional-cluster', 'external-cluster' and 'cluster-cluster' categories

^g Includes mental health, behaviour, hyperactivity/inattention (ADHD), wellbeing, quality of life, bullying, social and emotional learning, and self-esteem (body image)

^h Summary excludes the two CRTs that did not use the cohort design

ⁱ Summary excludes the two CRTs that did not use the cohort design

^j Summary excludes the two CRTs that did not use the cohort design

^k Summary excludes the two CRTs that did not use the cohort design

Table 3 Cluster-level characteristics used to balance the randomisation (N = 64)

Characteristic	Statistic
Deprivation (school or area in which school is based)	
Yes – Percentage of pupils eligible for free school meals, n (%)	21 (33)
Yes – Townsend Index [86] ^a , n (%)	2 (3)
Yes – Income Deprivation Affecting Children Index (IDACI) [87] ^b , n (%)	1 (2)
Yes – Index of Multiple Deprivation (IMD) [87] ^c , n (%)	1 (2)
Yes – Unspecified ^d , n (%)	6 (9)
Cluster size	
Yes, n (%)	23 (36)
Geographic area of school	
Yes, n (%)	13 (20)
Pupil ethnicity summary	
Yes, n (%)	5 (8)
Co-educational status of school	
Yes, n (%)	5 (8)
School performance	
Yes, n (%)	5 (8)
School type	
Yes, n (%)	2 (3)
Other ^e	
Yes, n (%)	24 (38)

^a Townsend Index quantifies material deprivation within a population

^b Income Deprivation Affecting Children Index (IDACI) is the proportion of all children aged 0 to 15 living in income deprived families in different local areas across England

^c Index of Multiple Deprivation (IMD) measures relative deprivation for small areas (or neighbourhoods) in England

^d Did not state which measure of deprivation used

^e Other balancing factors include: Percentage of students who actively commuted to school; School; English-speaking versus Welsh-speaking school; Local sexual health services; Number of students in year group; Date of entry of school into study; School in urban versus rural area; Percentage of children speaking English as an additional language; Quality and quantity of current school sex education; Local authority; Percentage of pupils staying on after age 16 years; Special educational need status; Whether school has existing policy similar to the intervention; School expressed preference for allocation (control versus intervention versus no preference); Health-promoting school status; Percentage of children in year group of interest with no dental decay; Frequency and timetabling of personal, social, and health education lessons; Preferred timetabling of the intervention; Facilitator of the intervention (Regional Project Manager)

Other study characteristics of methodological interest

A median (IQR) of 31.5 (21 to 50) clusters, 29 (15 to 50) schools and 1308 (604 to 3201) pupils were recruited. The CRT studies that used a cohort design and reported both targeted and achieved recruitment figures at the cluster (n = 45) and pupil (n = 43) levels achieved those recruitment targets in 89% and 77% of studies, respectively. Some authors noted challenges with recruitment at the cluster [45, 47, 50] and pupil [24, 55] levels. Based on the 33 studies that provided data, the median (IQR) percentage of pupils categorised as “White” was 76.8% (51.5% to 86.2%). Thirty out of 62 (48%) studies that provided information reported that at least one cluster was lost to follow-up. Missing data resulting from entire school drop-out was highlighted as a problem in some reports (e.g., [42, 48, 54]). The median follow-up at the pupil level was 79.9%.

Only 26 (41%) studies overall, and 18 of the 37 (49%) studies published after 2010, reported the ICC from the analysis of the primary outcome; the specific ICC

values are reported in Table 4. The median (range) ICC for school clusters was 0.028 (0.0005 to 0.21). For many studies that reported both values there was a marked difference between the observed school-level ICC in the study data and the corresponding assumed value of the ICC in the sample size calculation (Fig. 3). The median (range) of the differences between the observed ICC and the assumed ICC was -0.006 (-0.117 to 0.16) indicating that: on average, the observed ICC was slightly smaller than the assumed ICC; at one extreme, the observed ICC in one study was 0.117 smaller than the assumed value [25]; and at the other extreme, the observed ICC in one study was 0.16 larger than the assumed value [68]. The intra-class correlation coefficient of agreement between the observed and assumed ICCs was 0.24.

Seven studies [24, 26, 44, 59, 68, 71, 74] that reported ICCs had a binary primary outcome, but none of these stated whether the ICC was calculated on the proportions scale or the logistic scale [3]. It is possible that five of these studies [24, 26, 68, 71, 74] that used

Table 4 Reported intra-cluster correlation coefficients for primary outcomes (N = 26)

Author	Year	Cluster unit	Outcome	Health area	Outcome type	ICC estimate
Stallard [53]	2012	year group	Symptoms of low mood (depression)	socioemotional function	continuous	0.012
Chisholm [22]	2016	class	Stigma of mental illness	socioemotional function	continuous	0.1
Obsuth [71]	2017	school	School exclusion	socioemotional function	binary	0.028
Connolly [82]	2018	school	Prosocial behaviour	socioemotional function	continuous	0.116
Ford [32]	2019	school	Mental health / behaviour	socioemotional function	continuous	0.121
Axford [44]	2020	school	Victimisation (being bullied) occurring at least twice a month in the last 2 months	socioemotional function	binary	0.019
Campbell [26]	2008	school	Smoking in the past week	smoking	binary	0.017
Conner [74]	2019	school	Ever smoking	smoking	binary	0.017
McKay [24]	2018	school	Heavy episodic drinking in the previous 30 days (> = 6 units for males and > = 4.5 units for females)	alcohol misuse	binary	0.121
Crocker [40]	2012	school	Child's eating habits	obesity	continuous	0.07
Fairclough [56]	2013	school	Waist circumference (cm)	obesity	continuous	0.06
Lloyd [57]	2018	school	BMI z score	obesity	continuous	0.014
Breheeny [43]	2020	school	BMI z-score at 12 months	obesity	continuous	0.001
Jago [41]	2015	school	Mean weekday minutes of moderate to vigorous physical activity per day	physical activity	continuous	0.0005
Harrington [58]	2018	school	Minutes per day of moderate- to vigorous physical activity	physical activity	continuous	0.02
Norris [67]	2018	school	Sedentary behaviour during the school day in minutes	physical activity	continuous	0.080
James ^a [28]	2004	class	Consumption of carbonated drinks over 3 days (in glasses)	nutrition	continuous	-0.009
Christian [25]	2014	school	Combined daily fruit and vegetable intake (grams per day)	nutrition	continuous	0.003
Redmond [81]	1999	school	Proportion of teeth sites with caries at 6 months	dental health	continuous	0.16
Worthington [31]	2001	school	Plaque score	dental health	continuous	0.023
Milsom [59]	2006	school	Whether the child has active caries in their first permanent molars	dental health	binary	0.027
Mulvaney [68]	2006	school	Use of visibility aid (reflective and fluorescent slap wrap) while cycling	injury	binary	0.21
Kendrick [79]	2007	school	Knowledge score for fire and burn prevention	safety	continuous	0.187
Hubbard [76]	2016	school	Number of recognised cancer warning signs	cancer	continuous	0.038
Henderson [27]	2007	school	Terminations of pregnancy by age 20	obstetrics	count	0.005
Giles [23]	2014	school	Intention to breastfeed	obstetrics	continuous	0.12

^a The estimated intra-cluster correlation coefficient in James (2004) was negative. True negative values are generally considered implausible in the context of cluster randomised trials

mixed effects ("multi-level") models [91] to analyse the data reported the ICC on the logistic scale, which could potentially account for some of the differences between the observed and assumed ICCs. Further scrutiny of the data, however, revealed marked differences for only two of the aforementioned studies: 0.21 for the observed ICC versus 0.05 for the assumed ICC in Mulvaney and colleagues [68], and 0.028 versus 0.1, respectively, in Obsuth and colleagues [71].

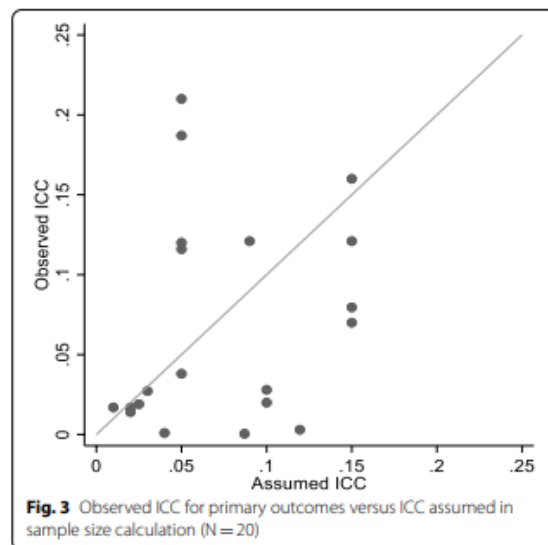
Discussion

The number of UK school-based CRTs evaluating the effects of interventions on pupil health outcomes has increased in recent years, reflecting growing

recognition of the role that schools can play in improving the health of children [10, 92–95]. The findings of this systematic review indicate a number of methodological considerations that are worthy of reflection.

Interpretation

Seventy two percent of the studies reported the level of clustering assumed in their sample size calculation, a little more than the 62% observed in a 2015 review of the reporting of sample size calculations in CRTs [96]. Our review found that the observed ICC in the study data often differed markedly from the ICC assumed in the sample size calculation. This will be partly due to sampling variation and adjustment for prognostic factors in



the analysis, but it may also reflect the lack of availability of good estimates of the ICC at the time of sample size calculation. Knowledge of the ICC for pupil health outcomes in the school setting is less well established than for patient health outcomes in the primary care setting where general practices are allocated as clusters [1, 97]. It has been reported that general practice-level ICCs for health outcomes are generally less than 0.05 [98]; in our review, only 13 of 23 studies that randomised school clusters and reported observed ICCs had values that were less than 0.05. School-based ICC estimates are widely available for educational outcomes [99], but these are markedly higher than those reported in this review for pupil health outcomes; this is to be expected given that the primary role of the school is to provide education. The importance of reporting ICCs from study data for planning future similar CRTs has long been established [100] and the 2012 CONSORT extension to CRTs includes a specific reporting item for this [101]. Only two-fifths (41%) of studies in this review, however, reported the ICC for the primary outcome; this figure rises to 48% (16/33) for studies published after 2012. Improved reporting of the ICC in the increasing number of CRTs in the school-based setting, and further papers written specifically to report ICCs [102, 103], will provide valuable knowledge. This review focussed on CRTs in the UK setting; a useful area to investigate is the extent to which school-based ICC estimates for health outcomes from other countries (e.g., [102, 104]) are similar to those in the UK.

Representativeness of school and pupil characteristics in school-based trials is important for external validity

and inclusiveness. For most studies in this review, schools were recruited from only one or two geographic regions/counties. A median 23% of participating pupils were in a minority ethnic group, lower than the national percentages reported by the UK Department for Education (33.5% of primary school pupils and 31.3% of secondary school pupils) [105]. The study reports generally provided little information on specific aspects of the recruitment process, such as why some schools declined to participate and details of their characteristics. Many of the studies evaluated interventions that involved classroom lessons and necessitated teachers being trained to deliver the intervention. Additionally, the teachers reported pupil outcomes in some studies [32, 34, 60, 73, 82]. Insufficient school resources to deliver the intervention and the wider trial may be a barrier to participation and result in lack of representation of certain types of schools.

Eighty percent of the studies used some form of restricted allocation to balance the randomisation on cluster-level characteristics, which is higher than previous methodological reviews of CRTs [106–109]. The percentage of pupils in the school that are eligible for free school meals was often used as a balancing factor, perhaps partly because this information is readily available from the UK Department for Education [110]. School characteristics that are predictive of the study outcomes, account for within-cluster correlation or influence effectiveness of the intervention are candidates on which to balance the randomisation [1, 111]; previous school-based CRTs could be used to identify such factors.

Strengths

This systematic review used a defined search strategy tailored to identify school-based CRTs. The strategy was developed following an iterative process and allowed us to achieve the right balance of sensitivity and specificity relevant to our available resources. Identifying reports of CRTs is a challenge given that many articles do not use the term 'cluster' in their title or abstract. Therefore, a search strategy was used which included terms such as 'group' and 'community' to improve sensitivity. The 'School' MeSH term was also used to identify publications that randomised any type of school-related unit. The piloting of our screening procedure and data extraction were conducted by two independent reviewers, improving accuracy. The review identified school-based CRTs with interventions spanning a variety of different health conditions/areas.

Limitations

A potential limitation of the review is that the search was limited to one database. MEDLINE was used because the

focus of the review was on describing the characteristics of trials that evaluate the impact of health interventions on pupil's health outcomes, but it is possible that we have not identified eligible publications that are not indexed in MEDLINE. Translating our search in the EMBASE, DARE, PsycINFO and ERIC databases for potential includes published in the last 3 years, however, revealed only one additional eligible school-based CRT.

Given resource constraints, we focussed the review on the UK, making the decision to collect rich data on CRT methodology in a single education system. As a result, the findings are readily applicable to a specific context. Despite being focussed on the UK, the findings of this review will be of global interest. Other high income countries, such as Australia, have a similar school system to the UK, and many of our findings may be applicable in those settings. Furthermore, some of the methodological challenges in the design of CRTs will be similar across different settings.

Future directions

The results provide a summary of the methodological characteristics of school-based CRTs with pupil health outcomes in the UK. To our knowledge, there has been no systematic review of the characteristics of school-based CRTs for evaluating interventions for improving education outcomes, despite the fact that the use of the CRT design is more established in that area. A comparison of methodology between health-based CRTs and education-based CRTs in the school setting would be valuable to both areas. The results in our review indicate that better information on the ICC is needed to design school-based CRTs with health outcomes. Cataloguing of ICCs from previous studies will help researchers choose better values for the assumed ICC when calculating sample size.

Conclusions

CRTs are increasingly used in the school setting for evaluating interventions for improving children's health and wellbeing. The emerging pool of published trials in the UK provides investigators and methodologists with relevant experiential knowledge for the design of future similar studies. This review of school-based CRTs has highlighted the need for more information on the ICCs to calculate the required sample size. Better reporting of the recruitment process in CRTs will help to identify common barriers to obtaining representative samples of schools and pupils. Finally, previous school-based CRTs may provide a useful source of data to identify the school-level characteristics that are strong predictors of pupil health

outcomes and, therefore, potentially good factors on which to balance the randomisation.

Abbreviations

CRT: Cluster randomised trial; DARE: Database of Abstracts of Reviews of Effects; EMBASE: Excerpta Medica Database; ERIC: Education Resources Information Center; ICC: Intra-cluster correlation coefficient; IQR: Interquartile range; MeSH: Medical Subject Headings; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PsycINFO: Psychological Information Database; SES: Socio-economic status; UK: United Kingdom.

Acknowledgements

Kitty Parker and Obioha Ukoumunne were supported by the National Institute for Health Research Applied Research Collaboration South West Peninsula. The views expressed in this publication are those of the author(s) and not necessarily those of the National Institute for Health Research or the Department of Health and Social Care.

Authors' contributions

KP, MN, ZMX, TF and OU conceived the study. ZMX and TF advised on the design of the study and contributed to the protocol. KP, MN and OU contributed to the design of the study, wrote the protocol and designed the data extraction form. KP and OU undertook data extraction. KP conducted the analyses of the data. All authors had full access to all the data. KP took primary responsibility for writing the manuscript. All authors provided feedback on all versions of the paper. All authors read and approved the final manuscript.

Funding

This research was funded by the National Institute for Health Research Applied Research Collaboration South West Peninsula.

Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available because they are also being used for a wider ongoing programme of research but are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

No competing interests to declare.

Author details

¹NIHR Applied Research Collaboration South West Peninsula (PenARC), University of Exeter, Room 2.16, South Cloisters, St Luke's Campus, 79 Heavitree Rd, Exeter EX1 2LU, UK. ²College of Medicine and Health, University of Exeter, St Luke's Campus, Heavitree Road, Exeter EX1 2LU, UK. ³School of Health and Social Care, University of Essex, Colchester CO4 3SQ, UK. ⁴Department of Psychiatry, University of Cambridge, LS Clifford Allbutt Building, Cambridge Biomedical Campus Box 58, Cambridge CB2 0AH, UK.

Received: 18 May 2021 Accepted: 14 July 2021

Published online: 26 July 2021

References

1. Eldridge SM, Kerry S. A Practical Guide to Cluster Randomised Trials in Health Services Research. Chichester: John Wiley & Sons; 2012.
2. Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. Chichester: Wiley; 2000.

3. Eldridge SM, Ukoumunne OC, Carlin JB. The intra-cluster correlation coefficient in cluster randomized trials: A review of definitions. *Int Stat Rev.* 2009;77(3):378–94.
4. Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol.* 1978;108(2):100–2.
5. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol.* 2006;35(5):1292–300.
6. Campbell MJ, Walters S. *How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research.* Chichester: John Wiley and Sons; 2014.
7. Hayes R, Moulton L. *Cluster Randomised Trials.* Florida: CRC Press; 2009.
8. Murray DM. *Design and Analysis of Group-Randomized Trials.* New York: Oxford University Press; 1998.
9. Walliser S, Hill SR, Bero LA. Characteristics and quality of reporting of cluster randomized trials in children: reporting needs improvement. *J Clin Epidemiol.* 2011;64(12):1331–40.
10. Goesling B. A practical guide to cluster randomized trials in school health research. *J School Health.* 2019;89(11):916–25.
11. Thomson D, Hartling L, Cohen E, Vandermeer B, Tjosvold L, Klassen TP. Controlled trials in children: quantity, methodological quality and descriptive characteristics of pediatric controlled trials published 1948–2006. *PLoS One.* 2010;5(9):e13106.
12. Torgerson D, Torgerson C. *Designing Randomised Trials in Health, Education and the Social Sciences: An Introduction.* London: Palgrave Macmillan UK; 2008.
13. Spybrook J, Zhang Q, Kelcey B, Dong N. Learning from cluster randomized trials in education: An assessment of the capacity of studies to determine what works, for whom, and under what conditions. *Educ Eval Policy Anal.* 2020;42(3):354–74.
14. Felzmann H. Ethical issues in school-based research. *Res Ethics Rev.* 2009;5(3):104–9.
15. Parker K, Nunns MP, Xiao Z, Ford T, Ukoumunne OC. Characteristics and practices of school-based cluster randomised controlled trials for improving health outcomes in pupils in the UK: a systematic review protocol. *BMJ Open.* 2021;11(2):e044143.
16. Taljaard M, McGowan J, Grimshaw J, Brehaut J, McRae A, Eccles M, et al. Electronic search strategies to identify reports of cluster randomized trials in MEDLINE: Low precision will improve with adherence to reporting standards. *BMC Med Res Methodol.* 2010;10:15.
17. The EndNote Team. *EndNote. EndNote X9 version ed.* Philadelphia, PA: Clarivate; 2013.
18. StataCorp. *Stata Statistical Software: Release 16 ed.* College Station, TX: StataCorp LLC; 2019.
19. Lakshman RR, Sharp SJ, Ong KK, Forouhi NG. A novel school-based intervention to improve nutrition knowledge in children: cluster randomised controlled trial. *BMC Public Health.* 2010;10:123.
20. Marcano-Olivier M, Pearson R, Ruparell A, Horne PJ, Viktor S, Erjavec M. A low-cost Behavioural Nudge and choice architecture intervention targeting school lunches increases children's consumption of fruit: a cluster randomised trial. *Int J Behav Nutr Phys Act.* 2019;16(1):1–9.
21. Tymms PB, Curtis SE, Routen AC, Thomson KH, Bolden DS, Bock S, et al. Clustered randomised controlled trial of two education interventions designed to increase physical activity and well-being of secondary school students: the MOVE Project. *BMJ Open.* 2016;6(1):e009318.
22. Chisholm K, Patterson P, Torgerson C, Turner E, Jenkinson D, Birchwood M. Impact of contact on adolescents' mental health literacy and stigma: the SchoolSpace cluster randomised controlled trial. *BMJ Open.* 2016;6(2):e009435.
23. Giles M, McClenahan C, Armour C, Millar S, Rae G, Mallett J, et al. Evaluation of a theory of planned behaviour-based breastfeeding intervention in Northern Irish Schools using a randomized cluster design. *Br J Health Psychol.* 2014;19(1):16–35.
24. McKay M, Agus A, Cole J, Doherty P, Foxcroft D, Harvey S, et al. Steps Towards Alcohol Misuse Prevention Programme (STAMPP): a school-based and community-based cluster randomised controlled trial. *BMJ Open.* 2018;8(3):e019722.
25. Christian MS, Evans CE, Nykjaer C, Hancock N, Cade JE. Evaluation of the impact of a school gardening intervention on children's fruit and vegetable intake: a randomised controlled trial. *Int J Behav Nutr Phys Act.* 2014;11(1):1–15.
26. Campbell R, Starkey F, Holliday J, Audrey S, Bloor M, Parry-Langdon N, et al. An informal school-based peer-led intervention for smoking prevention in adolescence (ASSIST): a cluster randomised trial. *Lancet.* 2008;371(9624):1595–602.
27. Henderson M, Wight D, Raab G, Abraham C, Parkes A, Scott S, et al. Impact of a theoretically based sex education programme (SHARE) delivered by teachers on NHS registered conceptions and terminations: final results of cluster randomised trial. *BMJ.* 2007;334(7585):133.
28. James J, Thomas P, Cavan D, Kerr D. Preventing childhood obesity by reducing consumption of carbonated drinks: cluster randomised controlled trial. *BMJ.* 2004;328(7450):1237.
29. Sahota P, Rudolf MC, Dixey R, Hill AJ, Barth JH, Cade J. Randomised controlled trial of primary school based intervention to reduce risk factors for obesity. *BMJ.* 2001;323(7320):1029.
30. Howlin P, Gordon RK, Pasco G, Wade A, Charman T. The effectiveness of Picture Exchange Communication System (PECS) training for teachers of children with autism: a pragmatic, group randomised controlled trial. *J Child Psychol Psychiatry.* 2007;48(5):473–81.
31. Worthington HV, Hill KB, Mooney J, Hamilton FA, Blinkhorn AS. A cluster randomized controlled trial of a dental health education program for 10-year-old children. *J Public Health Dentistry.* 2001;61(1):22–7.
32. Ford T, Hayes R, Byford S, Edwards V, Fletcher M, Logan S, et al. The effectiveness and cost-effectiveness of the Incredible Years® Teacher Classroom Management programme in primary school children: results of the STARS cluster randomised controlled trial. *Psychol Med.* 2019;49(5):828–42.
33. Sayal K, Taylor JA, Valentine A, Guo B, Sampson CJ, Sellman E, et al. Effectiveness and cost-effectiveness of a brief school-based group programme for parents of children at risk of ADHD: a cluster randomised controlled trial. *Child Care Health Dev.* 2016;42(4):521–33.
34. Humphrey N, Barlow A, Wigelsworth M, Lendrum A, Pert K, Joyce C, et al. A cluster randomized controlled trial of the Promoting Alternative Thinking Strategies (PATHS) curriculum. *J School Psychol.* 2016;58:73–89.
35. Conrod PJ, O'Leary-Barrett M, Newton N, Topper L, Castellanos-Ryan N, Mackie C, et al. Effectiveness of a selective, personality-targeted prevention program for adolescent alcohol use and misuse: a cluster randomized controlled trial. *JAMA Psychiat.* 2013;70(3):334–42.
36. Hodgkinson A, Abbott J, Hurley MA, Lowe N, Qualter P. An educational intervention to prevent overweight in pre-school years: a cluster randomised trial with a focus on disadvantaged families. *BMC Public Health.* 2019;19(1):1–13.
37. Sharpe H, Patalay P, Vostanis P, Belsky J, Humphrey N, Wolpert M. Use, acceptability and impact of booklets designed to support mental health self-management and help seeking in schools: results of a large randomised controlled trial in England. *Eur Child Adolesc Psychiatry.* 2017;26(3):315–24.
38. Pine C, McGoldrick P, Burnside G, Curnow M, Chesters R, Nicholson J, et al. An intervention programme to establish regular toothbrushing: understanding parents' beliefs and motivating children. *Int Dental J.* 2000;50(6):312–23.
39. Nutbeam D, Macaskill P, Smith C, Simpson JM, Catford J. Evaluation of two school smoking education programmes under normal classroom conditions. *BMJ.* 1993;306(6870):102–7.
40. Croker H, Lucas R, Wardle J. Cluster-randomised trial to evaluate the 'Change for Life' mass media/social marketing campaign in the UK. *BMC Public Health.* 2012;12(1):1–14.
41. Jago R, Edwards MJ, Sebire SJ, Tomkinson K, Bird EL, Banfield K, et al. Effect and cost of an after-school dance programme on the physical activity of 11–12 year old girls: The Bristol Girls Dance Project, a school-based cluster randomised controlled trial. *Int J Behav Nutr Phys Act.* 2015;12(1):1–15.
42. Stephenson J, Strange V, Forrest S, Oakley A, Copas A, Allen E, et al. Pupil-led sex education in England (RIPPLE study): cluster-randomised intervention trial. *The Lancet.* 2004;364(9431):338–46.
43. Breheny K, Passmore S, Adab P, Martin J, Hemming K, Lancashire ER, et al. Effectiveness and cost-effectiveness of The Daily Mile on childhood weight outcomes and wellbeing: a cluster randomised controlled trial. *Int J Obesity.* 2020;44(4):812–22.
44. Axford N, Bjornstad G, Clarkson S, Ukoumunne OC, Wrigley Z, Matthews J, et al. The effectiveness of the KiVa bullying prevention

- program in Wales, UK: Results from a pragmatic cluster randomized controlled trial. *Prev Sci*. 2020;21(5):615–26.
45. Diedrichs PC, Atkinson MJ, Steer RJ, Garbett KM, Rumsey N, Halliwell E. Effectiveness of a brief school-based body image intervention 'Dove Confident Me: Single Session' when delivered by teachers and researchers: Results from a cluster randomised controlled trial. *Behav Res Ther*. 2015;74:94–104.
 46. Murphy S, Moore G, Tapper K, Lynch R, Clarke R, Raisanen L, et al. Free healthy breakfasts in primary schools: a cluster randomised controlled trial of a policy intervention in Wales. *UK Public Health Nutr*. 2011;14(2):219–26.
 47. Breslin G, Shannon S, Rafferty R, Fitzpatrick B, Belton S, O'Brien W, et al. The effect of sport for LIFE: all island in children from low socio-economic status: a clustered randomized controlled trial. *Health Qual Life Outcomes*. 2019;17(1):1–12.
 48. Foulkes J, Knowles Z, Fairclough S, Stratton G, O'Dwyer M, Ridgers N, et al. Effect of a 6-week active play intervention on fundamental movement skill competence of preschool children: a cluster randomized controlled trial. *Perceptual Motor Skills*. 2017;124(2):393–412.
 49. Moore L, Tapper K. The impact of school fruit tuck shops and school food policies on children's fruit consumption: a cluster randomised trial of schools in deprived areas. *J Epidemiol Commun Health*. 2008;62(10):926–31.
 50. Rowland D, DiGiuseppi C, Gross M, Afolabi E, Roberts I. Randomised controlled trial of site specific advice on school travel patterns. *Arch Dis Childhood*. 2003;88(1):8–11.
 51. Milsom K, Blinkhorn A, Walsh T, Worthington H, Kearney-Mitchell P, Whitehead H, et al. A cluster-randomized controlled trial: fluoride varnish in school children. *J Dental Res*. 2011;90(11):1306–11.
 52. Evans C, Greenwood DC, Thomas JD, Cleghorn CL, Kitchen MS, Cade JE. SMART lunch box intervention to improve the food and nutrient content of children's packed lunches: UK wide cluster randomised controlled trial. *J Epidemiol Commun Health*. 2010;64(11):970–6.
 53. Stallard P, Sayal K, Phillips R, Taylor JA, Spears M, Anderson R, et al. Classroom based cognitive behavioural therapy in reducing symptoms of depression in high risk adolescents: pragmatic cluster randomised controlled trial. *BMJ*. 2012;345:e6058.
 54. Scott S, O'Connor TG, Futh A, Matias C, Price J, Doolan M. Impact of a parenting program in a high-risk, multi-ethnic community: The PALS trial. *J Child Psychol Psychiatry*. 2010;51(12):1331–41.
 55. Markham WA, Bridle C, Grimshaw G, Stanton A, Aveyard P. Trial protocol and preliminary results for a cluster randomised trial of behavioural support versus brief advice for smoking cessation in adolescents. *BMC Res Notes*. 2010;3(1):1–10.
 56. Fairclough SJ, Hackett AF, Davies IG, Gobbi R, Mackintosh KA, Warburton GL, et al. Promoting healthy weight in primary school children through physical activity and nutrition education: a pragmatic evaluation of the CHANGE! randomised intervention study. *BMC Public Health*. 2013;13(1):1–14.
 57. Lloyd J, Creanor S, Logan S, Green C, Dean SG, Hillsdon M, et al. Effectiveness of the Healthy Lifestyles Programme (HeLP) to prevent obesity in UK primary-school children: a cluster randomised controlled trial. *Lancet Child Adolescent Health*. 2018;2(1):35–45.
 58. Harrington DM, Davies MJ, Bodicoat DH, Charles JM, Chudasama YV, Gorely T, et al. Effectiveness of the 'Girls Active' school-based physical activity programme: A cluster randomised controlled trial. *Int J Behav Nutr Phys Act*. 2018;15(1):1–18.
 59. Milsom K, Blinkhorn A, Worthington H, Threlfall A, Buchanan K, Kearney-Mitchell P, et al. The effectiveness of school dental screening: a cluster-randomized control trial. *J Dental Res*. 2006;85(10):924–8.
 60. Hislop MD, Stokes KA, Williams S, McKay CD, England ME, Kemp SP, et al. Reducing musculoskeletal injury and concussion risk in school-boy rugby players with a pre-activity movement control exercise programme: a cluster randomised controlled trial. *Br J Sports Med*. 2017;51(15):1140–6.
 61. Kendrick D, Royal S. Cycle helmet ownership and use; a cluster randomised controlled trial in primary school children in deprived areas. *Arch Dis Child*. 2004;89(4):330–5.
 62. Bonell C, Allen E, Warren E, McGowan J, Bevilaqua L, Jamal F, et al. Effects of the Learning Together intervention on bullying and aggression in English secondary schools (INCLUSIVE): a cluster randomised controlled trial. *Lancet*. 2018;392(10163):2452–64.
 63. Adab P, Pallan MJ, Lancashire ER, Hemming K, Frew E, Barrett T, et al. Effectiveness of a childhood obesity prevention programme delivered through schools, targeting 6 and 7 year olds: cluster randomised controlled trial (WAVES study). *BMJ*. 2018;360:k211.
 64. Kipping RR, Howe LD, Jago R, Campbell R, Wells S, Chittleborough CR, et al. Effect of intervention aimed at increasing physical activity, reducing sedentary behaviour, and increasing fruit and vegetable consumption in children: active for Life Year 5 (AFLYS) school based cluster randomised controlled trial. *BMJ*. 2014;348:g3256.
 65. Aveyard P, Cheng K, Almond J, Sherratt E, Lancashire R, Lawrence T, et al. Cluster randomised controlled trial of expert system based on the transtheoretical ("stages of change") model for smoking prevention and cessation in schools. *BMJ*. 1999;319(7215):948–53.
 66. Patterson E, Brennan M, Linskey K, Webb D, Shields M, Patterson C. A cluster randomised intervention trial of asthma clubs to improve quality of life in primary school children: the School Care and Asthma Management Project (SCAMP). *Arch Dis Childhood*. 2005;90(8):786–91.
 67. Norris E, Dunsmuir S, Duke-Williams O, Stamatakis E, Shelton N. Physically active lessons improve lesson activity and on-task behavior: A cluster-randomized controlled trial of the "Virtual Traveller" intervention. *Health Educ Behav*. 2018;45(6):945–56.
 68. Mulvaney CA, Kendrick D, Watson MC, Coupland CA. Increasing child pedestrian and cyclist visibility: cluster randomised controlled trial. *J Epidemiol Commun Health*. 2006;60(4):311–5.
 69. Rees G, Bakhshi S, Surujal-Harry A, Stasinopoulos M, Baker A. A computerised tailored intervention for increasing intakes of fruit, vegetables, brown bread and wholegrain cereals in adolescent girls. *Public Health Nutr*. 2010;13(8):1271–8.
 70. Graham A, Moore L, Sharp D, Diamond I. Improving teenagers' knowledge of emergency contraception: cluster randomised controlled trial of a teacher led intervention. *BMJ*. 2002;324(7347):1179.
 71. Obsuth I, Sutherland A, Cope A, Pilbeam L, Murray AL, Eisner M. London Education and Inclusion Project (LEIP): Results from a cluster-randomized controlled trial of an intervention to reduce school exclusion and antisocial behavior. *J Youth Adolescence*. 2017;46(3):538–57.
 72. Hardman M, Davies G, Duxbury J, Davies R. A cluster randomised controlled trial to evaluate the effectiveness of fluoride varnish as a public health measure to reduce caries in children. *Caries Res*. 2007;41(5):371–6.
 73. Shemilt I, Harvey I, Shepstone L, Swift L, Reading R, Mugford M, et al. A national evaluation of school breakfast clubs: evidence from a cluster randomized controlled trial and an observational analysis. *Child Care Health Dev*. 2004;30(5):413–27.
 74. Conner M, Grogan S, West R, Simms-Ellis R, Scholtens K, Sykes-Muskett B, et al. Effectiveness and cost-effectiveness of repeated implementation intention formation on adolescent smoking initiation: A cluster randomized controlled trial. *J Consulting Clin Psychol*. 2019;87(5):422.
 75. Griffin TL, Jackson DM, McNeill G, Aucott LS, MacDiarmid JI. A brief educational intervention increases knowledge of the sugar content of foods and drinks but does not decrease intakes in Scottish children aged 10–12 years. *J Nutr Educ Behav*. 2015;47(4):367–73.
 76. Hubbard G, Stoddart I, Forbat L, Neal RD, O'Carroll RE, Haw S, et al. School-based brief psycho-educational intervention to raise adolescent cancer awareness and address barriers to medical help-seeking about cancer: a cluster randomised controlled trial. *Psycho-Oncol*. 2016;25(7):760–71.
 77. Cunningham CJ, Elton R, Topping GV. A randomised control trial of the effectiveness of personalised letters sent subsequent to school dental inspections in increasing registration in unregistered children. *BMC Oral Health*. 2009;9(1):1–8.
 78. Stallard P, Skryabina E, Taylor G, Phillips R, Daniels H, Anderson R, et al. Classroom-based cognitive behaviour therapy (FRIENDS): a cluster randomised controlled trial to Prevent Anxiety in Children through Education in Schools (PACES). *Lancet Psychiatry*. 2014;1(3):185–92.
 79. Kendrick D, Groom L, Stewart J, Watson M, Mulvaney C, Casterton R. "Risk Watch": Cluster randomised controlled trial evaluating an injury prevention program. *Inj Prev*. 2007;13(2):93–9.
 80. Evans CE, Ransley JK, Christian MS, Greenwood DC, Thomas JD, Cade JE. A cluster-randomised controlled trial of a school-based

- fruit and vegetable intervention: Project Tomato. *Public Health Nutr.* 2013;16(6):1073–81.
81. Redmond CA, Blinkhorn FA, Kay EJ, Davies RM, Worthington HV, Blinkhorn AS. A cluster randomized controlled trial testing the effectiveness of a school-based dental health education program for adolescents. *J Public Health Dentistry.* 1999;59(1):12–7.
 82. Connolly P, Miller S, Kee F, Sloan S, Gildea A, McIntosh E, et al. A cluster randomised controlled trial and evaluation and cost-effectiveness analysis of the Roots of Empathy schools-based programme for improving social and emotional well-being outcomes among 8-to 9-year-olds in Northern Ireland. *Public Health Research.* 2018;6(4).
 83. HM Government. Types of School [Available from: <https://www.gov.uk/types-of-school>]. Accessed 1 May 2021.
 84. Raab GM, Butcher I. Balance in cluster randomized trials. *Stat Med.* 2001;20(3):351–65.
 85. Moulton LH. Covariate-based constrained randomization of group-randomized trials. *Clin Trials.* 2004;1(3):297–305.
 86. Dale A, Marsh C. *The 1991 Census User's Guide.* London: HM Stationery Office; 1993.
 87. HM Government. *The English Indices of Deprivation 2019.* 2019 [Available from: <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019>]. Accessed 1 May 2021.
 88. Eldridge S, Kerry S, Torgerson DJ. Bias in identifying and recruiting participants in cluster randomised trials: what can be done? *BMJ.* 2009;339:b4006.
 89. Bolzern J, Mnyama N, Bosanquet K, Torgerson DJ. A review of cluster randomized trials found statistical evidence of selection bias. *J Clin Epidemiol.* 2018;99:106–12.
 90. Hayes R, Bennett S. Simple sample size calculation for cluster-randomized trials. *Int J Epidemiol.* 1999;28(2):319–26.
 91. Goldstein H, Browne W, Rasbash J. Multilevel modelling of medical data. *Stat Med.* 2002;21(21):3291–315.
 92. HM Government. *Childhood obesity: a plan for action 2019* [Available from: <https://www.gov.uk/government/publications/tackling-obesity-government-strategy>]. Accessed 1 May 2021.
 93. Bonell C, Humphrey N, Fletcher A, Moore L, Anderson R, Campbell R. Why schools should promote students' health and wellbeing. *BMJ.* 2014;348:g3078.
 94. Bonell C, Farah J, Harden A, Wells H, Parry W, Fletcher A, et al. Systematic review of the effects of schools and school environment interventions on health: evidence mapping and synthesis. *Public Health Res.* 2013;1(1).
 95. Bartlett R, Wright T, Olarinde T, Holmes T, Beamon ER, Wallace D. Schools as Sites for Recruiting Participants and Implementing Research. *J Commun Health Nurs.* 2017;34(2):80–8.
 96. Rutterford C, Taljaard M, Dixon S, Copas A, Eldridge S. Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review. *J Clin Epidemiol.* 2015;68(6):716–23.
 97. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *J Clin Epidemiol.* 2004;57(8):785–94.
 98. Campbell MJ. Cluster randomized trials in general (family) practice research. *Stat Methods Med Res.* 2000;9(2):81–94.
 99. Hedges LV, Hedberg EC. Intraclass correlation values for planning group-randomized trials in education. *Educ Eval Policy Anal.* 2007;29(1):60–87.
 100. Ukoumunne O, Gulliford M, Chinn S, Sterne J, Burney P. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technol Assessment.* 1999;3(5).
 101. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. *BMJ.* 2004;328(7441):702–8.
 102. Shackleton N, Hale D, Bonell C, Viner RM. Intra class correlation values for adolescent health outcomes in secondary schools in 21 European countries. *SSM - Popul Health.* 2016;2:217–25.
 103. Hale DR, Patalay P, Fitzgerald-Yau N, Hargreaves DS, Bond L, Görzig A, et al. School-level variation in health outcomes in adolescence: analysis of three longitudinal studies in England. *Prev Sci.* 2014;15(4):600–10.
 104. Murray DM, Short BJ. Intraclass correlation among measures related to tobacco use by adolescents: estimates, correlates, and applications in intervention studies. *Addict Behav.* 1997;22(1):1–12.
 105. HM Government. *Schools, pupils and their characteristics: January 2019.* 2019 [Available from: <https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2019>]. Accessed 1 May 2021.
 106. Murray DM, Pals SL, George SM, Kuzmichev A, Lai GY, Lee JA, et al. Design and analysis of group-randomized trials in cancer: A review of current practices. *Prev Med.* 2018;111:241–7.
 107. Diaz-Ordaz K, Froud R, Sheehan B, Eldridge S. A systematic review of cluster randomised trials in residential facilities for older people suggests how to improve quality. *BMC Med Res Methodol.* 2013;13(1):1–10.
 108. Ivers N, Taljaard M, Dixon S, Bennett C, McRae A, Taleban J, et al. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000–8. *BMJ.* 2011;343.
 109. Froud R, Eldridge S, Diaz Ordaz K, Marinho VCC, Donner A. Quality of cluster randomized controlled trials in oral health: a systematic review of reports published between 2005 and 2009. *Commun Dentistry Oral Epidemiol.* 2012;40:3–14.
 110. HM Government. *Get in formation about schools* [Available from: <https://www.gov.uk/government/organisations/department-for-education>]. Accessed 1 May 2021.
 111. de Hoop E, Teerenstra S, van Gaal BG, Moerbeek M, Borm GF. The "best balance" allocation led to optimal balance in cluster-controlled trials. *J Clin Epidemiol.* 2012;65(2):132–7.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



Appendix 3 – Systematic review of the characteristics of school-based feasibility cluster randomised trials of interventions for improving the health of pupils in the UK

Parker et al. *Pilot and Feasibility Studies* (2022) 8:132
<https://doi.org/10.1186/s40814-022-01098-w>

Pilot and Feasibility Studies

REVIEW

Open Access



Systematic review of the characteristics of school-based feasibility cluster randomised trials of interventions for improving the health of pupils in the UK

Kitty Parker^{1*}, Saskia Eddy², Michael Nunns³, ZhiMin Xiao⁴, Tamsin Ford⁵, Sandra Eldridge² and Obioha C. Ukoumunne⁶

Abstract

Background: The last 20 years have seen a marked increase in the use of cluster randomised trials (CRTs) in schools to evaluate interventions for improving pupil health outcomes. Schools have limited resources and participating in full-scale trials can be challenging and costly, given their main purpose is education. Feasibility studies can be used to identify challenges with implementing interventions and delivering trials. This systematic review summarises methodological characteristics and objectives of school-based cluster randomised feasibility studies in the United Kingdom (UK).

Methods: We systematically searched MEDLINE from inception to 31 December 2020. Eligible papers were school-based feasibility CRTs that included health outcomes measured on pupils.

Results: Of 3285 articles identified, 24 were included. School-based feasibility CRTs have been increasingly used in the UK since the first publication in 2008. Five (21%) studies provided justification for the use of the CRT design. Three (13%) studies provided details of a formal sample size calculation, with only one of these allowing for clustering. The median (IQR; range) recruited sample size was 7.5 (4.5 to 9; 2 to 37) schools and 274 (179 to 557; 29 to 1567) pupils. The most common feasibility objectives were to estimate the potential effectiveness of the intervention ($n = 17$; 71%), assess acceptability of the intervention ($n = 16$; 67%), and estimate the recruitment/retention rates ($n = 15$; 63%). Only one study was used to assess whether cluster randomisation was appropriate, and none of the studies that randomised clusters before recruiting pupils assessed the possibility of recruitment bias. Besides potential effectiveness, cost-effectiveness, and the intra-cluster correlation coefficient, no studies quantified the precision of the feasibility parameter estimates.

Conclusions: Feasibility CRTs are increasingly used in schools prior to definitive trials of interventions for improving health in pupils. The average sample size of studies included in this review would be large enough to estimate pupil-level feasibility parameters (e.g., percentage followed up) with reasonable precision. The review highlights the

*Correspondence: kp477@exeter.ac.uk

¹ NIHR Applied Research Collaboration South West Peninsula, University of Exeter, Room 2.16, South Cloisters, St Luke's Campus, 79 Heavitree Rd, Exeter EX1 2LU, UK

Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

need for clearer sample size justification and better reporting of the precision with which feasibility parameters are estimated. Better use could be made of feasibility CRTs to assess challenges that are specific to the cluster design.

Trial registration: PROSPERO: CRD42020218993.

Keywords: Children, Cluster randomised trials, Feasibility study, Pilot study, Public health, Randomised trials, Research methods, Schools, Systematic review, Trial methodology

Background

Cluster randomised trials (CRTs) are studies in which clusters (groups) of individuals are allocated to trial arms, and outcomes are measured on the individual participants [1]. These clusters might be geographical locations (e.g., cities), organisations (e.g., workplaces) or social units (e.g., households). Clusters may be chosen as the randomisation unit for different reasons, including logistical reasons, to prevent contamination that could otherwise occur between trial arms if individuals were randomised, or because the intervention is designed to be administered at the cluster level [2]. CRTs are often used to investigate complex interventions. They usually require more participants and can be more complicated to design, conduct and analyse than individually randomised controlled trials (RCTs) [1–6]. Therefore, it is important to assess the feasibility of the study processes and design uncertainties before a definitive CRT of intervention effectiveness is conducted.

Prior to a definitive trial, a feasibility study can be used to determine whether the research is something that can be done, whether it should be done and how it should be done [7]. Feasibility studies focus on areas of uncertainty in trial delivery, such as the randomisation process, recruitment and follow-up rates, acceptability to the participants of the trial processes and the intervention itself, implementation of the intervention, data collection processes, selection of outcome measures, potential harms related to the intervention and trial, knowledge of parameters that will inform the sample size calculation for the definitive trial, and potential effectiveness of the intervention. The randomised pilot trial is a type of feasibility study that involves conducting the future definitive trial or part of it on a smaller scale [7]. For ease of understanding, this paper refers to randomised pilot trials as feasibility studies. Other types of feasibility study include non-randomised parallel group and single-arm trials, which also focus on developing trial methodology and interventions, and testing processes prior to a full-scale RCT [7, 8]. However, such designs cannot be used to test specific uncertainties such as the randomisation process and the willingness of participants to be randomised. Feasibility CRTs differ from those done in advance of individually RCTs in that they may be used

to address concerns that are specific to CRTs, including evaluating the possibility for recruitment bias in studies where clusters are randomised before individual participants are recruited [9] and obtaining estimates of the intra-cluster correlation coefficient (ICC) of the primary outcome to support the calculation of the sample size for the definitive trial, although some authors caution that the resulting estimates will often be imprecise due to the small number of clusters typically included in such studies [10]. Other general feasibility considerations apply at both the cluster and individual levels, such as ease of recruitment, rate of loss to follow-up and acceptability of the intervention. Methodological considerations that are unique to the conduct of feasibility CRTs include the need to take account for clustering when calculating the sample size for and reporting the precision of feasibility parameter estimates from such studies [10].

In recent years, CRTs have been increasingly used to evaluate interventions for improving educational outcomes in schools [11] and complex interventions for improving child health outcomes [12–14]. Schools provide a natural environment in which to recruit and deliver public health interventions to children due to the amount of time they spend there [13]. The CRT design is suited to the natural clustered structure found in schools (pupils within classes within schools), but there are challenges to delivering trials in this setting. For example, schools and teachers often have stretched and limited resources, and implementing an intervention and participating in a trial can be challenging, given that the primary focus of schools is the education of pupils. A recent systematic review of definitive school-based CRTs found that 52% of the studies required a member of school staff to deliver components of the intervention [14]. Obtaining a representative sample of schools is important for external validity and inclusiveness [13], but recruitment of schools and pupils is also a challenge. Another potential feasibility issue regards which type of cluster to randomise in the school setting for a given trial, such as entire schools, year groups, classrooms or teachers. For example, there may be a choice between randomising schools and randomising classrooms; the former would be better to minimise the chance of contamination between trial arms but the latter would have the advantage of a smaller design effect

[1] and, therefore, greater power for a fixed total number of recruited pupils [15]. In comparison to the primary care setting, CRTs for evaluating health interventions have only relatively recently been used in schools in the UK and, therefore, there is a smaller pool of experience available from previous studies [1, 14]. Given these uncertainties, feasibility trials have an important role to play in the design and execution of definitive school-based CRTs.

Authors have previously discussed the growing literature described as ‘feasibility’ or ‘pilot’ studies, and the associated methodological challenges [7]. The characteristics of feasibility studies generally [10, 16, 17] and cluster randomised feasibility studies specifically [18, 19] have been summarised, but, to date, no systematic review has focussed on the characteristics of school-based feasibility CRTs for improving pupil health outcomes. The aim of this systematic review is to summarise the key design features and report the feasibility-related objectives of school-based feasibility CRTs in the United Kingdom (UK) that measure health outcomes on pupils. It follows our previous systematic review of full-scale definitive CRTs in the school setting [14]. Through summarising the design features of these studies, the findings of this review will highlight particular areas where improvements could be made to the conduct of feasibility CRTs. The reporting of their feasibility objectives will help identify areas in which better use of such studies could be made to address uncertainties that are specific to the CRT design.

Methods

Data sources and search methods

This review has been reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [20] as evidenced in the PRISMA checklist (see Additional file 1: Table S1) and was registered with PROSPERO (ID CRD: 42,020,218,993; www.crd.york.ac.uk/prospero).

Peer-reviewed school-based feasibility CRTs, indexed on MEDLINE (through Ovid), were the source of data for the review. MEDLINE was systematically searched from inception to 31 December 2020. A pragmatic decision was made to search MEDLINE only due to time constraints and available resources. The search strategy (Table 1) was developed using terms from the MEDLINE search strategy by Taljaard et al. [21] to identify CRTs, and this was combined with *school* concept terms, including the ‘Schools’ MeSH term. This was the same search strategy used in our previous systematic review to identify definitive school-based CRTs [14]. The search was limited to English language papers.

Table 1 Systematic review search strategy

Search strategy
Terms for Randomised Controlled trials:
1. random:mp
2. trial:ab, kw, ti
Cluster design-related terms:
3. *cluster*:ab, kw, ti
4. *communit*:ab, kw, ti
5. group*adj2 random*.ab, kw, ti
6. 3 OR 4 OR 5
School terms:
7. exp Schools/
8. School*.ab, kw, ti
9. 7 OR 8
Final search stages:
10. 1 AND 2 AND 6 AND 9
11. 10 limited to English language

Inclusion and exclusion criteria

The review included school-based feasibility CRTs that measured health outcomes on pupils and were conducted in the UK. It focussed on the UK to align with available resources and to summarise data from a single education system relevant to the research team. The population of included studies was pupils attending pre-school, primary or secondary school in the UK. ‘Pre-school’ was defined as an organisation offering early childhood education to children before they begin compulsory education (i.e., primary school). This included nursery schools and kindergartens. Eligible clusters could be any school-related unit (e.g., schools, classes, year groups). Studies that randomised school-related units as well as other types of clusters (e.g., towns, hospitals, households) were eligible for inclusion in the review as long as the results of the study were shown separately for the school clusters (i.e., the authors did not pool results across the different types of clusters). Any health-related intervention(s) were eligible. The primary outcome had to be measured on pupils and be health related. Studies with education-related primary outcomes were excluded. All types of CRT design were eligible, including parallel group, factorial, crossover and stepped wedge trials.

Only randomised external feasibility studies were included in the systematic review. The definition of feasibility study used to identify eligible papers was that used by Eldridge et al. [7] which states “A feasibility study asks whether something can be done, should we proceed with it, and if so, how.” Therefore, eligible studies had to be assessing some element of feasibility in the intervention and/or trial methodology, ahead of a definitive trial. This was determined by looking for the terms, ‘pilot’

'feasibility' or 'explanatory' in the title and abstract and by examining the aims and objectives of each study. Internal pilot studies that are part of the actual definitive trial, where the data from the pilot phase are included in the main analysis [22] were excluded. Non-randomised parallel group feasibility studies and single-arm feasibility studies were excluded. Definitive CRTs were not eligible for inclusion in this review.

If there was more than one publication of the results for an eligible feasibility CRT, the paper presenting quantitative results related to the feasibility objectives was designated the key study report (index paper) and used for data extraction. Papers that did not report the results of the feasibility objectives were excluded along with protocol/design articles, cost-effectiveness/economic evaluations and process evaluations.

Sifting and validation

Titles and abstracts were downloaded into Endnote [23] and screened by two independent reviewers (KP & SEd/OU) for eligibility against inclusion criteria. Studies for which inclusion status was uncertain were included for full-text screening. Full-text articles were assessed against inclusion criteria by two reviewers (KP & SEd) using a pre-piloted coding method. Any uncertainties were resolved by consulting a third reviewer (OU).

Data extraction

The data extraction form was pre-piloted in Microsoft Excel by KP and SEd. One investigator (KP) extracted data from all included studies. A second reviewer (SEd or OU) independently extracted data for validation. If there was uncertainty regarding a particular article, the data obtained were checked by another member of the team (MN) and resolved by further discussion.

The items of information extracted are listed as follows:

- *Publication details*: year of publication, journal name, funding source and trial registration status.
- *Setting characteristics*: country (England, Scotland, Wales, Northern Ireland) in which the trial took place, school level, types of school recruited and participant information.
- *Intervention information*: health area, intervention description and type of control arm.
- *Primary outcome information*: name of primary outcome.

- *Study design*: justification for using cluster trial design, type of cluster, method of randomisation, timing of randomisation of clusters relative to recruitment of pupils, number of trial arms, allocation ratio and length of follow-up.

- *Sample size information*: justification for sample size, targeted number of schools, clusters and pupils; number of recruited schools, clusters and pupils.

- *Objectives of feasibility study*: test randomisation process (yes/no), test willingness to be randomised (at cluster and/or individual levels) (yes/no), estimate recruitment rate (at cluster and/or individual levels) (yes/no), estimate retention/follow-up rate (at cluster and/or individual levels) (yes/no), test implementation of the intervention (yes/no), test compliance with the intervention (yes/no), assess acceptability of the intervention (at cluster and/or individual levels) (yes/no), assess acceptability of trial procedures (at cluster and/or individual levels) (yes/no), test the feasibility of blinding procedures (yes/no), test data collection process (yes/no), test outcome measures (yes/no), estimate standard deviation for continuous outcomes (or control arm rate for binary outcomes) (yes/no), test consent procedures (yes/no), identify potential harms (yes/no), estimate potential effectiveness of intervention (yes/no), estimate costs of delivering the intervention (yes/no), estimate the intra-cluster correlation coefficient (ICC) of the primary outcome (yes/no) and calculate the sample size required for the definitive trial (yes/no). Only formal feasibility objectives were extracted; these were obtained from the Background and Methods sections of the included articles.

- *Ethics and consent procedures*: Was ethical approval provided? (yes/no).

- *Other design characteristics of methodological interest*: analysis method used to estimate potential effectiveness of the intervention, baseline cluster-level characteristics, ICC estimates (and 95% confidence intervals (CIs)) and whether study concluded that a definitive trial is feasible (yes/yes (with modifications)/no).

Data analysis

Study characteristics were described using medians, interquartile ranges (IQRs) and ranges for continuous variables,

and numbers and percentages for categorical variables, using Stata 17 software [24]. Formal quality assessment of the papers was not performed as it was not necessary for summarising characteristics of studies. However, some of the data extracted and summarised in the review are indicative of the reporting quality of included studies based on the items in the CONSORT extension for both CRTs [25] and pilot studies [26]. This includes details on the rationale for using the CRT design, the rationale for the target sample size and ethical approval procedures.

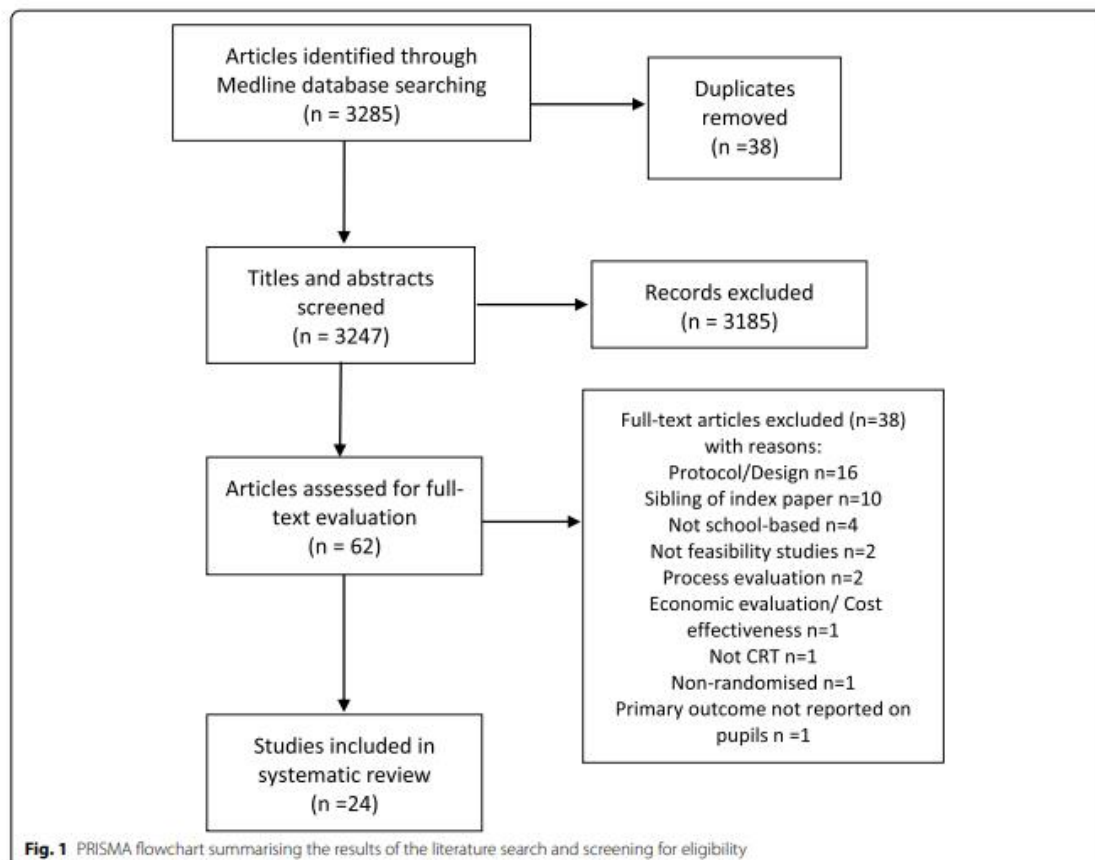
Results

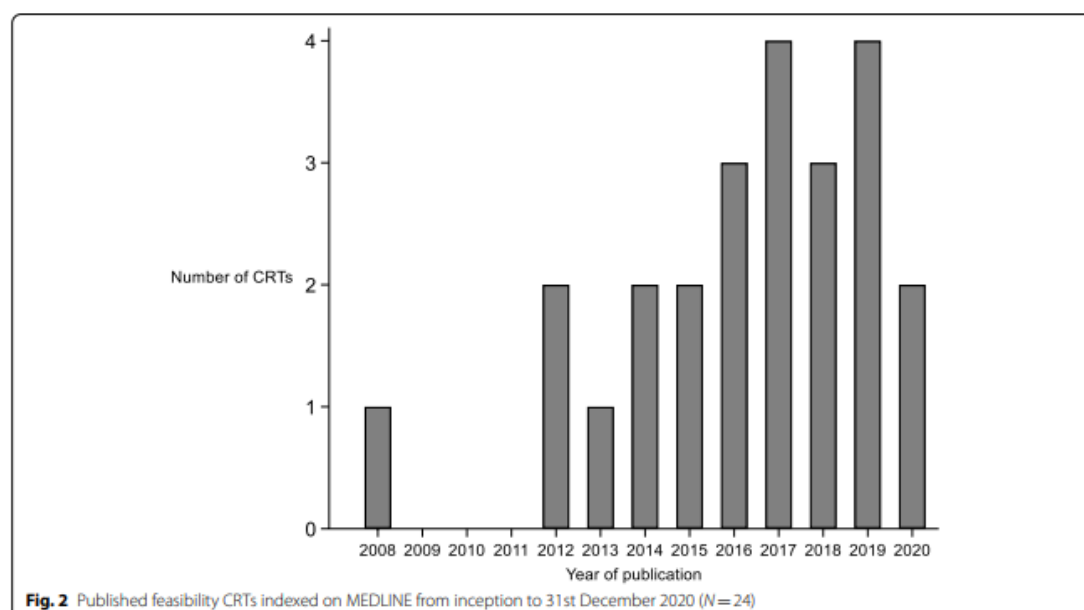
Search results

After deduplication, 3247 articles were identified through MEDLINE, 62 were eligible for full-text screening and 24 were included in the review [27–50]. Out of 38 excluded studies, 28 did not meet the inclusion criteria, and 10 met inclusion criteria but were excluded as they described the same study as a designated ‘index paper’. The PRISMA flow diagram [20] is shown in Fig. 1.

Study characteristics

School-based feasibility CRTs for health interventions on pupils have been increasingly used in the UK since the first publication in 2008 (Fig. 2). Included articles were published across 11 different journals: *Pilot and Feasibility Studies* (n=5), *International Journal of Behavioural Nutrition and Physical Activity* (n=4), *Public Health Research* (n=4), *BMJ Open* (n=3), *Health Technology Assessment* (n=2), *Archives of Disease in Childhood* (n=1), *BMC Public Health* (n=1), *British Journal of Cancer* (n=1), *British Journal of Psychiatry* (n=1), *Prevention Science* (n=1) and *Trials* (n=1). Ten articles described their study as a ‘pilot trial’, six as a ‘feasibility trial’, four as a ‘feasibility study’, two as an ‘exploratory trial’, one as a ‘pilot feasibility trial’ and one as a ‘pilot study’. Twelve (50%) studies were funded by the *National Institute for Health Research*. Eight (33%) studies were registered prospectively, thirteen (54%) retrospectively, and three (13%) did not state registration status.





Tables 2 and 3 summarise the characteristics of included studies.

Setting

Three quarters of studies ($n = 18$; 75%) took place in England. Just over half ($n = 13$; 54%) took place exclusively in secondary schools, 8 (33%) took place exclusively in primary schools, 2 (8%) exclusively in pre-schools and 1 (4%) study included both primary and secondary schools. Fifteen (63%) studies provided information about the types of schools included in their sample and, of these, 14 (93%) included “state” schools.

Intervention and control type

Eleven (46%) studies delivered interventions for improving physical activity, 4 (17%) in physical activity and nutrition, 2 (8%) in alcohol misuse, 2 (8%) in sexual health and 1 (4%) in each of illicit drug misuse, bullying, behavioural/social difficulties, body image, and dating and relationship violence.

The main types of intervention components included resources and materials for schools ($n = 11$; 46%), classroom lessons ($n = 10$; 42%) and physical activity lessons ($n = 5$; 21%). Almost all studies ($n = 23$, 96%) had intervention components that had to be delivered to entire clusters (‘cluster-cluster’ interventions [1] (pages 25 to 30))—e.g., classroom-delivered lessons [48] and physical activity sessions [27]. Two (8%) had intervention

components that were directed at individual pupils (‘individual-cluster’ interventions [1])—e.g., goal-setting [40, 50]. Eighteen (75%) had intervention components that were delivered by a professional or person internal to the cluster (‘professional-cluster’ interventions [1])—e.g., teachers [34], member of school staff [27] and fellow pupils/peers [46]. Eight studies (33%) had intervention components that were delivered by someone external to the cluster (‘external-cluster’ interventions [1])—e.g., ‘active play practitioners’ [38], researchers [41] and dance teachers [36].

The most common type of control arm was usual care ($n = 21$; 88%). Two (8%) studies used an active control arm, and one (4%) study had two control arms (a usual care arm and an active control arm).

Study design

Justification for the use of the CRT design was provided in only 5 (21%) studies. The reasons given were that the intervention was designed to be delivered to entire clusters [30, 47, 48] and to minimise contamination between trial arms [44, 50]. Twenty-three (96%) studies randomised schools and the remaining study randomised classrooms [48]. In the latter study [48], random allocation was carried out at the level of the classroom for ‘pragmatic considerations’. Thirteen (54%) studies used some form of restricted allocation to balance cluster characteristics between the trial arms.

Table 2 Characteristics of included studies (N = 24)

Author	Year of publication	School level	Cluster unit	Health area
Kipping [39]	2008	Primary	Schools	Physical activity and nutrition
Jago [36]	2012	Secondary	Schools	Physical activity
Lloyd [40]	2012	Primary	Schools	Physical activity and nutrition
Sharpe [48]	2013	Secondary	Classes	Body image
Jago [37]	2014	Primary	Schools	Physical activity
Newbury-Birch [44]	2014	Secondary	Schools	Alcohol misuse
Bonell [28]	2015	Secondary	Schools	Bullying
Segrott [47]	2015	Primary	Schools	Alcohol misuse
Barber [27]	2016	Pre-school	Schools	Physical activity
Corder [31]	2016	Secondary	Schools	Physical activity
Wright [50]	2016	Primary and secondary	Schools	Behavioural/social difficulties (Autism)
Forster [33]	2017	Secondary	Schools	Sexual health (Cancer)
Ginja [35]	2017	Primary	Schools	Physical activity
McSweeney [42]	2017	Pre-school	Schools	Physical activity and nutrition
White [49]	2017	Secondary	Schools	Illicit drug misuse
Carlin [29]	2018	Secondary	Schools	Physical activity
Lohan [41]	2018	Secondary	Schools	Sexual health
Sebire [46]	2018	Secondary	Schools	Physical activity
Corepal [32]	2019	Secondary	Schools	Physical activity
Gammon [34]	2019	Secondary	Schools	Physical activity
Johnstone [38]	2019	Primary	Schools	Physical activity
Sahota [45]	2019	Primary	Schools	Physical activity and nutrition
Clemes [30]	2020	Primary	Schools	Physical activity
Meiksin [43]	2020	Secondary	Schools	Dating and relationship violence

Most studies ($n=21$; 88%) had two trial arms and most allocated clusters in a 1:1 ratio ($n=17$; 71%). The median (IQR; range) length of follow up was 7 (3 to 12; 2 to 24) months.

Twelve (50%) studies recruited pupils before randomisation of clusters, four (17%) randomised clusters before recruiting pupils, and in eight (33%) studies, it was unclear whether or not randomisation occurred before pupils were recruited. Only 13 (54%) studies reported baseline characteristics of the schools.

Ethical approval

Ethical approval was obtained and reported in 22 (92%) studies. One study stated that ethical approval was sought but the local research committee said it was not required as the study did not involve patients or NHS staff. The remaining study did not state whether ethical approval was obtained.

Sample size

Of the 24 studies included in this review, three (13%) provided details of a formal sample size calculation. One of these studies based their sample size on being able to estimate feasibility parameters (e.g., participation rates,

questionnaire response rates) with a specified level of precision [33], and the other two studies based their sample size on power to detect a specified intervention effect [29, 48]. Only one (4%) study allowed for clustering in their sample size calculation [48]. Nineteen studies provided informal justification for their sample size calculation, based on one or more reasons: seven (29%) studies based their target sample size on recommendations from previous articles, six (25%) studies stated that a formal sample size calculation was not needed, four (17%) studies said their target sample size was determined by resource and/or time constraints, three (13%) studies provided a general statement that their sample size was considered sufficient to address the objectives of the feasibility CRT, and one (4%) study aimed to recruit as many clusters and participants as possible. Two (8%) studies did not provide any justification for their choice of sample size.

The median (IQR; range) target sample size was 7.5 (5 to 8; 2 to 20) schools, 7.5 (5 to 8; 2 to 20) clusters and 320 (150 to 1200; 50 to 1852) pupils. The median (IQR; range) achieved sample size was 7.5 (4.5 to 9; 2 to 37) schools, 8 (5.5 to 9.5; 2 to 37) clusters and 274 (179 to 557; 29 to 1567) pupils. Two studies included

Table 3 Summary of methodological characteristics of included studies (N = 24)

Characteristic	N	Statistic
Setting		
<i>Country</i>	24	
England, n (%)		18 (75)
Scotland, n (%)		1 (4)
Wales, n (%)		2 (8)
Northern Ireland, n (%)		3 (13)
<i>School types that were included [51](Accessed 1st September 2021)^a</i>	15	
State, n (%)		14 (93)
Academy, n (%)		3 (20)
Voluntary aided, n (%)		1 (7)
Foundation, n (%)		1 (7)
Faith, n (%)		1 (7)
Grammar, n (%)		1 (7)
Independent, n (%)		1 (7)
Intervention		
<i>Type of intervention [1]^b</i>	24	
Individual-cluster, n (%)		2 (8)
Professional-cluster, n (%)		18 (75)
External-cluster, n (%)		8 (33)
Cluster-cluster, n (%)		23 (96)
Multifaceted, n (%)		21 (88)
<i>Intervention components^c</i>	24	
Resources and materials for schools, n (%)		11 (46)
Classroom lessons, n (%)		10 (42)
Physical activity lessons, n (%)		5 (21)
Incentive scheme, n (%)		4 (17)
Change in school/classroom environment, n (%)		4 (17)
Peer support, n (%)		3 (13)
Support for parents/guardians, n (%)		3 (13)
Goal setting, n (%)		2 (8)
Staff training, n (%)		2 (8)
Home activities, n (%)		2 (8)
Extracurricular physical activity, n (%)		2 (8)
Parent's evenings, n (%)		1 (4)
Drama workshops, n (%)		1 (4)
Funding, n (%)		1 (4)
School action group formation, n (%)		1 (4)
School club sessions, n (%)		1 (4)
Screening, n (%)		1 (4)
Feedback, n (%)		1 (4)
Motivational interviews, n (%)		1 (4)
Interactive sessions, n (%)		1 (4)
Discussions with parents/guardians, n (%)		1 (4)
Gamification (competitive) techniques, n (%)		1 (4)
<i>Type of control group</i>	24	
Usual care, n (%)		21 (88)
Active, n (%)		2 (8)
Two control groups (one usual care and one active control), n (%)		1 (4)

Table 3 (continued)

Characteristic	N	Statistic
Study design		
<i>Justification for CRT design</i>	24	
Yes, n (%)		5 (21)
<i>Type of randomisation</i>	24	
Completely randomised, n (%)		11 (46)
Minimisation, n (%)		5 (21)
Stratified, n (%)		4 (17)
Matched pair, n (%)		3 (13)
Constrained [52, 53], n (%)		1 (4)
<i>Number of trial conditions</i>	24	
Two, n (%)		21 (88)
Three, n (%)		2 (8)
Four, n (%)		1 (4)
<i>Length of follow-up</i>	24	
Up to 6 months, n (%)		11 (46)
7 to 12 months, n (%)		8 (33)
13 to 18 months, n (%)		3 (13)
More than 18 months, n (%)		1 (4)
Not stated, n (%)		1 (4)
<i>Were pupils recruited before randomisation of clusters?</i>	24	
Pupils recruited before randomisation, n (%)		12 (50)
Pupils recruited after randomisation, n (%)		4 (17)
Unclear, n (%)		8 (33)
<i>Were baseline cluster-level characteristics reported?</i>	24	
Yes, n (%)		13 (54)
Ethical approval		
<i>Was ethical approval obtained?</i>	24	
Yes, n (%)		22 (92)
No, n (%)		1 (4)
Not stated, n (%)		1 (4)
Sample size		
<i>Type of justification for sample size</i>	24	
Formal sample size calculation ^d , n (%)		3 (13)
Other justification, n (%)		19 (79)
Not stated, n (%)		2 (8)
<i>Target number of schools, median (IQR; range)</i>	18	7.5 (5 to 8; 2 to 20)
<i>Target number of clusters, median (IQR; range)</i>	18	7.5 (5 to 8; 2 to 20)
<i>Target number of pupils, median (IQR; range)</i>	13	320 (150 to 1200; 50 to 1852)
<i>Achieved number of schools, median (IQR; range)</i>	24	7.5 (4.5 to 9; 2 to 37)
<i>Achieved number of clusters, median (IQR; range)</i>	24	8 (5.5 to 9.5; 2 to 37)
<i>Achieved number of pupils, median (IQR; range)</i>	24	274 (179 to 557; 29 to 1567)
<i>Achieved mean cluster size, median (IQR; range)</i>	24	35.9 (24 to 89.4; 1.4 to 237.7)
Objectives of the feasibility study		
<i>Feasibility objectives</i>	24	
Test randomisation process, n (%)		3 (13)
Test data collection process, n (%)		8 (33)
Test willingness to be randomised (at cluster level and/or individual levels), n (%)		4 (17)
Estimate recruitment percentage (at cluster level and/or individual levels), n (%)		15 (63)
Estimate follow-up percentage (at cluster level and/or individual levels), n (%)		15 (63)

Table 3 (continued)

Characteristic	N	Statistic
Test implementation of intervention, n (%)		10 (42)
Test compliance with intervention, n (%)		6 (25)
Assess acceptability of intervention (at cluster level and/or individual levels), n (%)		16 (67)
Assess acceptability of trial procedures (at cluster level and/or individual levels), n (%)		6 (25)
Test the feasibility of blinding procedures, n (%)		0 (0)
Test outcome measures, n (%)		14 (58)
Estimate standard deviation of continuous outcomes or control arm rate for binary outcomes, n (%)		1 (4)
Test consent procedures, n (%)		0 (0)
Identify potential harms, n (%)		3 (13)
Assess potential effectiveness of intervention, n (%)		17 (71)
Estimate intervention cost, n (%)		7 (29)
Estimate the ICC of the primary outcome, n (%)		2 (8)
Estimate sample size for definitive trial, n (%)		5 (21)
Other study characteristics of methodological interest		
<i>Analysis method for estimating potential effectiveness</i>	24	
Individual-level analysis that allows for clustering, n (%)		9 (38)
Cluster-level analysis, n (%)		4 (17)
Did not account for clustering, n (%)		4 (17)
Not stated, n (%)		3 (13)
Did not estimate potential effectiveness, n (%)		4 (17)
<i>P-value reported for effectiveness</i>	24	
Yes, n (%)		8 (33)

^a Some studies included more than one school type. This is the number of studies that included specific types of school. State schools receive funding through their local authority or directly from the government. The most common ones are local authority, foundation and voluntary aided school which are all funded by the local authority. Academies are run by government and not-for-profit trusts, and are independent of local authority. Grammar schools are run by local authorities but intake is based on assessment of the pupils' academic ability. Special schools cater for pupils with special educational needs. Faith schools follow the national curriculum but can decide what they teach in religious studies. Independent schools follow the national curriculum but charge fees for attending pupils

^b Intervention type has been described using the typology of Eldridge and Kerry [1]. 'Individual-cluster' interventions contain components that are aimed at the individual level (e.g., goal setting). 'Professional-cluster' interventions contain components that are delivered by a professional or person internal to the cluster (e.g., teacher, pupils). 'External-cluster' interventions contain components that require people external to the cluster to deliver the intervention (e.g., research staff, community support consultant). 'Cluster-cluster' interventions contain components that have to be delivered at the cluster level (e.g., classroom lessons). 'Multifaceted' interventions contain components across more than one of the 'individual-cluster', 'professional-cluster', 'external-cluster' and 'cluster-cluster' categories

^c Examples of each intervention component are provided for ease of understanding. Resources and materials (e.g., a resource box comprising food models, food mats, food cards, DVDs, and books); Classroom lessons (e.g., interactive film-based sexual-health lesson); Physical activity lessons (e.g., active play sessions, brisk walking programme during the school day); Incentive schemes (e.g., lottery-based incentive scheme to promote active travel to school); Peer support (e.g., informal peer-led smoking prevention); Change in school/classroom environment (e.g., sit-stand desks to replace standard desks, challenging attitudes and perceived norms concerning gender stereotypes and dating and relationship violence); Support for parents/guardians (e.g., information sheets about health eating habits); Goal setting (e.g., goal setting to engage and support schools); Staff training (e.g., staff training in restorative school action group formation); Home activities (e.g., home activities that encourage pupils to be more active, eat more nutritious foods, and spend less time in screen-based activities); Extracurricular physical activity (e.g., staff delivered after-school physical activity programme); Drama workshops (e.g., interactive drama workshops); School action group formation (e.g., to address bullying and aggression within schools); School club sessions (e.g., health eating club); Screening (e.g., alcohol screening and brief intervention to reduce hazardous drinking in younger adolescents); Feedback (e.g., feedback about pupil's drinking habits); Motivational interviews (e.g., motivational interviewing techniques to prevent alcohol misuse); Interactive sessions (e.g., interactive sessions with school learning mentors to prevent alcohol misuse); Discussions with parents/guardians (e.g., guided discussions conducted with parents); Gamification (competitive) techniques (e.g., gamification techniques to promote physical activity)

^d In one study, the sample size was based on being able to estimate feasibility parameters with a pre-specified level of precision. Two studies based their sample size on a definitive test of intervention effectiveness

just 2 schools, with 1 school allocated to each trial arm [34, 35]. The studies that reported both targeted and achieved recruitment numbers at the cluster ($n = 18$) and pupil ($n = 13$) levels achieved those targets in 94% and 46% of studies, respectively.

Objectives of feasibility study

Formal feasibility objectives were specified by all 24 studies (summarised in Table 3). Of the 18 objectives assessed in this review, the median (IQR; range) number addressed per study was 5 (4 to 7.5; 1 to 11). The most

common objectives were to estimate the potential effectiveness of the intervention ($n=17$; 71%; including two studies that sought to undertake a definitive test of effectiveness [29, 48]), assess acceptability of the intervention ($n=16$; 67%), estimate the recruitment rate ($n=15$; 63%), estimate the retention/follow-up rate ($n=15$; 63%) and test outcome measures ($n=14$; 58%). Two studies included estimation of the intra-cluster correlation coefficient of the primary outcome to be used in the planned definitive study as a formal objective of the feasibility study. No studies tested the feasibility of blinding or consent procedures. All studies reported additional feasibility outcomes beyond those formally stated as objectives.

The following feasibility objectives were stated specifically at the level of the cluster: assess acceptability of the intervention ($n=10$; 42%), estimate retention/follow-up rate ($n=7$; 29%), estimate recruitment rate ($n=6$; 25%), assess willingness to be randomised ($n=4$; 17%) and assess acceptability of the trial procedures ($n=3$; 13%). One (4%) feasibility CRT had the formal objective of assessing the appropriateness of cluster randomisation [50]. None of the feasibility studies used their research to assess the type of cluster that should be randomised. Of the 4 studies that randomised clusters before recruiting pupils, none investigated the possibility of recruitment bias.

Analyses were undertaken to investigate if the target sample size differed according to whether or not the studies addressed specific feasibility objectives. Many objectives were only formally stated in a small number of studies; therefore, it was hard to identify clear patterns in the data. The twelve studies that assessed potential effectiveness aimed to recruit a median (IQR; range) of 7 (3.5 to 8; 2 to 20) schools, similar to the targeted recruitment in the remaining studies (7.5 (6 to 8; 5 to 12)).

All studies reported estimates of feasibility parameters, but, other than for potential intervention effectiveness,

cost-effectiveness and the intra-cluster correlation coefficient, no studies quantified the precision of these estimates. Five of the eight (63%) studies that reported estimates of the ICC for the provisional primary outcome of the planned definitive study provided 95% confidence intervals (95 CIs) for these. Table 4 reports the ICC estimates. As expected the 95% confidence intervals were generally wide given that the sample size is small for estimating the ICC. Notably, however, the upper bound for two ICC estimates was only 0.03, which provides useful information on plausible true values of the parameter despite those studies having only 6 [46] and 19 [39] clusters.

Of the 20 studies that reported intervention effect estimates, nine (45%) used an adjusted individual-level analysis method to allow for clustering, 4 (20%) used a cluster-level analysis method, four (20%) did not allow for clustering and three (15%) did not state the analytical method. Eight studies reported p values with the intervention effect estimate, contrary to published guidance for feasibility studies [25, 26].

Eleven (46%) studies concluded that the definitive trial was feasible, 11 (46%) said the definitive trial would be feasible with modifications and two (8%) said that the planned study was not feasible. Through searching the literature and personal correspondence with the authors, it was established that of the 24 feasibility CRTs included in the review, 11 are known to have progressed to definitive trials [28, 29, 31, 36, 39–41, 44, 46, 49, 50]. Of these, nine had concluded that the definitive trial was feasible, and two had concluded that the definitive trial would be feasible with modifications.

Discussion

Main findings

This is the first systematic review to summarise the characteristics and objectives of school-based

Table 4 Reported intra-cluster correlation coefficients for primary outcomes ($N=8$)

Author (Year)	Cluster unit	Health area	Outcome	Outcome type	ICC (95% CI)
Jago (2012) [36]	Schools	Physical activity	MVPA (minutes per weekday)	Continuous	0.018 (< 0.001 to 0.087)
Jago (2014) [37]	Schools	Physical activity	MVPA (minutes per weekday)	Continuous	0.0653 (0.00091 to 0.12977)
Kipping (2008) [39]	Schools	Physical activity and nutrition	Minutes spent on screen-based activities	Continuous	0.01 (0 to 0.03)
Lloyd (2012) [40]	Schools	Physical activity and nutrition	BMI SD score	Continuous	0.04 (0 to 0.15)
Sahota (2019) [45]	Schools	Physical activity and nutrition	Healthy nutrition and physical activity knowledge	Continuous	0.07 (Not provided)
Sebire (2018) [46]	Schools	Physical activity	MVPA (minutes per weekday)	Continuous	< 0.0001 (0.0 to 0.03)
Segrott (2015) [47]	Schools	Alcohol misuse	Drinking initiation	Binary	0.112 (Not provided)
White (2017) [49]	Schools	Illicit drug misuse	Lifetime illicit drug use	Binary	0.003 (Not provided)

BMI Body mass index, *CI* Confidence interval, *ICC* Intra-cluster correlation coefficient, *MVPA* Moderate to vigorous physical activity, *SD* Standard deviation

feasibility CRTs of interventions to improve pupil health outcomes in the UK. The review found an increase in such studies since the earliest included paper was published in 2008. This mirrors the increase in definitive CRTs in this area reported in our parallel review [14] and highlights the rising popularity of health-based CRTs in the school-setting. The increase in feasibility CRTs may partly be due to the publication of the 2006 MRC guidelines for the evaluation of complex interventions [54] which highlights the importance of conducting feasibility studies ahead of full-scale trials. The relatively large number of feasibility CRTs with interventions for increasing physical activity indicates the growing importance of adolescent physical activity as a public health priority, and the use of schools in order to deliver these types of intervention [55]. The review of school-based definitive CRTs also reflected the increasing use of the design to evaluate physical activity interventions [14]. Based on what was observed in the review of definitive school-based CRTs, there were fewer than expected feasibility studies in the area of socioemotional functioning. This is despite the increased awareness of the prevalence of these health conditions and research funding in this area [56].

A previous review of feasibility CRTs found that, among other objectives, assessing the implementation of the intervention ($n=9$, 50%) was the most common [18]. The studies included in the current review sought to address a range of feasibility objectives; most commonly estimating potential effectiveness of the intervention, assessing acceptability of the intervention, estimating the recruitment and follow-up rates and testing the outcome measures. It was notable, however, that few studies formally stated objectives that were related to uncertainties that are unique to the cluster design. This finding is similar to another review of feasibility CRTs which also stated that few studies investigated issues specific to the complexities of the design [19]. None of the 4 studies that randomised clusters before recruiting pupils investigated the potential for recruitment bias as a feasibility objective. In the current review, only one study assessed whether a cluster design was needed, and none used the research to decide on the type of school-based cluster (e.g., school versus classroom) that was best to randomise. It may be the case that the need for cluster randomisation and the appropriate type of cluster to allocate had a strong theoretical basis, negating the need for empirical justification, but only 5 of the 24 studies provided a rationale for the cluster design even though the CONSORT extension for CRTs [25] recommends reporting this.

The studies included in this review were heterogeneous in their formal feasibility objectives, and this may have influenced specific features of their design, such as

sample size and length of follow-up. The designs may also have been influenced by other factors such as budget, time and practical constraints.

Only three (13%) studies in the review reported details of a formal calculation for the target sample size [29, 33, 48], and only one accounted for clustering in the sample size calculation [48]. These results are similar to that found in a previous systematic review of feasibility CRTs which reported that only one of the 18 studies reported a formal sample size calculation based on the primary feasibility objective [18]. A quarter of the included papers in the current review stated that a formal sample size calculation was not needed, and some authors have argued that it is not always appropriate in feasibility studies [16]. In a recent review of current practice in feasibility studies, only 36% reported sample size calculations [57]. Also, when surveyed, some journal editors stated they were willing to accept pilot studies for publication that did not report a sample size calculation [57]. The precision with which parameters are estimated in feasibility CRTs should be reported, especially given the small number of clusters that are typically included in such studies. Despite this, apart from when assessing the effectiveness of the intervention, cost-effectiveness and estimating the ICC, this was not done by any papers in the current review. Correspondingly, a formal sample size calculation based on the feasibility objectives that allows for clustering [10] is appropriate to estimate parameters precisely and, therefore, minimise the uncertainty regarding the assumptions that are made for the subsequent definitive study [16, 57].

Our review found the median number of clusters recruited (eight) was similar to a previous review of feasibility studies [18]. Based on results from a simulation study, it has been suggested that as many as 30 or more clusters may be required in a feasibility CRT in order to avoid downwardly biased and imprecise estimates of the number of clusters required to test the intervention effect in the subsequent definitive CRT; this is largely due to the imprecision with the ICC is estimated in the feasibility study [10]. The current review found only one study that recruited more than 30 clusters [50], and it is difficult to achieve this level of recruitment due to funding and practical constraints. Smaller feasibility studies may, however, still provide informative estimates of many parameters. Two of the feasibility studies in the review, despite including only 6 [46] and 19 [39] clusters, were able to estimate the intra-cluster correlation coefficient with a 95% confidence interval upper bound of 0.03, which could rule out the need for unattainably large sample sizes in the definitive study. Many studies report feasibility objectives in the form of percentages (e.g., follow-up rates, intervention adherence rates). Eldridge and colleagues [10] provide formulae for calculating the

sample size required in feasibility CRTs to estimate percentages based on individual-level characteristics (e.g., whether the pupil was followed up) with a confidence interval of specified width, whilst allowing for clustering. Assuming the ICC for the feasibility characteristic is 0.05, a study with 8 schools and 240 pupils (an average sample size based on the findings in the current review) is large enough to estimate the percentage with a margin of error no greater than 10 percentage points based on a 95% confidence interval. There will generally be little precision for estimating percentages based on cluster-level characteristics since this is determined by the, typically, small number of schools (clusters) in feasibility studies.

Another important reason to recruit sufficient clusters to feasibility CRTs is to assess how the intervention might be implemented and the trial delivered in a range of different types of cluster [18]. Parameter estimates will only be useful to the extent that the clusters and individuals in the feasibility study are broadly representative and reflect the diversity of the population from which the sample in the definitive trial will be drawn [18]. In the context of school-based trials, important aspects of representativeness include single sex versus co-educational schools, state versus independent schools, and deprived versus non-deprived areas. In the current review, only 54% of studies reported baseline characteristics of the schools, although this is higher than found in a previous systematic review of feasibility CRTs where only 11% of studies reported baseline cluster-level characteristics [18].

The current systematic review found that of the 13 studies that reported both targeted and achieved numbers of pupils recruited, those targets were only achieved in 46% of studies. Our previous systematic review of definitive school-based CRTs found that only 77% of studies achieved their target recruitment of pupils [14]. The facilitators and barriers to the recruitment and retention of pupils to school-based CRTs have been discussed in detail in the literature [58–60], including the type of intervention being offered and the perceived benefits of the study (e.g., sexual education) [58, 60], lack of time [58], incompatibility of the intervention with the needs of pupils or parents or with the school's ethos [58] and a lack of incentivisation [59].

Strengths and limitations

A strength of the review is that a predefined search strategy was used to identify feasibility cluster randomised trials in the school setting. The protocol was publicly available prior to conducting the review. Screening, piloting of the data extraction form and data extraction were conducted by two independent reviewers. A pragmatic decision was made to limit the review to the UK in order to align with available resources and to make it more focused.

A limitation is the decision to use only the MEDLINE database. MEDLINE was chosen as health-based studies were the focus of this review. We acknowledge that further articles may have been found by searching other databases, grey literature and through citation searching. The search strategy was translated in EMBASE, DARE, PsycINFO and ERIC databases to search for additional eligible school-based CRTs published between 2017 and 2020 and resulted in identification of only one further unique eligible article. Therefore, we feel the pragmatic approach to only use MEDLINE to perform this search did not result in omission of a significant body of relevant evidence.

The systematic review only included feasibility studies that used the cluster randomised trial design and not other types, such as non-randomised parallel group and single-arm feasibility studies. We focussed on CRTs because we were interested in studies that could be used to assess a wide range of uncertainties for definitive CRTs, but we acknowledge that the systematic review may, therefore, not include some relevant knowledge of practice in non-randomised feasibility studies. While the approach used was not comprehensive, it enabled us to efficiently identify studies of interest that were undertaken in advance of planned definitive CRTs.

A further limitation of the review is that data were not extracted on consent procedures used by the included studies. As found in our previous review of definitive school-based CRTs [14], this information was inconsistently reported across studies making it challenging to summarise. This highlights the need for more comprehensive reporting of the consent procedures in these studies.

Conclusions

Cluster randomised feasibility studies are increasingly used in the school setting to test feasibility prior to definitive trials. Although these studies usually include few schools, the average sample size of those included in this review would be large enough to estimate percentages based on pupil characteristics that are used to address feasibility objectives (e.g., the percentage followed up) with a reasonable level of precision. The review has highlighted the need for clearer justification for the target sample size of school-based feasibility CRTs and to report the precision with which feasibility parameters are estimated in these studies. The characteristics of the recruited schools in feasibility CRTs could be better described to help understand the extent to which the feasibility parameter estimates are applicable to the planned definitive trial and other future similar trials. Finally, better use could be made of feasibility CRTs in the area of school-based pupil health research to assess challenges that are specific to the cluster trial design.

Abbreviations

BMI: Body mass index; CI: Confidence interval; CRT: Cluster randomised trial; DARE: Database of Abstracts of Reviews of Effects; EMBASE: Excerpta Medica Database; ERIC: Education Resources Information Center; ICC: Intra-cluster correlation coefficient; IQR: Interquartile range; MEDLINE: Medical Literature Analysis and Retrieval System Online; MeSH: Medical Subject Headings; MVPA: Moderate to vigorous physical activity; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PsycINFO: Psychological Information Database; SD: Standard deviation; UK: United Kingdom.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40814-022-01098-w>.

Additional file 1: Table S1. PRISMA checklist.

Acknowledgements

Kitty Parker and Obioha Ukoumunne were supported by the National Institute for Health Research Applied Research Collaboration South West Peninsula. Saskia Eddy received a Doctor of Philosophy (PhD) studentship from Barts Charity. SEd received training from the Medical Research Council (MRC)—National Institute of Health Research (NIHR) Trials Methodology Research Partnership (TMRP). Barts Charity, the MRC, the NIHR and the TMRP have no role in the study design, collection, management, analysis or interpretation of data, writing of the report or the decision to submit the report for publication. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Authors' contributions

KP, SEd, MN, ZMX, TF, SE and OU conceived the study. MN, ZMX, TF and SE advised on the design of the study and contributed to the protocol. KP, SEd and OU contributed to the design of the study, wrote the protocol and designed the data extraction form. KP, SEd and OU undertook data extraction. KP conducted the analyses of the data. All authors had full access to all the data. KP took primary responsibility for writing the manuscript. All authors provided feedback on all versions of the paper. The authors read and approved the final manuscript.

Funding

This research was funded by the National Institute for Health Research Applied Research Collaboration South West Peninsula. Saskia Eddy received a Doctor of Philosophy (PhD) studentship from Barts Charity.

Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available because they are also being used for a wider ongoing programme of research but are available from the corresponding author on reasonable request.

Declarations**Ethics approval and consent to participate**

Not applicable.

Consent for publication

Not applicable.

Competing interests

Sandra Eldridge is a member of the editorial board of the journal *Pilot and Feasibility Studies*.

Author details

¹NIHR Applied Research Collaboration South West Peninsula, University of Exeter, Room 2.16, South Cloisters, St Luke's Campus, 79 Heavitree Rd, Exeter EX1 2LLJ, UK. ²Wolfson Institute of Population Health, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK. ³College of Medicine and Health, University of Exeter, Exeter, UK. ⁴School of Health and Social Care, University of Essex, Colchester, UK.

⁵Department of Psychiatry, University of Cambridge, Cambridge, UK. ⁶NIHR Applied Research Collaboration South West Peninsula, University of Exeter, Exeter, UK.

Received: 20 October 2021 Accepted: 20 June 2022

Published online: 02 July 2022

References

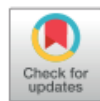
- Eldridge S, Kerry S. A practical guide to cluster randomised trials in health services research. Chichester: Wiley; 2012.
- Donner A, Klar N. Design and analysis of cluster randomization trials in health research. Chichester: Wiley; 2000.
- Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol*. 1978;108(2):100–2.
- Campbell M, Walters S. How to design, analyse and report cluster randomised trials in medicine and health related research. Chichester: Wiley; 2014.
- Hayes R, Moulton L. Cluster randomised trials. Florida: CRC Press; 2009.
- Murray D. Design and analysis of group-randomized trials. New York: Oxford University Press; 1998.
- Eldridge SM, Lancaster GA, Campbell MJ, Thabane L, Hopewell S, Coleman CL, et al. Defining feasibility and pilot studies in preparation for randomised controlled trials: development of a conceptual framework. *PLoS ONE*. 2016;11(3): e0150205.
- Lancaster GA, Thabane L. Guidelines for reporting non-randomised pilot and feasibility studies. *Pilot and Feasibility Studies*. 2019;5(1):1–14.
- Eldridge S, Kerry S, Torgerson DJ. Bias in identifying and recruiting participants in cluster randomised trials: what can be done? *BMJ*. 2009;339: b4006.
- Eldridge SM, Costelloe CE, Kahan BC, Lancaster GA, Kerry SM. How big should the pilot study for my cluster randomised trial be? *Stat Methods Med Res*. 2016;25(3):1039–56.
- Spybrook J, Zhang Q, Kelcey B, Dong N. Learning from cluster randomized trials in education: an assessment of the capacity of studies to determine what works, for whom, and under what conditions. *Educ Eval Policy Anal*. 2020;42(3):354–74.
- Wallester S, Hill SR, Bero LA. Characteristics and quality of reporting of cluster randomized trials in children: reporting needs improvement. *J Clin Epidemiol*. 2011;64(12):1331–40.
- Goesling B. A practical guide to cluster randomized trials in school health research. *J Sch Health*. 2019;89(11):916–25.
- Parker K, Nunns M, Xiao Z, Ford T, Ukoumunne OC. Characteristics and practices of school-based cluster randomised controlled trials for improving health outcomes in pupils in the United Kingdom: a methodological systematic review. *BMC Med Res Methodol*. 2021;21(1):152.
- Fazzari MJ, Kim MY, Heo M. Sample size determination for three-level randomized clinical trials with randomization at the first or second level. *J Biopharm Stat*. 2014;24(3):579–99.
- Billingham SA, Whitehead AL, Julious SA. An audit of sample sizes for pilot and feasibility trials being undertaken in the United Kingdom registered in the United Kingdom Clinical Research Network database. *BMC Med Res Methodol*. 2013;13(1):1–6.
- Thabane L, Ma J, Chu R, Cheng J, Ismail A, Rios LP, et al. A tutorial on pilot studies: the what, why and how. *BMC Med Res Methodol*. 2010;10(1):1–10.
- Chan CL, Leyrat C, Eldridge SM. Quality of reporting of pilot and feasibility cluster randomised trials: a systematic review. *BMJ Open*. 2017;7(11): e016970.
- Kristunas CA, Hemming K, Eborall H, Eldridge S, Gray LJ. The current use of feasibility studies in the assessment of feasibility for stepped-wedge cluster randomised trials: a systematic review. *BMC Med Res Methodol*. 2019;19(1):12.
- Moher D, Liberati A, Tetzlaff J, Altman DG, Gro P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009;6(7):e1000097.
- Taljaard M, McGowan J, Grimshaw JM, Brehaut JC, McRae A, Eccles MP, et al. Electronic search strategies to identify reports of cluster randomized trials in MEDLINE: low precision will improve with adherence to reporting standards. *BMC Med Res Methodol*. 2010;10(1):1–8.

22. Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. *J Eval Clin Pract.* 2004;10(2):307–12.
23. The EndNote Team. EndNote. EndNote X9 version ed. Philadelphia: Clarivate; 2013.
24. StataCorp. Release 17. College Station: StataCorp LLC; 2021.
25. Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. *BMJ.* 2012;345: e5661.
26. Thabane L, Hopewell S, Lancaster GA, Bond CM, Coleman CL, Campbell MJ, et al. Methods and processes for development of a CONSORT extension for reporting pilot randomized controlled trials. *Pilot Feasibility Stud.* 2016;2:25.
27. Barber SE, Jackson C, Hewitt C, Ainsworth HR, Buckley H, Akhtar S, et al. Assessing the feasibility of evaluating and delivering a physical activity intervention for pre-school children: a pilot randomised controlled trial. *Pilot Feasibility Stud.* 2016;2(1):12.
28. Bonell C, Fletcher A, Fitzgerald-Yau N, Hale D, Allen E, Elbourne D, et al. Initiating change locally in bullying and aggression through the school environment (INCLUSIVE): a pilot randomised controlled trial. *Health Technol Assess.* 2015;19(53).
29. Carlin A, Murphy MH, Nevill A, Gallagher AM. Effects of a peer-led Walking In Schools intervention (the WISH study) on physical activity levels of adolescent girls: a cluster randomised pilot study. *Trials.* 2018;19(1):31.
30. Cledes SA, Bingham DD, Pearson N, Chen Y-L, Edwardson CL, McEachan RRC, et al. Stand out in class: restructuring the classroom environment to reduce sitting time – findings from a pilot cluster randomised controlled trial. *Int J Behav Nutr Phys Act.* 2020;17(1):55.
31. Corder K, Brown HE, Schiff A, van Sluijs EMF. Feasibility study and pilot cluster-randomised controlled trial of the GoActive intervention aiming to promote physical activity among adolescents: outcomes and lessons learnt. *BMJ Open.* 2016;6(11): e012335.
32. Corepal R, Best P, O'Neill R, Kee F, Badham J, Dunne L, et al. A feasibility study of 'The StepSmart Challenge' to promote physical activity in adolescents. *Pilot Feasibility Stud.* 2019;5(1):132.
33. Forster AS, Cornelius V, Rockcliffe L, Marlow LA, Bedford H, Waller J. A cluster randomised feasibility study of an adolescent incentive intervention to increase uptake of HPV vaccination. *Br J Cancer.* 2017;117(8):1121–7.
34. Gammon C, Morton K, Atkin A, Corder K, Daly-Smith A, Quarmby T, et al. Introducing physically active lessons in UK secondary schools: feasibility study and pilot cluster-randomised controlled trial. *BMJ Open.* 2019;9(5): e025080.
35. Ginja S, Arnott B, Araujo-Soares V, Namdeo A, McCall E. Feasibility of an incentive scheme to promote active travel to school: a pilot cluster randomised trial. *Pilot and Feasibility Studies.* 2017;3(1):57.
36. Jago R, Sebire SJ, Cooper AR, Haase AM, Powell J, Davis L, et al. Bristol girls dance project feasibility trial: outcome and process evaluation results. *Int J Behav Nutr Phys Act.* 2012;9(1):83.
37. Jago R, Sebire SJ, Davies B, Wood L, Edwards MJ, Banfield K, et al. Randomised feasibility trial of a teaching assistant led extracurricular physical activity intervention for 9 to 11 year olds: Action 3:30. *Int J Behav Nutr Phys Act.* 2014;11:114.
38. Johnstone A, Hughes AR, Bonnar L, Booth JN, Reilly JJ. An active play intervention to improve physical activity and fundamental movement skills in children of low socio-economic status: feasibility cluster randomised controlled trial. *Pilot and Feasibility Studies.* 2019;5(1):45.
39. Kipping RR, Payne C, Lawlor DA. Randomised controlled trial adapting US school obesity prevention to England. *Arch Dis Child.* 2008;93(6):469–73.
40. Lloyd JJ, Wyatt KM, Creanor S. Behavioural and weight status outcomes from an exploratory trial of the Healthy Lifestyles Programme (HeLP): a novel school-based obesity prevention programme. *BMJ Open.* 2012;2(3): e000390.
41. Lohan M, Aventin A, Clarke M, Curran RM, McDowell C, Agus A, et al. Can Teenage men be targeted to prevent teenage pregnancy? A feasibility cluster randomised controlled intervention trial in schools. *Prev Sci.* 2018;19(8):1079–90.
42. McSweeney L, Araujo-Soares V, Rapley T, Adamson A. A feasibility study with process evaluation of a preschool intervention to improve child and family lifestyle behaviours. *BMC Public Health.* 2017;17(1):248.
43. Meiksin R, Crichton J, Dodd M, Morgan GS, Williams P, Willmott M, et al. A school intervention for 13- to 15-year-olds to prevent dating and relationship violence: the project respect pilot cluster RCT. *Public Health Res.* 2020;8(5).
44. Newbury-Birch D, Scott S, O'Donnell A, Coulton S, Howel D, McCall E, et al. A pilot feasibility cluster randomised controlled trial of screening and brief alcohol intervention to prevent hazardous drinking in young people aged 14–15 years in a high school setting (SIPS JR-HIGH). *Public Health Res.* 2014;2(6).
45. Sahota P, Christian M, Day R, Cocks K. The feasibility and acceptability of a primary school-based programme targeting diet and physical activity: the PhunkyFoods Programme. *Pilot Feasibility Stud.* 2019;5(1):152.
46. Sebire SJ, Jago R, Banfield K, Edwards MJ, Campbell R, Kipping R, et al. Results of a feasibility cluster randomised controlled trial of a peer-led school-based intervention to increase the physical activity of adolescent girls (PLAN-A). *Int J Behav Nutr Phys Act.* 2018;15(1):50.
47. Segrott J, Rothwell H, Hewitt G, Playle R, Huang C, Murphy S, et al. Preventing alcohol misuse in young people: an exploratory cluster randomised controlled trial of the Kids, Adults Together (KAT) programme. *Public Health Res.* 2015;3(15).
48. Sharpe H, Schober I, Treasure J, Schmidt U. Feasibility, acceptability and efficacy of a school-based prevention programme for eating disorders: cluster randomised controlled trial. *Br J Psychiatry.* 2013;203(6):428–35.
49. White J, Hawkins J, Madden K, Grant A, Er V, Angel L, et al. Adapting the ASSIST model of informal peer-led intervention delivery to the Talk to FRANK drug prevention programme in UK secondary schools (ASSIST + FRANK): intervention development, refinement and a pilot cluster randomised controlled trial. *Public Health Res.* 2017;5(7).
50. Wright B, Marshall D, Adamson J, Ainsworth H, Ali S, Allgar V, et al. Social Stories™ to alleviate challenging behaviour and social difficulties exhibited by children with autism spectrum disorder in mainstream schools: design of a manualised training toolkit and feasibility study for a cluster randomised controlled trial with nested qualitative and cost-effectiveness components. *Health Technol Assess.* 2016;20(6).
51. HM Government. Types of School. [Available from: <https://www.gov.uk/types-of-school>]. Accessed 01 Sept 2021.
52. Raab GM, Butcher I. Balance in cluster randomized trials. *Stat Med.* 2001;20(3):351–65.
53. Moulton LH. Covariate-based constrained randomization of group-randomized trials. *Clin Trials.* 2004;1(3):297–305.
54. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ.* 2008;337: a1655.
55. Guthold R, Stevens GA, Riley LM, Bull FC. Global trends in insufficient physical activity among adolescents: a pooled analysis of 298 population-based surveys with 1- 6 million participants. *Lancet Child Adolesc Health.* 2020;4(1):23–35.
56. Sadler K, Vizard T, Ford T, Marcheselli F, Pearce N, Mandalia D, et al. Mental health of children and young people in England, 2017. Leeds: NHS Digital; 2018.
57. Arain M, Campbell MJ, Cooper CL, Lancaster GA. What is a pilot or feasibility study? A review of current practice and editorial policy. *BMC Med Res Methodol.* 2010;10(1):67.
58. Aventin A, Lohan M, Maguire L, Clarke M. Recruiting faith- and non-faith-based schools, adolescents and parents to a cluster randomised sexual-health trial: experiences, challenges and lessons from the mixed-methods Jack Feasibility Trial. *Trials.* 2016;17(1):365.
59. Henderson M, Wight D, Nixon C, Hart G. Retaining young people in a longitudinal sexual health survey: a trial of strategies to maintain participation. *BMC Med Res Methodol.* 2010;10(1):9.
60. Pound B, Riddell M, Byrnes G, Kelly H. Perception of social value predicts participation in school-based research. *Aust N Z J Public Health.* 2000;24(5):543–5.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix 4 – Intra-cluster correlation coefficients from school-based cluster randomised trials of interventions for improving health outcomes on pupils



Journal of Clinical Epidemiology 158 (2023) 18–26

Journal of
Clinical
Epidemiology

ORIGINAL ARTICLE

Intracluster correlation coefficients from school-based cluster randomized trials of interventions for improving health outcomes in pupils

Kitty Parker^{a,*}, Michael Nunns^b, ZhiMin Xiao^c, Tamsin Ford^d, Obioha C. Ukoumunne^a

^aNIHR Applied Research Collaboration South West Peninsula, Department of Health and Community Sciences, Faculty of Health and Life Sciences, University of Exeter, Room 2.16, South Cloisters, St Luke's Campus, 79 Heavitree Rd, Exeter EX1 2LU, UK

^bFaculty of Health and Life Sciences, University of Exeter, St Luke's Campus, Heavitree Road, Exeter EX1 2LU, UK

^cSchool of Health and Social Care, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK

^dDepartment of Psychiatry, University of Cambridge, L5 Clifford Allbutt Building, Cambridge Biomedical Campus Box 58, Cambridge CB2 0AH, UK

Accepted 22 March 2023; Published online 28 March 2023

Abstract

Background and Objectives: To summarize intracluster correlation coefficient (ICC) estimates for pupil health outcomes from school-based cluster randomized trials (CRTs) across world regions and describe their relationship with study design characteristics and context.

Methods: School-based CRTs reporting ICCs for pupil health outcomes were identified through a literature search of MEDLINE (via Ovid). ICC estimates were summarized both overall and for different categories of study characteristics.

Results: Two hundred and forty-six articles reporting ICC estimates were identified. The median (interquartile range) ICC was 0.031 (0.011 to 0.08) at the school level ($N = 210$) and 0.063 (0.024 to 0.1) at the class level ($N = 46$). The distribution of ICCs at the school level was well described by the beta and exponential distributions. Besides larger ICCs in definitive trials than feasibility studies, there were no clear associations between study characteristics and ICC estimates.

Conclusion: The distribution of school-level ICCs worldwide was similar to previous summaries from studies in the United States. The description of the distribution of ICCs will help to inform sample size calculations and assess their sensitivity when designing future school-based CRTs of health interventions. © 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Keywords: Children; Cluster randomized trials; Intracluster correlation coefficient; Public health; Randomized trials; Schools

1. Background

Cluster randomized trials (CRTs) are studies in which clusters (groups) of individuals are randomized to trial arms and outcomes are measured on individuals [1]. CRTs are increasingly undertaken in schools to evaluate public health

interventions for improving outcomes of children and adolescents [2–5]. Schools provide a natural environment in which to recruit and study children and deliver interventions to improve their health due to the amount of time they spend there [3,6,7]. CRTs may be undertaken in schools because many of the interventions examined in such studies

Funding: This research was funded by the National Institute for Health and Care Research Applied Research Collaboration South West Peninsula.

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Conflict of interest: The authors of this manuscript have no conflicts of interest to declare.

Availability of data and materials: The datasets generated and/or analyzed during the current study are not publicly available because they are also being used for a wider ongoing program of research but are available from the corresponding author on reasonable request.

Competing interests: Not applicable.

Author Contributions: Kitty Parker: Conceptualisation; Methodology; Software; Formal analysis; Investigation; Writing - Original Draft;

Visualisation; Project administration. Michael Nunns: Conceptualisation; Methodology; Writing - Review & Editing; Supervision. ZhiMin Xiao: Conceptualisation; Writing - Review & Editing; Supervision. Tamsin Ford: Conceptualisation; Writing - Review & Editing; Supervision. Obioha Ukoumunne: Conceptualisation; Methodology; Software; Validation; Writing - Review & Editing; Supervision.

* Corresponding author. NIHR Applied Research Collaboration South West Peninsula, Department of Health and Community Sciences, Faculty of Health and Life Sciences, University of Exeter, Room 2.16, South Cloisters, St Luke's Campus, 79 Heavitree Rd, Exeter EX1 2LU, UK. Tel: +01392-727588; fax: +01392-723686.

E-mail address: kp477@exeter.ac.uk (K. Parker).

<https://doi.org/10.1016/j.jclinepi.2023.03.020>

0895-4356/© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

What is new?**Key findings**

- Few studies outside the United States (US) have summarized intraclass correlation coefficients (ICC) for pupil health outcomes and explored their size in relation to design characteristics in school-based cluster randomized trials (CRTs).

What this adds to what was known?

- This study collated 260 ICCs for school-related clusters from CRTs worldwide to inform sample size calculation for future trials.

What is the implication and what should change now?

- Two-thirds of school-level ICCs were no greater than 0.05 and three-quarters were under 0.08.
- The ICC distribution was similar to previous summaries from US-based studies and larger for definitive trials than feasibility studies.
- There was little evidence of relationships between ICC estimates and region, health outcome area, and educational level.

are designed to be delivered to entire schools or classrooms [4], interventions are theorized to affect change at those levels, and randomizing clusters (for example, schools, classes) helps to minimize contamination between trial arms that may otherwise occur if individuals are allocated [1,6,7].

CRTs require more participants than individually randomized trials because observations on individuals in the same cluster are usually more similar than those from different clusters [1]. Due to this lack of independence between individuals within clusters, if standard sample size formulae are used this may result in an underpowered study [1]. Correlation between pupils within clusters needs to be accounted for when designing and analyzing data from CRTs. In the sample size calculation, this is done by inflating the number of participants required in an individually randomized trial by the design effect (DE):

$$DE = 1 + (\bar{n} - 1)\rho$$

where \bar{n} is the mean number of participants providing outcome data in each cluster (cluster size) and ρ is the intraclass correlation coefficient (ICC) of the outcome [1]. The ICC quantifies the similarity of observations on individuals within clusters. For continuous outcomes, it can be defined as the proportion of the total variability in the outcome that is between clusters as opposed to between individuals within clusters:

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

where σ_b^2 is the between-cluster variance component and σ_w^2 is the within-cluster variance component [1]. Under this definition the ICC can take values between zero and one. The larger the ICC, the greater the sample size required. Similarity between participants from the same cluster can also be quantified by the between-cluster coefficient of variation (CV) of the outcome (the ratio of the between-cluster standard deviation to the outcome mean [6]):

$$CV = \frac{\sigma_b}{\mu}$$

where σ_b is the between-cluster standard deviation and μ is the mean outcome across the clusters [6]. The CV can then be incorporated into a modified design effect formula.

In the context of school-based CRTs there are several reasons for the similarity of outcomes between pupils within schools. First, in some countries, pupils and their parents/guardians have some influence regarding the school they attend [8]. Schools are likely to attract pupils with similar characteristics and who are more likely to share similar behaviours [3]. Second, pupils interact in the school setting and may influence the behaviour of their peers in the same schools or classrooms [8]. Finally, the school itself can influence the behaviours of pupils through its physical environment, ethos and policies [9,10].

At the time of sample size calculation the ICC is usually unknown and specification of a suitable value for the outcome and type of cluster should be informed by the empirical literature [1]. Researchers have reported ICCs for pupil health outcomes to be generally smaller than those for educational outcomes in schools [11–13]. This might be expected given that the main purpose of schools is to provide education [8]. Although ICCs for health outcomes in health care settings are well established, particularly in primary care [1,14,15], there is a relative lack of reported estimates in the school setting. Several studies have provided estimates of ICCs from school-based CRTs or surveys for outcomes related to substance use [8,16–24], nutrition [25–27], physical activity [24,27–29], and mental health and behaviour [12,24,30], but the vast majority of these were undertaken in the United States. It is not known whether these estimates are transferable to other regions and education systems, and outcome areas such as infectious diseases and dental health are not well represented. Furthermore, although patterns in the size of the ICC have been investigated [14,15,31–34], little is known about the extent to which ICCs from school-based CRTs differ by study characteristics.

A summary of ICCs for a range of health outcomes in different settings would aid the design of future school-based CRTs by providing plausible values that can be used

in sample size calculations. Estimates from CRTs specifically, rather than surveys, are potentially especially relevant as they may better reflect the level of variation in outcomes across the types of schools that tend to participate in health-related trials [1] (p177).

1.1. Objectives

This paper collates and summarizes ICC estimates for health outcomes from school-based CRTs and examines the relationship between the size of the ICC and study characteristics.

2. Methods

2.1. Data sources and search methods

A systematic searching approach was used to identify papers reporting ICC estimates from school-based CRTs. MEDLINE (Ovid) was exclusively searched for published peer-reviewed articles reporting school-based CRTs from inception to 18th October 2021. The search strategy was developed based on a strategy by Taljaard and colleagues [35] used to identify CRTs, combined with school-related terms (Table 1).

2.2. Inclusion and exclusion criteria

Eligible articles reported school-based studies with a CRT design, including articles reporting baseline data, follow-up outcomes, or secondary data analyses that used the data to address additional questions that were unrelated to the main trial objectives. To be eligible, the article had to report the estimate of an ICC/CV for at least one health outcome measured on pupils. The eligible study population was pupils attending pre-primary, primary, lower secondary

and higher secondary educational settings according to the United Nations Educational, Scientific and Cultural Organisation (UNESCO) International Standard Classification of Education (ISCED) system [36]. Eligible clusters were any school-related unit (e.g., schools, classes/classrooms, year groups, teachers). Any intervention(s) were considered. Articles were excluded if they randomised after-school clubs, school-based health centres or childcare centres. Articles that only reported protocol/design information, process evaluations, economic evaluations/cost-effectiveness analyses, statistical analysis plans, commentaries and mediation/mechanism analyses were also excluded.

If more than one publication of the same eligible study was identified, the key study report (index paper) for data extraction was determined by identifying the article that first published the outcomes.

2.3. Sifting and validation

Titles and abstracts were screened by two independent researchers (KP & OU) for eligibility against the inclusion criteria. Any studies for which the reviewers were uncertain of inclusion status were progressed to full text screening. Two independent researchers (KP & OU) examined the full text of each article against the inclusion criteria. Any disagreements over inclusion were resolved through discussion with a third researcher (MN).

2.4. Data extraction

One researcher (KP) extracted data from all included articles, while a second (OU) independently validated the process. Any uncertainty regarding the data extraction was resolved through discussion, or consultation with a third researcher (MN). The information extracted is specified in Table 2.

The ICC/CV estimate(s) of one pupil health outcome was extracted from each article, as estimates for multiple outcomes from the same study would likely be correlated and contribute relatively little additional information to the analyses in this paper which are focussed on comparing the ICC/CV across different study scenarios. Where estimates were reported for the chosen outcome at multiple levels (for example, school and class) these were all extracted. The criteria used to select the ICC/CV when multiple estimates were reported for a given paper are presented in Table 3. Where studies reported both unadjusted and adjusted ICCs, the former was extracted on the basis that this would be of more general use to future researchers who may want to adjust their estimate of the intervention effect for a specific set of prognostic factors. Where the ICC for a given outcome was reported for multiple time points the ICC for the earliest wave was extracted, as the ICC estimate would be less likely to be impacted by the intervention. For a similar reason, where

Table 1. Search strategy using MEDLINE (through Ovid)

Search strategy
Terms for Randomized Controlled trials:
1. random:.mp.
2. trial.ab, kw, ti.
Cluster design-related terms:
3. "cluster*".ab, kw, ti.
4. "communit*".ab, kw, ti.
5. group*adj2 random*.ab, kw, ti.
6. 3 OR 4 OR 5
School terms:
7. exp Schools/
8. School*.ab, kw, ti.
9. 7 OR 8
Final search stages:
10. 1 AND 2 AND 6 AND 9
11. 10 limited to English language

Table 2. Data extracted

Aspect	Information extracted
Publication details	Author surname, year of publication, title of article, type of study (that is, definitive or feasibility study).
Setting information	Country in which the study took place (for example, France), stage of education (for example, primary, secondary), gender of pupils, age(s) of pupils at baseline.
Study design	Type of cluster unit allocated, cluster unit of ICC/CV estimate.
Sample size information	ICC/CV assumed in the sample size calculation, number of clusters and pupils that provided outcome data, number of classes per school.
Health outcome information	Health area of outcome (for example, physical activity), outcome description (for example, amount of moderate-to-vigorous physical activity), outcome type (for example, continuous, binary), timing (months postrandomization) at which outcome was measured.
ICC information	ICC/CV of the outcome (and 95% CIs where provided), analytical method used to calculate ICC/CV (for example, multilevel model [37], marginal model using Generalized Estimating Equations [38]), whether the ICC/CV estimate was pooled across trial arms, whether the ICC/CV estimate was unadjusted or adjusted for prognostic factors, whether the ICC/CV estimate was adjusted for the baseline value of the outcome, whether the ICC/CV was estimated from an analysis of change scores between baseline and follow-up, whether a repeated measures analysis was used to estimate the ICC.

the ICC was reported separately for the control and intervention arms the former was chosen.

2.5. Data analysis

Study characteristics were summarized using medians, interquartile ranges (IQRs), and ranges for continuous variables and numbers and percentages for categorical variables. Mann–Whitney and Kruskal–Wallis tests were used to compare the ICC estimates across subgroups. Analyses were undertaken using Stata 17 [39].

3. Results

3.1. Search results

Three thousand six hundred and thirty-two articles were identified through searching MEDLINE. One thousand five hundred and ninety articles were included in the full text screening stage and 246 articles were identified as eligible for inclusion in the review. One paper reported an estimate of the between-cluster coefficient of variation of the outcome, but this was negative and therefore the paper

was not included. The PRISMA flow diagram is presented in Figure 1.

3.2. Publication characteristics

Worldwide, the rate of publication of articles reporting ICC estimates from school-based CRTs that evaluate interventions for improving pupil health outcomes has increased since the first publication in 1999; 44 articles were published between 1999 and 2010, compared to 25 in 2021 alone. Of the 246 included studies, 226 (91.9%) were definitive trials and 20 (8.1%) were feasibility studies. The settings of included studies spanned all regions of the world and different stages of education. The majority of studies ($n = 227$; 92.3%) included males and females. In most of the studies schools were the units of randomization ($n = 220$; 89.4%); classes were randomized in 23 (9.3%) studies; and school buildings [40], student groups [41] and year groups [42] were randomized in one study each. The studies spanned a range of different health outcome areas, the most common being socioemotional functioning and its influences ($n = 53$; 21.5%), physical activity

Table 3. Criteria used to select which ICC/CV to extract

Aspect	Criteria
Outcome measure	In the first instance, the ICC/CV for the primary health outcome was selected. If there was more than one primary health outcome, the ICC/CV for the first primary outcome presented in the Results section of the paper was selected. If no primary health outcome was declared, the ICC/CV for the health outcome on which the sample size calculation was based was selected. If no primary health outcome was declared and the sample size was not based on a health outcome, the ICC/CV for the first health outcome reported in the Results section of the paper was selected.
Time point at which outcome was measured	In the first instance, the ICC/CV from the baseline time point was selected. If this was not reported, the ICC/CV from the earliest time point of measurement was selected.
Unadjusted vs. adjusted ICC/CV	If the study presented both unadjusted ICCs/CVs estimates and estimates that are adjusted for prognostic factors, the unadjusted ICC/CV was extracted.
Control versus intervention arm	If the ICC/CV was reported separately for the intervention and control arms, the ICC/CV from the control arm was selected.

Table 4. Median (IQR; range) school-level ICC by region, outcome area and education stage

Characteristic	N	Median ICC (IQR; range)	P value
Region			0.26
Europe ^a	45	0.04 (0.014 to 0.08; 0 to 0.47)	
USA and Canada	44	0.033 (0.010 to 0.073; 0 to 0.286)	
UK ^b	40	0.029 (0.01 to 0.106; 0 to 0.45)	
Australia and New Zealand	27	0.02 (0.01 to 0.03; 0 to 0.16)	
Asia ^c	21	0.05 (0.013 to 0.118; 0 to 0.31)	
Central and South America ^d	17	0.05 (0.016 to 0.09; 0.0001 to 0.36)	
Africa ^e	16	0.05 (0.018 to 0.127; 0.0005 to 0.21)	
Health outcome area			0.76
Socioemotional functioning and its influences ^f	39	0.05 (0.02 to 0.097; 0 to 0.217)	
Physical activity	30	0.035 (0.013 to 0.059; 0 to 0.19)	
Adiposity	26	0.027 (0.014 to 0.041; 0.004 to 0.19)	
Smoking	19	0.055 (0.017 to 0.11; 0 to 0.286)	
Alcohol use	10	0.055 (0.02 to 0.098; 0 to 0.121)	
Dental/oral health	10	0.051 (0.027 to 0.119; 0 to 0.31)	
General health	10	0.025 (0.014 to 0.045; 0.001 to 0.18)	
Infectious disease	9	0.042 (0.004 to 0.070; 0.0001 to 0.21)	
Nutrition	8	0.06 (0.010 to 0.097; 0 to 0.36)	
Violence	8	0.048 (0.014 to 0.085; 0.002 to 0.13)	
Education stage			0.40
Preprimary education only ^g	13	0.048 (0.03 to 0.063; 0 to 0.097)	
Primary education only ^h	81	0.04 (0.013 to 0.094; 0 to 0.47)	
Secondary education only ⁱ	81	0.03 (0.01 to 0.07; 0 to 0.31)	

^a Included countries stated as follows: Finland, The Netherlands, Denmark, Belgium, Norway, Germany, Estonia, Poland, Spain, Switzerland, Cyprus, Italy, Greece, Hungary, Sweden, Austria, Majorca, France, Ireland, Romania, Slovenia.

^b Included countries stated as follows: England, Northern Ireland, Scotland, Wales.

^c Included countries stated as follows: Israel, China, Iran, India, Japan, Bangladesh, Nepal, Taiwan, Peru, Pakistan, Thailand, Indonesia, Hong Kong.

^d Included countries stated as follows: Jamaica, Brazil, Ecuador, Chile, Haiti, Belize.

^e Included countries stated as follows: Uganda, South Africa, Kenya, Tanzania, Burundi.

^f Includes mental health, behaviour, neurodiversity, wellbeing, quality of life, bullying, social and emotional learning, body image, and self-esteem.

^g Includes preschools, kindergartens, educational childcare centres, and head-start schools.

^h Includes elementary schools, middle schools (Grade 6).

ⁱ Includes secondary schools, middle schools (\geq Grade 7), high schools, junior high schools, lower secondary schools, higher/upper secondary schools, vocational schools, intermediate vocational schools, secondary-level vocational schools, and continuation schools.

(0.016 to 0.07); $P = 0.50$). Also, for continuous outcomes, the median school-level ICC was identical for studies that did ($N = 11$) and did not ($N = 124$) analyze change scores (0.04; $P = 0.37$). The median (IQR) school-level ICC was lower for studies that estimated the ICC from a repeated measures analysis ($N = 37$) compared with those that did not ($N = 173$) (0.027 (0.01 to 0.057) vs. 0.036 (0.013 to 0.088)), but with little evidence of a systematic difference ($P = 0.15$). Finally, for binary outcomes, there was weak evidence that the median (IQR) ICC was higher for studies that use multilevel logistic regression to estimate this parameter on the logistic scale ($N = 42$) than those that use other methods to estimate it on the proportions (natural) scale ($N = 14$) (0.049 (0.014 to 0.109) vs. 0.014 (0.007 to 0.023); $P = 0.08$). The direction of this difference is consistent with the fact that the ICC on the logistic scale is generally larger than on the proportions scale [44]. [Appendix 4](#)

summarises the relationship between ICC estimates and the prevalence for binary outcomes.

4. Discussion

To our knowledge, this is the first paper to report the distribution of ICCs for pupil health outcomes from school-based CRTs worldwide. 260 ICC estimates from 246 school-based CRTs were extracted for outcomes spanning a range of health areas. There were few clear patterns regarding the relationship of the ICC with aspects of the design and analysis. Indeed, comparison of the ICC across categories of the study features examined was characterized by overlap in the distributions, although the differences in medians would be large in terms of the impact they would have on the sample size requirement for a CRT. Imprecision

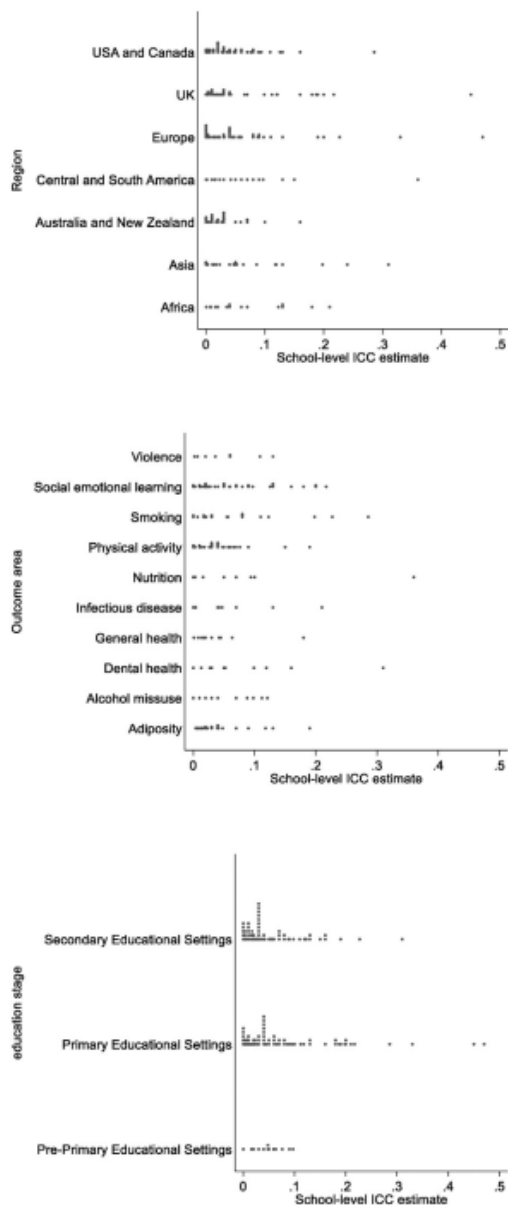


Fig. 3. Dot plots of school-level ICCs by region, outcome area, and education stage.

in the ICC estimates may have reduced the power to detect differences between subgroups defined by design and analysis characteristics.

The large number of different outcomes represented (Appendix 1) partly accounts for the variation in the estimates, although there was even a marked variation in ICC estimates across studies for the same outcome (that is, amount of MVPA and BMI). Sampling variability, the methodological context of the trials and the models

specified to estimate the parameter will also contribute to variability in the ICC estimates. Given the clinical and methodological heterogeneity across CRTs, an individual ICC estimate for a given outcome from a single study may have poor generalizability [28], and it has been recommended that researchers use the distribution of ICCs from many studies to model the sensitivity of sample size calculations [1,14,34]. Distributions of ICC estimates for health outcomes in primary care-based clusters have been found to be well described by the beta distribution [14,45]. The beta distribution was a good fit to the school-level ICCs reported in this paper as was the exponential distribution. The distribution parameters of these ICC estimates are of value for constructing informative priors when using a Bayesian framework to incorporate uncertainty about the ICC in sample size calculations for school-based CRTs [46,47].

There was little difference between ICC estimates that were adjusted for the baseline outcome measurement and those that were not. This may be due to differences in aspects of the design and setting across the studies and the fact that adjustment for individual-level prognostic factors may increase or decrease the ICC depending on the extent to which the between-cluster and within-cluster components of variance are reduced following adjustment [48].

The ICC from a repeated measures analysis using outcome data from across all study waves does not necessarily estimate the same parameter as an ICC for the outcome at a specific study wave. The correlation between observations from the same cluster from different waves may be smaller than the correlation between observations from the same cluster at the same study wave [34,49]. In this study, however, there was little evidence that the school-level ICC is lower for studies that estimate the ICC from a repeated measures analysis than those that do not, although the median was lower for the former set of studies.

Previously reported summaries of school-based ICCs for pupil health outcomes have largely used data from trials and surveys in the United States [12,16–22,25–30]. The distribution of school-level ICCs worldwide in the current paper was broadly similar to those previous summaries, with most estimates less than 0.05 and few greater than 0.1. Only the distribution for the Australia/New Zealand region was notably different (smaller).

The median ICC for pupil health outcomes was 0.031 at the school level and 0.063 at the class level. The difference is intuitive given the greater opportunity for interaction within classes as opposed to between classes within the same school and that the ICC has been reported to be larger when the natural cluster size is smaller [20,50]. The median ICC was markedly smaller for feasibility studies than in definitive trials. This may reflect that schools recruited in feasibility studies are a more restricted and less representative subset of the wider types of schools that are recruited in larger definitive studies [1] (p180/181). There was little evidence of a relationship between the ICC for pupil health

outcomes and stage of education. Previously, it has been reported that there is a tendency for ICCs for educational outcomes to be larger for lower education grades [48].

4.1. Strengths and limitations

This is the first study to collate and summarize ICCs for pupil outcomes across different health areas from school-based CRTs worldwide. The study used a systematic searching approach with dual screening and data validation. The sample of 246 CRTs was not sufficiently large to describe the ICC within different combinations of categories of the study design parameters (for example, only one combination of region and health outcome area provided at least 10 school-level ICC estimates). Partly for this reason, when investigating geographic variation in the ICC, we grouped countries into regions which will have obscured differences between individual countries. Based on empirical evidence from a European-based survey, it has been suggested that the ICCs assumed in the sample size calculation for school-based trials should be country-specific and outcome-specific [8]. As more school-based CRTs are undertaken the pool of reported ICCs will increase, enabling a more detailed examination with greater power to detect ICC patterns in relation to key study characteristics.

A potential limitation was the decision to use only the MEDLINE database. Although findings from a previous systematic review of similar studies indicated that few additional studies would have been found by searching other databases (specifically, EMBASE, DARE, PsychINFO, and ERIC) [4], we acknowledge that further articles may have been found by searching the grey literature. Additionally, some older articles may have been missed because the titles and abstracts did not refer to using a cluster design.

It was decided to extract the ICC estimate for only one outcome from each study even when multiple ones were reported. We anticipated that ICCs would be more similar within studies and wanted to avoid a scenario where a small number of studies that reported many ICCs had a disproportionate impact on the observed distribution of ICCs.

5. Conclusions

The 260 reported ICC estimates from studies spanning all world regions and different health outcome areas, and the summaries of their distribution are a valuable resource to researchers for calculating sample size for future school-based CRTs. The ICCs had a similar distribution to published summaries of the parameter from studies based in the United States. Better reporting of the ICC in CRTs, in keeping with CONSORT guidance [51], will provide a larger pool of data that can be used to explore the distribution of ICC values and the factors that determine them in greater detail.

Acknowledgments

Kitty Parker and Obioha Ukoumunne were supported by the National Institute for Health and Care Research (NIHR) Applied Research Collaboration South West Peninsula. The NIHR had no role in the study design, collection, management, analysis or interpretation of data, writing of the report, or the decision to submit the report for publication. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jclinepi.2023.03.020>.

References

- [1] Eldridge S, Kerry S. A practical guide to cluster randomised trials in health services research. Chichester: John Wiley & Sons; 2012.
- [2] Walleiser S, Hill SR, Bero LA. Characteristics and quality of reporting of cluster randomized trials in children: reporting needs improvement. *J Clin Epidemiol* 2011;64:1331–40.
- [3] Goesling B. A practical guide to cluster randomized trials in school health research. *J Sch Health* 2019;89(11):916–25.
- [4] Parker K, Nunns M, Xiao Z, Ford T, Ukoumunne OC. Characteristics and practices of school-based cluster randomised controlled trials for improving health outcomes in pupils in the United Kingdom: a methodological systematic review. *BMC Med Res Methodol* 2021;21:152.
- [5] Parker K, Eddy S, Nunns M, Xiao Z, Ford T, Eldridge S, et al. Systematic review of the characteristics of school-based feasibility cluster randomised trials of interventions for improving the health of pupils in the UK. *Pilot Feasibility Stud* 2022;8(1):132.
- [6] Hayes R, Moulton L. Cluster Randomised Trials. Florida: CRC Press; 2009.
- [7] Murray DM. Design and Analysis of Group-Randomized Trials. New York: Oxford University Press; 1998.
- [8] Shackleton N, Hale D, Bonell C, Viner RM. Intraclass correlation values for adolescent health outcomes in secondary schools in 21 European countries. *SSM Popul Health* 2016;2:217–25.
- [9] Bonell C, Parry W, Wells H, Jamal F, Fletcher A, Harden A, et al. The effects of the school environment on student health: a systematic review of multi-level studies. *Health Place* 2013;21:180–91.
- [10] Langford R, Bonell CP, Jones HE, Poulou T, Murphy SM, Waters E, et al. The WHO Health Promoting School framework for improving the health and well-being of students and their academic achievement. *Cochrane Database Syst Rev* 2014;(4):CD008958.
- [11] Bonell C, Jamal F, Harden A, Wells H, Parry W, Fletcher A. Systematic review of the effects of schools and school environment interventions on health: evidence mapping and synthesis. *Public Health Res* 2013;1(1).
- [12] Dong N, Reinke WM, Herman KC, Bradshaw CP, Murray DW. Meaningful effect sizes, intraclass correlations, and proportions of variance explained by covariates for planning two- and three-level cluster randomized trials of social and behavioral outcomes. *Eval Rev* 2016;40(4):334–77.
- [13] Hale DR, Fitzgerald-Yau N, Viner RM. A systematic review of effective interventions for reducing multiple health risk behaviors in adolescence. *Am J Public Health* 2014;104:e19–41.
- [14] Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care

- research to inform study design and analysis. *J Clin Epidemiol* 2004; 57:785–94.
- [15] Stuart B, Becque T, Moore M, Little P. Clustering of continuous and binary outcomes at the general practice level in individually randomised studies in primary care - a review of 10 years of primary care trials. *BMC Med Res Methodol* 2020;20:83.
- [16] Murray DM, Hannan PJ. Planning for the appropriate analysis in school-based drug-use prevention studies. *J Consult Clin Psychol* 1990;58(4):458–68.
- [17] Murray DM, Rooney BL, Hannan PJ, Peterson AV, Ary DV, Biglan A, et al. Intraclass correlation among common measures of adolescent smoking: estimates, correlates, and applications in smoking prevention studies. *Am J Epidemiol* 1994;140:1038–50.
- [18] Murray DM, Short B. Intraclass correlation among measures related to alcohol use by young adults: estimates, correlates and applications in intervention studies. *J Stud Alcohol* 1995;56(6):681–94.
- [19] Murray DM, Short BJ. Intraclass correlation among measures related to tobacco use by adolescents: estimates, correlates, and applications in intervention studies. *Addict Behav* 1997;22(1):1–12.
- [20] Siddiqui O, Hedeker D, Flay BR, Hu FB. Intraclass correlation estimates in a school-based smoking prevention study: outcome and mediating variables, by sex and ethnicity. *Am J Epidemiol* 1996; 144:425–33.
- [21] Murray DM, Clark M, Wagenaar AC. Intraclass correlations from a community-based alcohol prevention study: the effect of repeat observations on the same communities. *J Stud Alcohol* 2000;61(6):881–90.
- [22] Ennett ST, Flewelling RL, Lindrooth RC, Norton EC. School and neighborhood characteristics associated with school rates of alcohol, cigarette, and marijuana use. *J Health Soc Behav* 1997;38(1):55–71.
- [23] Resnicow K, Zhang N, Vaughan RD, Reddy SP, James S, Murray DM. When intraclass correlation coefficients go awry: a case study from a school-based smoking prevention study in South Africa. *Am J Public Health* 2010;100:1714–8.
- [24] Sellström E, Bremberg S. Is there a “school effect” on pupil outcomes? A review of multilevel studies. *J Epidemiol Community Health* 2006;60:149–55.
- [25] Murray DM, Phillips GA, Birnbaum AS, Lytle LA. Intraclass correlation for measures from a middle school nutrition intervention study: estimates, correlates, and applications. *Health Educ Behav* 2001; 28(6):666–79.
- [26] Juras R. Estimates of intraclass correlation coefficients and other design parameters for studies of school-based nutritional interventions. *Eval Rev* 2016;40(4):314–33.
- [27] Gray HL, Burgermaster M, Tipton E, Contento IR, Koch PA, Di Noia J. Intraclass correlation coefficients for obesity indicators and energy balance-related behaviors among New York city public elementary schools. *Health Educ Behav* 2016;43(2):172–81.
- [28] Murray DM, Catellier DJ, Hannan PJ, Treuth MS, Stevens J, Schmitz KH, et al. School-level intraclass correlation for physical activity in adolescent girls. *Med Sci Sports Exerc* 2004;36(5):876–82.
- [29] Murray DM, Stevens J, Hannan PJ, Catellier DJ, Schmitz KH, Dowda M, et al. School-level intraclass correlation for physical activity in sixth grade girls. *Med Sci Sports Exerc* 2006;38(5):926–36.
- [30] Hedberg EC. Academic and behavioral design parameters for cluster randomized trials in kindergarten: an analysis of the early childhood longitudinal study 2011 kindergarten cohort (ECLS-K 2011). *Eval Rev* 2016;40(4):279–313.
- [31] Murray DM, Blitstein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. *Eval Rev* 2003;27(1):79–103.
- [32] Gulliford MC, Adams G, Ukoumunne OC, Latinovic R, Chinn S, Campbell MJ. Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. *J Clin Epidemiol* 2005;58:246–51.
- [33] Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intra-cluster correlation coefficient in cluster randomized trials: the case of implementation research. *Clin Trials* 2005;2:99–107.
- [34] Korevaar E, Kasza J, Taljaard M, Hemming K, Haines T, Turner EL, et al. Intra-cluster correlations from the CLustered OUtcome Dataset bank to inform the design of longitudinal cluster trials. *Clin Trials* 2021;18:529–40.
- [35] Taljaard M, McGowan J, Grimshaw JM, Brehaut JC, McRae A, Eccles MP, et al. Electronic search strategies to identify reports of cluster randomized trials in MEDLINE: low precision will improve with adherence to reporting standards. *BMC Med Res Methodol* 2010;10:15.
- [36] UNESCO. International Standard Classification of Education: ISCED 2011. Montreal: UIS; 2012:85.
- [37] Goldstein H, Browne W, Rasbash J. Multilevel modelling of medical data. *Stat Med* 2002;21:3291–315.
- [38] Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986;73(1):13–22.
- [39] StataCorp. Stata. Release 17. College Station, TX: StataCorp LLC; 2021.
- [40] Lopata C, Thomeer ML, Rodgers JD, Donnelly JP, McDonald CA, Volker MA, et al. Cluster randomized trial of a school intervention for children with autism spectrum disorder. *J Clin Child Adolesc Psychol* 2019;48(6):922–33.
- [41] Nykänen M, Sund R, Vuori J. Enhancing safety competencies of young adults: a randomized field trial (RCT). *J Safety Res* 2018; 67:45–56.
- [42] Stallard P, Sayal K, Phillips R, Taylor JA, Spears M, Anderson R, et al. Classroom based cognitive behavioural therapy in reducing symptoms of depression in high risk adolescents: pragmatic cluster randomised controlled trial. *Br Med J* 2012;345:e6058.
- [43] Emery CA, Rose MS, McAllister JR, Meeuwisse WH. A prevention strategy to reduce the incidence of injury in high school basketball: a cluster randomized controlled trial. *Clin J Sport Med* 2007;17: 17–24.
- [44] Eldridge SM, Ukoumunne OC, Carlin JB. The intra-cluster correlation coefficient in cluster randomized trials: a review of definitions. *Int Stat Rev* 2009;77(3):378–94.
- [45] Turner RM, Omar RZ, Thompson SG. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Stat Med* 2001;20:453–72.
- [46] Turner RM, Thompson SG, Spiegelhalter DJ. Prior distributions for the intraclass correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clin Trials* 2005;2:108–18.
- [47] Jones BG, Streeter AJ, Baker A, Moyeed R, Creanor S. Bayesian statistics in the design and analysis of cluster randomised controlled trials and their reporting quality: a methodological systematic review. *Syst Rev* 2021;10(1):91.
- [48] Hedges LV, Hedberg EC. Intraclass correlation values for planning group-randomized trials in education. *Educ Eval Policy Anal* 2007; 29(1):60–87.
- [49] Li F, Hughes JP, Hemming K, Taljaard M, Melnick ER, Heagerty PJ. Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: an overview. *Stat Methods Med Res* 2021;30(2):612–39.
- [50] Gulliford MC, Ukoumunne OC, Chinn S. Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: data from the Health Survey for England 1994. *Am J Epidemiol* 1999;149:876–83.
- [51] Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. *Br Med J* 2012;345: e5661.

Appendix 5 - Chapter 3: A table comparing school year groups across nations in the UK [135]

UK Comparison Table of School Year Groups across the UK (April 2020)

	England		Northern Ireland		Wales		Scotland		Boarding	
	National Curriculum		Northern Ireland Curriculum		National Curriculum Wales		Curriculum for Excellence			
Born between	1 Sept – 31 Aug		1 Sept – 1 July	2 Jul – 31 Aug	1 Sept – 31 Aug		1 Sept – 29 Feb	1 Mar – 31 Aug	1 Sept – 31 Aug Prep & Public	1 Sept-31 Aug State Boarding
Age										
4-5	EYFS	Reception	P1	Nursery	Foundation Phase	Reception	P1 (Early level)	Nursery/Early level	None	None
5-6	KS1	Yr 1	P2	P1		Yr 1	P2(First level)	P1 (Early level)	None	None
6-7	KS1	Yr 2	P3	P2		Yr 2	P3 (First level)	P2 (First level)	None	None
7-8	KS2	Yr 3	P4	P3	Key Stage 2	Yr 3	P4 (First level)	P3 (First level)	Prep	Yr 3
8-9	KS2	Yr 4	P5	P4		Yr 4	P5 (Second level)	P4 (First level)	Prep	Yr 4
9-10	KS2	Yr 5	P6	P5		Yr 5	P6 (Second level)	P5 (Second level)	Prep	Yr 5
10-11	KS2	Yr 6	P7	P6		Yr 6	P7 (Second level)	P6 (Second level)	Prep	Yr 6
11-12	KS3	Yr 7	Yr 8	P7	Key Stage 3	Yr 7	S1 (Third/Fourth level)	P7 (Second Level)	Prep	Secondary Yr 7
12-13	KS3	Yr 8	Yr 9	Yr 8		Yr 8	S2 (Third/Fourth level)	S1 (Third/Fourth level)	Prep CE	Yr 8
13-14	KS3	Yr 9	Yr 10	Yr 9		Yr 9	S3 (Third/Fourth level)	S2 (Third/Fourth level)	Public	Yr 9
14-15	KS4	Yr 10	Yr 11	Yr 10	Key Stage 4	Yr 10	S4 (Senior phase)	S3 (Third/Fourth level)	Public	Yr 10
15-16	KS4	Yr 11	Yr 12	Yr 11		Yr 11	S5 (Senior phase)	S4 (Senior phase)	Public	Yr 11
<i>A Levels and SCE Highers – non-compulsory</i>										
16-17	AS	Yr 12	Yr13 Sixth form	Yr 12	Post 16	Yr 12	S6 (Senior phase)	S5 (senior phase)	Public	Yr12
17-18	A2	Yr 13	Yr14 Sixth form	Sixth form		Yr 13		S6 (Senior phase)	Public	Yr 13
18				Sixth form						

Yr	Year
EYFS	Early Years Foundation Stage
FS	Foundation stage
KS	Key Stage
Early	Early Years
First	First level
Second	Second level
P	Primary
S	Secondary
CE	Common Entrance
AS	1 st year A Level exams
A2	A level exams

Appendix 6 - Chapter 3: Further details on journals and funding sources (N=64)

Studies were published in journal including: *British Medical Journal* (n=9; 14%); *BMC Public Health* (n=4; 6%); *International Journal of Behavioural Nutrition and Physical Activity* (n=4; 6%); *Archives of Disease in Childhood* (n=3; 5%); *BMJ Open* (n=3; 5%); *Journal of Epidemiology and Community Health* (n=3; 5%); *Public Health Nutrition* (n=3; 5%); *The Lancet* (n=3; 5%); *Child: Care, Health and Development* (n=2; 3%); *Journal of Dental Research* (n=2; 3%); *Journal of Public Health Dentistry* (n=2; 3%); *BMC Oral Health* (n=1; 2%); *BMC Research Notes* (n=1; 2%); *Behaviour Research & Therapy* (n=1; 2%); *British Journal of Health Psychology* (n=1; 2%); *British Journal of Sports Medicine* (n=1; 2%); *Caries Research* (n=1; 2%); *European Child and Adolescent Psychiatry* (n=1; 2%); *Health Education and Behaviour* (n=1; 2%); *Health and Quality of Life outcomes* (n=1; 2%); *Injury Prevention* (n=1; 2%); *International Dental Journal* (n=1; 2%), *International Journal of Obesity* (n=1; 2%); *JAMA Psychiatry* (n=1; 2%); *Journal of Child Psychology* (n=1; 2%); *Journal of Child Psychology & Psychiatry* (n=1; 2%); *Journal of Consulting & Clinical psychology* (n=1; 2%), *Journal of Nutrition Education and Behaviour* (n=1; 2%); *Journal of School Psychology* (n=1; 2%); *Journal of Youth & Adolescence* (n=1; 2%); *Lancet Child & Adolescent Health* (n=1; 2%); *Lancet Psychiatry* (n=1; 2%); *Perceptual & Motor Skills* (n=1; 2%); *Prevention Science* (n=1; 2%); *Psycho-Oncology* (n=1; 2%); *Psychological Medicine* (n=1; 2%); *Public Health Nutrition* (n=1; 2%); *Public Health Research – NIHR* (n=1; 2%).

Funding sources were: *NIHR Public Health Research programme* (n=11; 17%); *Medical Research Council* (n=6; 9%); *Department of Health* (n=3; 5%); *Food Standards Agency* (n=3); *Economic and Social Research Council* (n=2; 3%); *Education Endowment Foundation* (n=2; 3%); *NIHR Health Technology Assessment Programme* (n=2; 3%); *Unilever* (n=2; 3%); *Action on Addiction* (n=1; 2%); *Bangor University* (n=1; 2%); *Big Lottery Wales* (n=1; 2%); *Birmingham City Council* (n=1; 2%); *Bournemouth Diabetes and Endocrine Centre* (n=1; 2%); *Broxtowe and Hucknall Primary Care Trust Injury Prevention Research Programme* (n=1; 2%); *Camden and Islington Health Authority* (n=1; 2%); *Cancer Research UK* (n=1; 2%); *Coca Cola Foundation* (n=1; 2%); *Department of Child Health, Queen's University Belfast* (n=1; 2%); *Department of Education* (n=1; 2%); *Eastern Health and Social Services Board* (n=1; 2%);

European Commission (n=1; 2%); GlaxoSmithKline, Aventis, and Pfizer (n=1; 2%); Guy's and St Thomas's Charitable Foundation (n=1; 2%); Health authorities of the West Midlands (n=1; 2%); Health Education Board for Scotland (n=1; 2%); Health Enterprise East, NHS innovations hub for East of England (n=1; 2%); Jacob's Foundation (n=1; 2%); Joseph Rowntree Foundation (n=1; 2%); Knowledge Economy Skills Scholarships (n=1; 2%); Lancashire County Council (n=1; 2%); Liverpool Area Based Grants and the SportsLinx Programme (n=1; 2%); Liverpool John Moores University (n=1; 2%); London Borough of Camden and Islington (n=1; 2%); NHS Executive North West R&D Directorate (n=1; 2%); NHS North Lancashire (n=1; 2%); NHS R&D S&W Studentship (n=1; 2%); NIHR Collaboration for Leadership in Applied Health Research and Care for Nottinghamshire, Derbyshire and Lincolnshire (n=1; 2%); NIHR Collaboration for Leadership in Applied Health Research and Care South West (n=1; 2%); NIHR's Collaborations for Leadership in Applied Health Research and Care West Midlands Initiative (n=1; 2%); NIHR National Coordinating Centre for Research Capacity Development (n=1; 2%); Northern Ireland Research and Development Office (n=1; 2%); Northern and Yorkshire Region Research and Development Unit (n=1; 2%); Nottinghamshire Fire and Rescue Service (n=1; 2%); Nottingham Health Authority (n=1; 2%); Primary Care Research Fund of the Chief Scientist Office, Scottish Executive (n=1; 2%); Psychiatry Research Trust (n=1; 2%); Royal College of General Practitioners' Scientific Foundation Board (n=1; 2%); Rugby Football Union (n=1; 2%); Scottish Government Detect Cancer Early Programme (n=1; 2%); South and East Belfast Health and Social Services Trust (n=1; 2%); Sugar Bureau (n=1; 2%); Teenage Cancer Trust (n=1; 2%); The Primary Care and Development Fund (n=1; 2%); The Three Guineas Trust (n=1; 2%); University College London (n=1; 2%); University of Aberdeen (n=1; 2%); Welsh Assembly Government (n=1; 2%).

Appendix 7 - Chapter 3: Cluster and individual-level characteristics adjusted for in the analysis of included studies

Cluster-level characteristics adjusted for in the analysis (N=27)

Characteristic	Statistic (N (%))
Deprivation (school or area in which school is based)	
Yes	17 (63)
School size ¹	
Yes	12 (44)
Baseline characteristic of the outcome	
Yes	8 (30)
Area	
Yes	6 (22)
Year group/key stage/ primary vs secondary	
Yes	3 (11)
School	
Yes	3 (11)
Co-educational status	
Yes	2 (7)
Cohort	
Yes	2 (7)
Ethnicity	
Yes	2 (7)
School performance	
Yes	2 (7)
School type	
Yes	2 (7)
Other ²	
Yes	17 (63)

¹ Included number of students and number of classes

² Other includes: Attitude of the school towards health promotion; Change in social emotional learning; Coaches' attitude; Continuing education (proportion staying on after age 16 years); Educational attainment; Existence of other safety programmes in the school; Existing policy on snacks at morning break; Expressed preference for allocation; Frequency of delivery of PSHE lessons; Hours of daylight; Local family planning services; Percentage of pupils with English as an additional language; Quality and quantity of current school sex education; Season; Special education needs status; Welsh language medium; Whether sex education was taught by a tutor or specialised team of teachers; Whether sex education was taught mainly in year 9 or in year 10.

Individual-level characteristics adjusted for in the analysis (N=45)

Characteristic	Statistic (N (%))
Baseline characteristic of the outcome	
Yes	35 (78)
Gender	
Yes	26 (58)
Deprivation (pupil)	
Yes	14 (31)
Age	
Yes	11 (24)
Ethnicity	
Yes	8 (18)
Car ownership	
Yes	2 (4)
Distance to school	
Yes	2 (4)
Home ownership/rental status	
Yes	2 (4)
Year group	
Yes	2 (4)
Other ¹	
Yes	18 (40)

¹ Other category included: Child's enrolment in an after-school play scheme; Dental attendance; Emotional problems; Family encouragement to wear helmet; Frequency of riding bike; High risk status; Language; Level of play; Maternal paid employment; Maturation status; Number of children living in their household; Parental warning about danger of not wearing a helmet; Playing experience; Playing position; Previous injury history; School leaver status; Survey respondent.

Appendix 8 - Chapter 4: Further details on journals and funding sources (N=24)

The 24 included articles in this systematic review were published across 11 different journals: *Pilot and Feasibility Studies* (n=5; 21%); *International Journal of Behavioural Nutrition and Physical Activity* (n=4; 17%); *Public Health Research* (n=4; 17%); *BMJ Open* (n=3; 13%); *Health Technology Assessment* (n=2; 8%); *Archives of Disease in Childhood* (n=1; 4%); *BMC Public Health* (n=1; 4%); *British Journal of Cancer* (n=1; 4%); *British Journal of Psychiatry* (n=1; 4%); *Prevention Science* (n=1; 4%); and *Trials* (n=1; 4%).

Twelve (50%) studies' main funding source was the *National Institute for Health Research (NIHR)*. Other main funding sources were; *National Prevention Research Initiative* (n=2; 8%); *Cancer Research UK Cancer Prevention Fellowship* (n=1; 4%); *Centre for Diet and Activity Research (CEDAR)* (n=1; 4%); *Department of Health* (n=1; 4%); *Department of Health Policy Research Programme* (n=1; 4%); *ESRC studentship awarded to FUSE* (n=1; 4%); *HSC R&D (NI) Enabling Research Award* (n=1; 4%); *Inspiring Scotland* (n=1; 4%); *Newcastle University Institute for Sustainability* (n=1; 4%); *Purely Nutrition* (n=1; 4%); *Vice Chancellor's Research Scholarship from the University of Ulster* (n=1; 4%).

Appendix 9 – Chapter 5: Intra-cluster correlation coefficients (ICCs) collated from published school-based cluster randomised trials interventions for improving health outcomes on pupils (N=260)

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Cunha [334]	2013	Brazil	Adiposity	Body Mass Index (raw score)	Continuous	Schools	0.07	Yes	No	No
Leme [335]	2016	Brazil	Adiposity	Body Mass Index (raw score)	Continuous	Schools	0.016	No	No	Yes
Lui, Z [336]	2019	China	Adiposity	Body Mass Index (raw score)	Continuous	Schools	0.04	Yes	No	No
Lloyd [241]	2012	England	Adiposity	Body Mass Index (raw score)	Continuous	Schools	0.04	No	No	No
Grydeland [337]	2014	Norway	Adiposity	Body Mass Index (raw score)	Continuous	Schools	0.02	No	No	No
Fitzgibbon [338]	2006	USA	Adiposity	Body Mass Index (raw score)	Continuous	Schools	0.048	No	Yes	No
Gray [339]	2016	USA	Adiposity	Body Mass Index (raw score)	Continuous	Schools	0.041	No	No	No
Sichieri [340]	2009	Brazil	Adiposity	Body Mass Index (raw score) (log transformed)	Continuous	Classes	0.024	No	Yes	No
Waters [341]	2018	Australia	Adiposity	Body Mass Index (z-score)	Continuous	Schools	0.008	Yes (cluster level)	No	No
Pena [342]	2021	Chile	Adiposity	Body Mass Index (z-score)	Continuous	Schools	0.015	No	No	Yes
Pena [342]	2021	Chile	Adiposity	Body Mass Index (z-score)	Continuous	Classes	0.026	No	No	Yes
Li [343]	2019	China	Adiposity	Body Mass Index (z-score)	Continuous	Schools	0.118	Yes	No	No
Adab [143]	2018	England	Adiposity	Body Mass Index (z-score)	Continuous	Schools	0.0211	No	No	No
Lloyd [180]	2018	England	Adiposity	Body Mass Index (z-score)	Continuous	Schools	0.014	Unclear	No	No
Hodgkinson [170]	2019	England	Adiposity	Body Mass Index (z-score)	Continuous	Schools	0.0396	No	Yes	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Breheeny [147]	2020	England	Adiposity	Body Mass Index (z-score)	Continuous	Schools	0.005	No	No	No
Viggiano [344]	2015	Italy	Adiposity	Body Mass Index (z-score)	Continuous	Schools	0.006	No	No	No
Robbins [345]	2020	USA	Adiposity	Body Mass Index (z-score)	Continuous	Schools	0.02	No	No	No
Lubans [346]	2010	Australia	Adiposity	Body Mass Index (z-score)	Continuous	Schools	0.03	No	No	No
Daly [347]	2016	Australia	Adiposity	Body composition (Bone densitometry (DXA))	Continuous	Schools	0.03	No	No	No
Martínez-Vizcaíno [348]	2020	Spain	Adiposity	Fat mass percentage	Continuous	Schools	0.09	Yes	No	No
Ten Hoor [349]	2018	The Netherlands	Adiposity	Fat mass percentage	Continuous	Schools	0.04	No	No	No
Bayer [350]	2009	Germany	Adiposity	Overweight (Body Mass Index)	Binary	Schools	0.023	No	No	No
Muckelbauer [351]	2009	Germany	Adiposity	Overweight (Body Mass Index)	Binary	Schools	0.011	Yes	No	No
Kriemler [352]	2010	Switzerland	Adiposity	Skinfolds (millimetres)	Continuous	Classes	0.06	Yes	No	No
Tarp [353]	2016	Denmark	Adiposity	Waist circumference (centimetres)	Continuous	Schools	0.13	No	Yes	No
Fairclough [160]	2013	England	Adiposity	Waist circumference (centimetres)	Continuous	Schools	0.004	Yes	No	No
Stavnsbo [354]	2020	Norway	Adiposity	Waist circumference (centimetres)	Continuous	Schools	0.19	Yes	Yes	No
Davis [355]	2019	USA	Adiposity	Waist circumference (centimetres)	Continuous	Schools	0.0126	No	No	No
Champion [356]	2016	Australia	Alcohol use	Alcohol knowledge (through questionnaire)	Continuous	Schools	0.07	Yes	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Tael-Öeren [357]	2019	Estonia	Alcohol use	Alcohol use initiation ("Have you ever tried an alcoholic beverage (more than a sip)? Yes/no".)	Binary	Schools	0.04	No	No	No
Martinez-Montilla [358]	2020	Spain	Alcohol use	Binge drinking in the previous 30 days	Binary	Schools	0	Yes	No	No
Segrott [248]	2015	Wales	Alcohol use	Drinking initiation ("Ever had an alcoholic drink")	Binary	Schools	0.112	No	No	No
Newton [282]	2016	Australia	Alcohol use	Frequency of drinking	Continuous	Schools	0.03	No	No	Yes
Teeson [280]	2017	Australia	Alcohol use	Frequency of drinking in the past 6 months	Continuous	Schools	0.01	No	No	Yes
Bodin [359]	2011	Sweden	Alcohol use	Frequent drunkenness	Binary	Schools	0.098	Yes	No	No
Sumnall [360]	2017	Northern Ireland, Scotland	Alcohol use	Heavy Episodic Drinking in the previous 30-days (defined as the consumption of ≥ 6 units (males)/ ≥ 4.5 units (females) on one or more occasions) (log transformed)	Binary	Schools	0.121	No	No	No
Koning [361]	2009	The Netherlands	Alcohol use	Heavy weekly drinking (Boys drinking at least three glasses and girls drinking at least two glasses every week)	Binary	Classes	0.036	No	No	No
D'Amico [362]	2012	USA	Alcohol use	Lifetime alcohol consumption	Binary	Schools	0.02	Yes	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Vallentin-Holbech [363]	2018	Denmark	Alcohol use	Overestimation of peers' lifetime binge drinking (defined as perceived prevalence among peers > actual prevalence in own grade and school +10% tolerance)	Binary	Schools	0.088	Yes	No	No
Haug [364]	2017	Switzerland	Alcohol use	Risky single-occasion drinking (defined as drinking at least 5 standard drinks on one occasion in men and 4 in women) in the past 30-days	Binary	Classes	0.091	Unclear	No	No
Cooper [277]	2006	Ecuador	Allergy	Atopy	Binary	Schools	0.01	Yes	No	No
Palacios [365]	2021	Haiti	Anaemia	Anaemia	Binary	Schools	0.08	No	No	No
Miller, G [366]	2012	China	Anaemia	Haemoglobin concentration	Continuous	Schools	0.086	Yes	No	No
Makris [367]	2019	Cyprus	Biomarker	Urinary biomarkers of exposure to pyrethroid pesticides (3-phenoxybenzoic acid) (log transformed)	Continuous	Schools	0	Yes	No	Yes
Hubbard [172]	2016	Scotland	Cancer	Number recognised of cancer warning signs (Cancer Awareness Measure)	Continuous	Schools	0.03	Yes	No	No
Azam [368]	2021	USA	Dating violence	Victimisation (unwanted sex)	Continuous	Schools	0.0006	No	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Pakpour [369]	2016	Iran	Dental health	Frequency of self-reported brushing ("How many times in the past month have you brushed your teeth?")	Continuous	Schools	0.31	No	No	Yes
Young [370]	2014	Hong Kong	Dental health	Gain in knowledge of emergency management of dental trauma	Continuous	Schools	0.1193	Yes	No	No
Milsom [185]	2006	England	Dental health	Has decayed (untreated) primary teeth	Binary	Schools	0.0271	No	No	No
Rodríguez [371]	2016	Chile	Dental health	Increment of caries (dicdas2–6mft, baseline)	Continuous	Schools	0.03	No	No	No
Haleem [372]	2012	Pakistan	Dental health	Oral health knowledge (through questionnaire)	Continuous	Schools	0.05	No	No	No
Nammontri [373]	2013	Thailand	Dental health	Oral health-related quality of life (Child Perception Questionnaire)	Continuous	Schools	0.013	Yes	No	No
Pakpour [374]	2014	Iran	Dental health	Oral health-related quality of life (Paediatric Quality of Life Inventory (PedsQL) Oral Health Scale)	Continuous	Schools	0.05227758	No	No	No
Worthington [206]	2001	England	Dental health	Plaque Scores	Continuous	Schools	0.099	No	No	No
Redmond [194]	1999	England	Dental health	Proportion of teeth sites with caries at 6 months	Continuous	Schools	0.16	Yes	No	No
Feng [275]	2007	China	Dental health	Volume of lesion (ΔQ) (product of fluorescence loss and area) at a 5% threshold	Continuous	Schools	0	No	No	No
Martiniuk [375]	2007	Canada	Epilepsy	Epilepsy knowledge (through questionnaire)	Continuous	Schools	0.16	No	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Rossetto [376]	2020	Australia	General health	Appropriate first aid given	Binary	Schools	0.03	No	No	No
Shah [278]	2001	Australia	General health	Asthma related quality of life improvement	Continuous	Schools	0.001	No	No	No
Tahlil [377]	2015	Indonesia	General health	Health Knowledge (through questionnaire)	Continuous	Classes	0.1	Yes	No	No
Lassander [378]	2021	Finland	General health	Health-Related Quality of Life - physical health (KINDL-R measure)	Continuous	Classes	0.06	No	No	Yes
Denbæk [379]	2018	Denmark	General health	Illness-related absenteeism in previous week	Binary	Schools	0.014	No	No	No
Denbæk [379]	2018	Denmark	General health	Illness-related absenteeism in previous week	Binary	Classes	0.065	No	No	No
Nsangi [380]	2017	Uganda	General health	Mean test score (percentage of correct answers) on the test of informed health choices taken at the end of the term	Continuous	Schools	0.18	No	No	No
Priest [381]	2014	New Zealand	General health	Number of absence episodes due to any illness	Count/rate	Schools	0.018	No	No	No
Kesztyüs [382]	2016	Germany	General health	Number of sick days	Continuous	Schools	0.045	Yes	No	No
Rosen [383]	2006	Israel	General health	Overall absenteeism	Continuous	Schools	0.0634	No	No	No
Phillips-Howard [384]	2016	Kenya	General health	School dropout (defined as non-attendance for one term with no return to school)	Binary	Schools	0.0084	No	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Süss-Havemann [385]	2020	Germany	General health	Self-efficacy helping in general (% of total variance) (through questionnaire)	Continuous	Classes	0.0338	No	No	Yes
Ssewamala [386]	2021	Uganda	General health	Self-rated health (through questionnaire)	Continuous	Schools	0.02	No	No	No
Woods-Townsend [387]	2021	England	General health	Theoretical health literacy score	Continuous	Schools	0.042	Yes	No	No
Berg [388]	2009	USA	Hearing	Standard threshold shift	Binary	Schools	0	No	No	No
Marlenga [389]	2011	USA	Hearing	Use of hearing protection device for occupation (agriculture)	Binary	Schools	0	No	No	No
Karki [390]	2021	Nepal	Heart disease	Definite or borderline Rheumatic heart disease according to World Health Organisation	Binary	Schools	0.24	No	No	No
Stebbins [269]	2011	USA	Infectious disease	All laboratory confirmed influenza cases	Count/rate	Schools	0.001	No	No	No
Freeman [391]	2013	Kenya	Infectious disease	Ascaris lumbricoides	Binary	Schools	0.04	No	No	No
Gyorkos [392]	2013	Peru	Infectious disease	Ascaris lumbricoides	Binary	Schools	0.042	No	No	No
Whelan [393]	2021	Australia	Infectious disease	Carriage of disease-causing Neisseria meningitidis	Binary	Schools	0.004	No	No	No
Dreibelbis [394]	2014	Kenya	Infectious disease	Diarrhoea in past week	Binary	Schools	0.0701	No	No	No
Liu, X [395]	2019	China	Infectious disease	Hand, foot and mouth disease	Count/rate	Schools	0.047	No	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Cunha [284]	2008	Brazil	Infectious disease	Leprosy (incidence rates)	Count/rate	Schools	0.00013568	Yes	No	No
Joachim [396]	2021	Germany	Infectious disease	Overall acceptance of surveillance methods for Covid-19 (the number of students with informed consent divided by the number of eligible students)	Binary	Classes	0.2	No	No	No
Karanja [397]	2017	Kenya	Infectious disease	Schistosoma mansoni	Binary	Schools	0.13	No	No	No
Watson-Jones [272]	2012	Tanzania	Infectious disease	Vaccine coverage (HPV) (Dose 1 (all schools))	Binary	Schools	0.21	No	No	No
Kovacs [398]	2011	Majorca	Injury	Back injury knowledge	Continuous	Schools	0.33	No	Yes	No
Iserbyt [399]	2017	Belgium	Injury	Basic life support performance	Continuous	Schools	0.04	No	No	No
Iserbyt [399]	2017	Belgium	Injury	Basic life support performance	Continuous	Classes	0.02	No	No	No
Glang [400]	2015	USA	Injury	Composite knowledge of sports concussion	Continuous	Schools	0.089	Unclear	No	No
Nauta [401]	2013	The Netherlands	Injury	Fall-related injuries	Binary	Schools	0.47	No	No	No
Emery [274]	2007	Canada	Injury	Injury rate among basketball players	Count/rate	Teams	0.06	Yes	No	No
Slauterbeck [402]	2019	USA	Injury	Injury to the lower extremity that occurred at a specific location (foot, ankle, leg, knee, thigh, groin, and hip)	Count/rate	Schools	0.03	No	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Emery [403]	2005	Canada	Motor skills	Static balance	Continuous	Schools	0.0358	No	Yes	No
Croker [155]	2012	England	Nutrition	Child Feeding Questionnaire	Continuous	Schools	0.07	No	No	No
James [175]	2004	England	Nutrition	Consumption of carbonated drinks over 3-days (in glasses)	Continuous	Classes	-0.009	No	No	No
De Bock [404]	2012	Germany	Nutrition	Fruit intake	Continuous	Schools	0.016	No	No	Yes
Wyse [405]	2021	Australia	Nutrition	Mean lunch order content of energy (kilograms (kg))	Continuous	Schools	0.1	No	No	No
Amaro [406]	2006	Italy	Nutrition	Nutrition knowledge (through questionnaire)	Continuous	Classes	0.16	No	No	No
Ochoa-Avilés [270]	2017	Ecuador	Nutrition	Nutritional value of dietary intake - Added sugar (grams/day)	Continuous	Schools	0.36	Yes	No	No
Kaufman-Shriqui [407]	2016	Israel	Nutrition	Packed lunch score (quality of packed lunch)	Continuous	Schools	0.05	No	Yes	No
Ezendam [408]	2012	The Netherlands	Nutrition	Snacks per day	Continuous	Schools	0	Yes	No	No
Christian [151]	2014	England	Nutrition	Total fruit and vegetable intake	Continuous	Schools	0.003	No	No	No
Juras [106]	2016	USA	Nutrition	Total fruit and vegetable intake (cup equivalents)	Continuous	Schools	0.094	No	No	No
Giles [163]	2014	Northern Ireland	Obstetrics	Intention to breast feed	Continuous	Schools	0.12	No	No	No
He [409]	2015	China	Ophthalmology	Myopia (3-year cumulative incidence)	Binary	Schools	0.023	No	No	No
Steenart [410]	2019	The Netherlands	Organ donation	Intention to register a decision regarding organ donation	Binary	Classes	0.1	No	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Steenart [410]	2019	The Netherlands	Organ donation	Intention to register a decision regarding organ donation	Binary	Schools	0.03	No	No	No
Hill [411]	2015	New Zealand	Pain	Episode of lower back pain over 9-month period	Binary	Schools	0	No	No	No
Shaygan [412]	2021	Iran	Pain	Pain intensity (11-point numerical rating scale)	Continuous	Schools	0.003	No	No	Yes
Rathleff [413]	2015	Denmark	Pain	Recovered from Patellofemoral pain	Binary	Schools	0	Yes	No	No
Lubans [414]	2020	Australia	Physical activity	20-meter shuttle run test	Continuous	Schools	0.02634	No	No	Yes
Andrade [415]	2014	Ecuador	Physical activity	20-meter shuttle run test	Continuous	Schools	0.15	No	No	No
Harris [416]	2021	New Zealand	Physical activity	20-meter shuttle run test	Continuous	Classes	0.14	Unclear	No	No
Muller [417]	2019	South Africa	Physical activity	20-meter shuttle run test	Continuous	Schools	0.04	No	Yes	No
Puder [418]	2011	Switzerland	Physical activity	20-meter shuttle run test	Continuous	Classes	0.07	No	No	No
Lubans [414]	2020	Australia	Physical activity	20-meter shuttle run test	Continuous	Classes	0.05153	No	No	Yes
Cardon [419]	2009	Belgium	Physical activity	Average activity levels	Continuous	Schools	0.059	Yes	No	No
Kolle [420]	2020	Norway	Physical activity	Daily mean physical activity level counts per minute (full day)	Continuous	Schools	0.04	No	No	Yes
McNeil [421]	2009	Canada	Physical activity	Increased participation in physical or skill-based activities	Binary	Schools	0.04	No	No	No
Schneider [422]	2020	Finland	Physical activity	Leisure-time physical activity engagement	Continuous	Schools	0.003	Unclear	No	No
De Bock [423]	2013	Germany	Physical activity	Mean accelerometry counts (count/15 seconds/day)	Continuous	Schools	0.048	No	No	Yes
Kipping [240]	2008	England	Physical activity	Minutes spent on screen-based activities	Continuous	Schools	0.01	Yes	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Norris [189]	2018	England	Physical activity	Moderate-to-vigorous Physical Activity (mins/school-day)	Continuous	Schools	0.0173062	No	No	No
Suchert [424]	2015	Germany	Physical activity	Moderate-to-vigorous physical activity (days/week)	Continuous	Schools	0.057	No	No	Yes
Sutherland [425]	2016	Australia	Physical activity	Moderate-to-vigorous physical activity (mins/day)	Continuous	Schools	0.03	No	No	Yes
Tymms [205]	2016	England	Physical activity	Moderate-to-vigorous physical activity (mins/day)	Continuous	Schools	0.19	No	No	No
Jago [237]	2012	England	Physical activity	Moderate-to-vigorous physical activity (mins/weekday)	Continuous	Schools	0.018	No	No	No
Jago [238]	2014	England	Physical activity	Moderate-to-vigorous physical activity (mins/weekday)	Continuous	Schools	0.06534	No	No	No
Jago [174]	2015	England	Physical activity	Moderate-to-vigorous physical activity (mins/weekday)	Continuous	Schools	0.0005	Yes	No	No
Jago [426]	2019	England	Physical activity	Moderate-to-vigorous physical activity (mins/weekday)	Continuous	Schools	0.01	Yes	No	No
Robbins [427]	2019	USA	Physical activity	Moderate-to-vigorous physical activity (per week)	Continuous	Schools	0.0126	No	No	No
Barber [228]	2015	England	Physical activity	Moderate-to-vigorous physical activity (mins/day)	Continuous	Schools	0	No	No	No
Harrington [167]	2018	England	Physical activity	Moderate-to-vigorous physical activity (mins/day)	Continuous	Schools	0.03	Yes	No	No
Toftager [428]	2014	Denmark	Physical activity	Overall physical activity (counts per minute)	Continuous	Schools	0.09	Yes	No	Yes

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Lonsdale [429]	2013	Australia	Physical activity	Perceived autonomy during Physical Education lessons (choice provided)	Continuous	Classes	0.04	No	No	Yes
Mendoza [430]	2011	USA	Physical activity	Percentage of trips made by active commuting over 1 school week (percent active commuting)	Continuous	Schools	0.04	No	No	Yes
Mendoza [431]	2017	USA	Physical activity	Percentage of trips made to school by cycling (% cycling)	Continuous	Schools	0.0005	No	No	Yes
Lonsdale [432]	2019	Australia	Physical activity	Proportion of Physical Education lesson time spent in Moderate-to-vigorous physical activity	Continuous	Classes	0.09	No	No	Yes
Lonsdale [432]	2019	Australia	Physical activity	Proportion of Physical Education lesson time spent in Moderate-to-vigorous physical activity	Continuous	Schools	0.07	No	No	Yes
Crammer [433]	2021	USA	Physical activity	Self-efficacy (through questionnaire)	Continuous	Schools	0.06	No	No	No
Nettleford [434]	2021	Canada	Physical activity	Self-reported physical activity over the previous 7 days (Physical Activity Questionnaire for Children)	Continuous	Schools	0.05	No	No	No
Naylor [435]	2008	Canada	Physical activity	Step count (average number of daily steps during 4 measurement sessions)	Continuous	Schools	0.03	No	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Mendoza [436]	2016	USA	Physical activity	TV viewing (mins/day, measured by 7-day TV diary)	Continuous	Schools	0.076	No	No	Yes
Bjelland [437]	2015	Belgium, Germany, Greece, Hungary, Norway	Physical activity	Time used for TV/DVD and computer/games console	Continuous	Schools	0	No	No	No
Cohen [438]	2015	Australia	Physical activity	Total physical activity (counts per minute)	Continuous	Classes	0.08	No	No	Yes
Whittemore [439]	2013	USA	Physical activity	Vigorous exercise	Continuous	Schools	0.03	No	No	Yes
Dzielska [440]	2020	Poland	Physical activity/Nutrition	Health Behaviour Index (HBI)	Continuous	Schools	0.031	No	No	No
Bavarian [441]	2016	USA	Physical activity/Nutrition	Healthy eating and exercise - How much of the time they "eat fresh fruits and vegetables," "drink or eat dairy products," and "exercise hard enough to...sweat and breathe hard."	Continuous	Schools	0.02	No	No	Yes
Kendrick [176]	2007	England	Safety	Knowledge score for fire and burn prevention	Continuous	Schools	0.187	No	No	No
Kendrick [177]	2004	England	Safety	Owens a cycle helmet?	Binary	Schools	0.04	Unclear	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Nykänen [268]	2018	Finland	Safety	Safety preparedness	Continuous	Student groups	0.076	Yes	No	No
Mulvaney [187]	2006	England	Safety	Use of any visibility aid (reflective or florescent) while cycling	Binary	Schools	0.45	Yes	No	No
Henderson [168]	2007	Scotland	Sexual health	Any abortion	Binary	Schools	0.005	No	No	No
Piotrowski [442]	2016	USA	Sexual health	Ever had sexual intercourse	Binary	Classes	0.01	No	No	No
Potter [443]	2016	USA	Sexual health	Initiation of vaginal sex by end of eighth grade	Binary	Schools	0.002	Yes (cluster level)	No	No
Stephenson [271]	2008	England	Sexual health	One or more abortions by age 20 years	Binary	Schools	0.0034	No	No	No
Constantine [444]	2015	USA	Sexual health	Rights with steady partner	Continuous	Classes	0	Yes	No	No
Rohrbach [445]	2015	USA	Sexual health	Sex (vaginal or anal sex) without birth control or condoms in the last 3 months	Binary	Classes	0	Yes	No	No
Mathews [446]	2016	South Africa	Sexual health	Sexual debut	Binary	Schools	0.016	Yes	No	No
Martiniuk [447]	2003	Belize	Sexual health	Sexual knowledge (through questionnaire)	Continuous	Classes	0.025	No	Yes	No
Lohan [242]	2017	Northern Ireland	Sexual health	Unprotected sex	Binary	Schools	0.01	No	No	No
Jemmott [448]	2010	South Africa	Sexual health	Unprotected vaginal intercourse in the past 3 months	Binary	Schools	0.007	No	No	Yes
Aarestrup [449]	2014	Denmark	Skin Cancer	Attitudes toward sunbed use	Continuous	Schools	0.06	No	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Brinker [276]	2020	Brazil	Skin Cancer	Daily sunscreen use in past 30-days	Binary	Schools	0.0016	No	No	No
Brinker [276]	2020	Brazil	Skin Cancer	Daily sunscreen use in past 30-days	Binary	Classes	0.0066	No	No	No
Hunter [450]	2010	USA	Skin Cancer	Directly observed hat use at school	Binary	Schools	0.003	No	No	No
Roetzheim [451]	2011	USA	Skin Cancer	Observed hat use at school	Binary	Schools	0.002	No	No	No
Buller [452]	2006	USA	Skin Cancer	Sun protection Behaviour Composite	Continuous	Schools	0.003	Yes	No	No
Onrust [453]	2018	The Netherlands	Smoking	Attitudes towards smoking (through questionnaire)	Continuous	Schools	0.08	Yes	No	No
Onrust [453]	2018	The Netherlands	Smoking	Attitudes towards smoking (through questionnaire)	Continuous	Classes	0.12	Yes	No	No
Andersen [454]	2015	Denmark	Smoking	Current smoking	Binary	Schools	0.055	No	No	No
Wen [283]	2010	China	Smoking	Ever smoking	Binary	Schools	0.017	No	No	No
Conner [152]	2019	England	Smoking	Ever smoking	Binary	Schools	0.017	No	No	No
Kiewik [455]	2016	The Netherlands	Smoking	Knowledge of smoking (through questionnaire)	Continuous	Classes	0.057	No	No	No
Caria [456]	2011	Austria, Belgium, Germany, Greece, Italy, Spain, and Sweden	Smoking	Lifetime smoking	Binary	Schools	0.08	No	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Hansen [457]	2011	Germany	Smoking	Lifetime smoking	Binary	Schools	0.03	No	No	No
Isensee [458]	2012	Germany	Smoking	Lifetime smoking ("How many cigarettes have you ever smoked in your life?" with response categories "none," "just a few puffs," "1 to 19 (<1 pack)," "20 to 100 (one to five packs)" or ">100 (more than five packs)," resulting in the categorisation of never smokers, experimenters (a few puffs to 100 cigarettes lifetime) and established smoking (>100 cigarettes lifetime))	Unclear	Schools	0.11	No	No	No
Isensee [458]	2012	Germany	Smoking	Lifetime smoking ("How many cigarettes have you ever smoked in your life?" with response categories "none," "just a few puffs," "1 to 19 (<1 pack)," "20 to 100 (one to five packs)" or ">100 (more than five packs)," resulting in the categorisation of never smokers, experimenters (a few puffs to 100 cigarettes lifetime) and established	Unclear	Classes	0.07	No	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
				smoking (>100 cigarettes lifetime))						
Ringwalt [459]	2009	USA	Smoking	Lifetime use of cigarettes	Binary	Schools	0.08	No	No	Yes
Krist [460]	2016	Germany	Smoking	Regular smoking (at least one cigarette per day)	Binary	Classes	0.262	Yes	No	No
Krist [460]	2016	Germany	Smoking	Regular smoking (at least one cigarette per day)	Binary	Schools	0.227	Yes	No	No
Siddiqi [461]	2019	Bangladesh	Smoking	Saliva Cotinine	Continuous	Schools	0	No	No	No
Huque [462]	2015	Bangladesh	Smoking	Smoke-free homes	Binary	Schools	0.198	No	No	Yes
Resnicow [104]	2010	South Africa	Smoking	Smoking - 30-day prevalence	Binary	Schools	0.123	No	No	No
Valdivieso [463]	2015	Spain	Smoking	Smoking - 30-day prevalence	Binary	Schools	0.0567	Yes	No	No
Allara [464]	2015	Italy	Smoking	Smoking - Past 30-days	Binary	Schools	0.021	No	No	No
Haug [465]	2017	Switzerland	Smoking	Smoking abstinence - 7-day point prevalence	Binary	Classes	0.135	No	No	No
Hiemstra [466]	2014	The Netherlands	Smoking	Smoking initiation	Binary	Schools	0	No	No	No
Sashegyi [467]	2000	Canada	Smoking	Smoking status	Binary	Schools	0.2857	No	No	No
Gordon [468]	2008	USA	Smoking	Smoking status	Binary	Schools	0.007	No	No	Yes
Hodder [469]	2017	Australia	Smoking	Tobacco use (ever)	Binary	Schools	0.0182	No	No	No
Campbell [149]	2008	England, Wales	Smoking	Weekly smoker (smokes every week)	Binary	Schools	0.03	No	No	No
Tokolahi [470]	2018	New Zealand	Social emotional functioning	Anxiety symptoms (Multidimensional Anxiety Scale for Children (MASC-10))	Continuous	Schools	0	Yes	No	Yes

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Makover [471]	2019	USA	Social emotional functioning	Anxiety symptoms (Short Mood and Feelings Questionnaire)	Continuous	Schools	0.07	No	No	Yes
Guo [472]	2015	Taiwan	Social emotional functioning	Behavioural intention	Continuous	Schools	0.011	No	No	Yes
McCoy [473]	2021	Brazil	Social emotional functioning	Behavioural problems (Strength and Difficulties Questionnaire (SDQ))	Continuous	Schools	0.13	Yes	No	No
Jenson [474]	2007	USA	Social emotional functioning	Bullying victimisation (dichotomised version of Bully Victim scale from the Revised Olweus Bully/Victim Questionnaire)	Continuous	Schools	0.02735562	No	No	Yes
Agley [475]	2021	USA	Social emotional functioning	Bullying victimization – physical (Bullying and Cyberbullying Scale for Adolescents)	Continuous	Classes	0.0661	No	No	Yes
Baker-Henningham [476]	2021	Jamaica	Social emotional functioning	Child Inhibitory Control	Continuous	Schools	0.09	Yes	No	No
Baker-Henningham [477]	2019	Jamaica	Social emotional functioning	Child behavioural difficulties (Strength and Difficulties Questionnaire (SDQ) - total score)	Continuous	Schools	0.06	No	No	No
Weisleder [478]	2018	Brazil	Social emotional functioning	Cognitive stimulation (StimQ)	Continuous	Schools	0.097	Yes	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Williford [479]	2013	Finland	Social emotional functioning	Cybervictimization (modified version of the Olweus Bully/Victim Questionnaire (OBVQ))	Ordinal	Schools	0.04	No	No	No
Williford [479]	2013	Finland	Social emotional functioning	Cybervictimization (modified version of the Olweus Bully/Victim Questionnaire (OBVQ))	Ordinal	Classes	0.09	No	No	No
Tak [480]	2016	The Netherlands	Social emotional functioning	Depression symptoms (Children's Depression Inventory)	Continuous	Schools	0.022	Yes	No	No
Perry [481]	2017	Australia	Social emotional functioning	Depression symptoms (Major Depression Inventory)	Continuous	Schools	0.017	No	No	Yes
Bradshaw [482]	2012	USA	Social emotional functioning	Disruptive behaviour (The Teacher Observation of Classroom Adaptation - Checklist)	Continuous	Schools	0.05	No	No	Yes
Lopata [267]	2019	USA	Social emotional functioning	Emotion recognition skills (Cambridge Mindreading Face-Voice Battery for Children)	Continuous	School buildings	0.28	No	Yes	No
Edridge [483]	2020	England	Social emotional functioning	Emotional difficulties (Me and My School questionnaire)	Continuous	Classes	0.17	No	No	Yes
Willoughby [484]	2021	Kenya	Social emotional functioning	Executive function	Continuous	Classes	0.14	No	No	No
Lubman [485]	2020	Australia	Social emotional functioning	Help-seeking behaviour - overall - sought help	Binary	Schools	0.02	Yes	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
O'Dea [486]	2021	Australia	Social emotional functioning	Help-seeking intentions for general mental health problems (General Help-Seeking Questionnaire)	Continuous	Schools	0.01	No	No	Yes
Morgan [487]	2018	USA	Social emotional functioning	Instructional Participation (Classroom Measure of Active Engagement)	Continuous	Schools	0.13	Yes	No	No
Link [488]	2020	USA	Social emotional functioning	Mental health knowledge and attitudes	Continuous	Classes	0.094	Yes	No	Yes
Baker-Henningham [489]	2012	Jamaica	Social emotional functioning	Observed conduct problems (log transformed)	Continuous	Schools	0.05	Yes	No	No
Boyd [490]	2018	USA	Social emotional functioning	Play Skills (Structured Play Assessment)	Continuous	Classes	0	No	No	Yes
Lewis [491]	2013	USA	Social emotional functioning	Positive affect (Positive and Negative Affect Scale for Children (PANAS))	Continuous	Schools	0.02	No	No	No
Tol [492]	2014	Burundi	Social emotional functioning	Posttraumatic Stress Disorder (PTSD) symptoms (Child Posttraumatic Symptom Scale)	Continuous	Schools	0.035	No	Yes	No
Kliewer [493]	2011	USA	Social emotional functioning	Problem Behaviour Frequency Scale Physical Aggression (log transformed)	Continuous	Classes	0	No	No	No
Connolly [153]	2018	Northern Ireland	Social emotional functioning	Prosocial behaviour (Strength and Difficulties Questionnaire (SDQ))	Continuous	Schools	0.217	Unclear	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Ford [494]	2021	England, Northern Ireland, Scotland, Wales	Social emotional functioning	Psychopathology (Strength and Difficulties Questionnaire (SDQ))	Continuous	Schools	0.024	No	No	No
Hart [495]	2018	Australia	Social emotional functioning	Quality of mental health first aid intentions - helpful intentions	Continuous	Schools	0	No	No	Yes
Volanen [496]	2020	Finland	Social emotional functioning	Resilience (Resilience scale (RS14))	Continuous	Classes	0.03	No	No	Yes
Mazzoli [497]	2021	Australia	Social emotional functioning	Response inhibition	Continuous	Classes	0	Yes	No	No
Shinde [498]	2018	India	Social emotional functioning	School climate (Beyond Blue School Climate Questionnaire (BBSCQ))	Continuous	Schools	0.13	No	No	No
Obsuth [191]	2017	England	Social emotional functioning	School exclusion	Binary	Schools	0.028	No	No	No
Valente [499]	2021	Brazil	Social emotional functioning	School experience (through questionnaire)	Continuous	Schools	0.023	Yes	No	No
Kirk [279]	2021	Australia	Social emotional functioning	Selective attention (Test of Everyday Attention for Children–Second Edition, (TEACh-2))	Continuous	Classes	0	No	No	Yes
Howard [500]	2020	Australia	Social emotional functioning	Self-Regulation (Head-Toes-Knees-Shoulders)	Continuous	Schools	0.02	No	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Katz [501]	2020	Canada	Social emotional functioning	Self-concept (Self-Description Questionnaire–General Subscale)	Continuous	Classes	0.08792402	No	No	Yes
Katz [501]	2020	Canada	Social emotional functioning	Self-concept (Self-Description Questionnaire–General Subscale)	Continuous	Schools	0.12553393	No	No	Yes
Tirlea [502]	2016	Australia	Social emotional functioning	Self-esteem (Rosenberg self-esteem scale)	Continuous	Schools	0.059	No	No	Yes
Golan [503]	2018	Israel	Social emotional functioning	Self-esteem (Rosenberg self-esteem scale)	Continuous	Classes	0.03	No	No	Yes
DiPerna [504]	2015	USA	Social emotional functioning	Social Skills composite (Social Skills Improvement System Rating Scale)	Continuous	Schools	0.08	No	No	No
DiPerna [504]	2015	USA	Social emotional functioning	Social Skills composite (Social Skills Improvement System Rating Scale)	Continuous	Classes	0.18	No	No	No
Humphrey [173]	2016	England	Social emotional functioning	Social and Emotional Competence Change Index (SECCI)	Continuous	Schools	0.2	No	No	No
Chisholm [150]	2016	England	Social emotional functioning	Stigma of mental illness (willingness to have contact with individuals who are experiencing mental illness) (Reported and Intended Behaviour Scale (RIBS))	Continuous	Classes	0.1	No	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Watanabe [505]	2016	Japan	Social emotional functioning	Subjective psychosomatic symptoms	Continuous	Schools	0.0125	Yes	Yes	No
Wasserman [506]	2015	Austria, Estonia, France, Germany, Hungary, Ireland, Italy, Romania, Slovenia, and Spain	Social emotional functioning	Suicide attempt(s)	Binary	Schools	0.003	No	No	Yes
Halliday [285]	2014	Kenya	Social emotional functioning	Sustained attention (Tests of everyday attention for children) (TEA-Ch battery)	Continuous	Schools	0.07	No	No	No
Stallard [202]	2012	England	Social emotional functioning	Symptoms of (low mood) depression (Short mood and feelings questionnaire)	Continuous	Year groups	0.012	Yes	No	No
Bartholomew [507]	2018	USA	Social emotional functioning	Time on task	Continuous	Schools	0.09	No	No	No
Bartholomew [507]	2018	USA	Social emotional functioning	Time on task	Continuous	Classes	0.14	No	No	No
Dray [508]	2017	Australia	Social emotional functioning	Total difficulties score (Strength and Difficulties Questionnaire (SDQ))	Continuous	Schools	0.16	No	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Ford [161]	2019	England	Social emotional functioning	Total difficulties score (Strength and Difficulties Questionnaire (SDQ))	Continuous	Schools	0.18	No	No	No
Streimann [509]	2020	Estonia	Social emotional functioning	Total difficulties score (Strength and Difficulties Questionnaire (SDQ))	Continuous	Schools	0.2	No	No	No
Calear [510]	2009	Australia	Social emotional functioning	Total score (Revised Children's Manifest Anxiety Scale)	Continuous	Classes	0.02	No	No	Yes
Newton [511]	2014	Australia	Social emotional functioning	Truancy (Number of days students were absent from school in the last year without parental permission)	Continuous	Schools	0.05	No	No	Yes
Gold [512]	2017	Australia	Social emotional functioning	Unhealthy use of music (Healthy-Unhealthy Music Scale)	Continuous	Schools	0.01	No	Yes	No
Axford [145]	2020	Wales	Social emotional functioning	Victimisation (being bullied) -occurring at least twice a month in the last 2 months	Binary	Schools	0.019	No	No	No
Mallick [513]	2018	South Africa	Speech and Language	Attitudes to children who stutter (Stuttering resource outcomes measure)	Continuous	Schools	0.0005	No	No	No
Champion [514]	2016	Australia	Substance misuse	Intentions to use ecstasy	Binary	Schools	0.01	No	No	No
White [250]	2017	Wales	Substance misuse	Lifetime illicit drug use	Binary	Schools	0.003	No	No	No
Miller, E [515]	2012	USA	Violence	Intentions to intervene when witnessing abusive behaviours	Continuous	Schools	0.036	Yes	No	No

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	Cluster unit	ICC estimate	Adjusted for baseline of the outcome	Analysed change scores	Used a repeated measures analysis
Fabbri [516]	2021	Tanzania	Violence	Past-week pupil experience of physical violence from teacher	Binary	Schools	0.13	No	No	No
Temple [517]	2021	USA	Violence	Physical Dating Violence - Perpetration	Binary	Schools	0.109	No	No	No
Wolfe [518]	2009	Canada	Violence	Physical Dating Violence Reported in the Past Year (All students)	Binary	Schools	0.02	No	No	No
Miller, E [519]	2020	USA	Violence	Positive bystander intervention behaviours	Continuous	Schools	0.007	No	No	No
Sanchez-Jimenez [281]	2018	Spain	Violence	Psychological aggression (Psychological Dating Abuse Scale)	Continuous	Schools	0.002	No	No	Yes
Devries [520]	2015	Uganda	Violence	Student self-reported past week physical violence at school	Binary	Schools	0.06	No	No	No
Beets [521]	2009	USA	Violence	Violent behaviours	Binary	Schools	0.06	No	No	No

Appendix 10 – Chapter 5: School- and class-level intra-cluster correlation coefficients (ICCs) are reported side-by-side for the 14 studies that reported at both those levels

Author	Year	Country	Health area of the outcome	Outcome description	Outcome type	School-level ICC estimate	Class-level ICC estimate
Pena [342]	2021	Chile	Adiposity	Body Mass Index (z-score)	Continuous	0.015	0.026
Denbæk [379]	2018	Denmark	General health	Illness-related absenteeism in previous week	Binary	0.014	0.065
Iserbyt [399]	2017	Belgium	Injury	Basic life support performance	Continuous	0.04	0.02
Steenart [410]	2019	The Netherlands	Organ donation	Intention to register a decision regarding organ donation	Binary	0.03	0.1
Lonsdale [432]	2019	Australia	Physical activity	Proportion of Physical Education lesson time spent in Moderate-to-vigorous physical activity	Continuous	0.07	0.09
Lubans [414]	2020	Australia	Physical activity	20-meter shuttle run test	Continuous	0.02634	0.05153
Brinker [276]	2020	Brazil	Skin cancer	Daily sunscreen use in past 30-days	Binary	0.0016	0.0066

Isensee [458]	2012	Germany	Smoking	Lifetime smoking ("How many cigarettes have you ever smoked in your life?" with response categories "none," "just a few puffs," "1 to 19 (<1 pack)," "20 to 100 (one to five packs)" or ">100 (more than five packs)," resulting in the categorisation of never smokers, experimenters (a few puffs to 100 cigarettes lifetime) and established smoking (>100 cigarettes lifetime))	Unclear	0.11	0.07
Krist [460]	2016	Germany	Smoking	Regular smoking (at least one cigarette per day)	Binary	0.227	0.262
Onrust [453]	2018	The Netherlands	Smoking	Attitudes towards smoking (through questionnaire)	Continuous	0.08	0.12

Williford [479]	2013	Finland	Social emotional functioning	Cybervictimization (modified version of the Olweus Bully/Victim Questionnaire (OBVQ))	Ordinal	0.04	0.09
DiPerna [504]	2015	USA	Social emotional functioning	Social Skills composite (Social Skills Improvement System Rating Scale)	Continuous	0.08	0.18
Bartholomew [507]	2018	USA	Social emotional functioning	Time on task	Continuous	0.09	0.14
Katz [501]	2020	Canada	Social emotional functioning	Self-concept (Self-Description Questionnaire–General Subscale)	Continuous	0.126	0.088

Appendix 11 - Chapter 6: Ethical approval for use of the datasets has been granted by the University of Exeter Medical School Research Ethics Committee

Estimating parameters to aid in the design of school-based cluster randomised trials of interventions for improving mental health outcomes on pupils

Project description:

This study will use data from school-based cluster randomised controlled trials of mental health interventions for improving pupils' outcomes to estimate parameters that can be used to aid the design of future similar studies. Specifically we will: (1) obtain estimates of the intra-cluster (intra-school) correlation coefficient and components of variance that are needed to calculate the sample size required for cluster randomised trials; and (2) identify the school-level (cluster-level) characteristics that are most strongly predictive of pupil mental health outcomes and are, therefore, suitable factors on which to stratify the randomisation of schools and incorporate in the analysis as adjustment (prognostic) factors when estimating the intervention effect in such studies. The findings of this study will aid researchers in ensuring their school-based cluster trials are large enough to evaluate pupil health interventions and improve the efficiency of their design and analysis.

Project Dates

4th Jan 2022 – 1st June 2022

Scope

Does your research involve only secondary data?

Yes

Does your project require external ethical review?

No

Please summarise the background to the project?

Cluster randomised trials (CRTs) are increasingly used in the school setting to evaluate public health interventions for improving outcomes on pupils [50]. Such studies involve allocation of entire clusters of individuals (e.g., schools, year groups,

classrooms) rather than the individuals (pupils) themselves on whom outcomes are measured.

A characteristic feature of CRTs is that observations on participants from the same cluster are usually more similar to each other than observations on participants from different clusters [4]. For example, pupils in the same school are more likely to have similar outcomes than those from different schools. This similarity, or lack of statistical independence, means that the usual methods for calculating sample size and analysing data in trials that randomise individuals should not be used in studies that randomise clusters [4]. Use of standard sample size formulae may result in an underpowered study, and the use of standard analytical methods to estimate the intervention effect may result in confidence intervals that are too narrow and p -values that are too small, thus exaggerating the impact of the intervention [3]. Therefore, accounting for the clustered design in sample size calculation and analysis is essential in CRTs. The similarity between observations from the same cluster for a given outcome is quantified by the intra-cluster correlation coefficient (ICC). The ICC is defined as the proportion of the total variation in the outcome that is between clusters as opposed to within clusters. Information about the ICC (or the between-cluster variance component and the within-cluster variance component) is invaluable when designing CRTs as they are needed to calculate the required sample size. They can be obtained from previous studies with a similar cluster structure and similar outcomes [10, 11], but there is a lack of published estimates relevant to school-based CRTs in the UK. The dissemination of ICCs based on school and classroom clusters would greatly aid the planning of future school-based CRTs.

CRTs usually include only a relatively small number of clusters. Consequently, simple randomisation may result in trial arms that are unbalanced with respect to cluster level characteristics (e.g., socioeconomic profile of the cluster) that may be related to the outcome of interest [4]. It is then harder to ascribe any resulting differences on the outcomes between the trial arms as resulting from the intervention itself. Restricted randomisation involves controlling the randomisation process to ensure the trial arms are similar (or balanced) with respect to key cluster characteristics that are expected to predict the study outcomes, thus enabling a fair comparison between the intervention and control arms. An example of restricted

randomisation is stratified randomisation where clusters are first grouped into strata based on factors that the investigators believe are necessary to balance between trial arms (e.g., geographical area, cluster size, socio-economic status). Within each stratum, clusters are then randomly assigned to the trial arms [2]. When estimating the intervention effect, inclusion in the analytical models of the factors used to balance the randomisation improves the precision of the resulting estimates if those factors are related to the outcome. Therefore, it is useful to know which cluster level characteristics are predictive of the trial outcomes so that the best ones are chosen to balance on in the randomisation and/or adjusted for in the analysis [3]. In a recent systematic review describing school-based CRTs with pupil health outcomes in the UK, 80% of studies used some form of restricted randomisation [117]. Despite this, there is little evidence on which factors *should* be balanced on for pupil outcomes in specific disease/health areas in CRTs in the school setting and justification for the choice of balancing factors is rarely provided. School-based CRTs with mental health outcomes have balance the randomisation on different school-level characteristics [145, 161, 202, 203, 298], but it is rarely reported whether the balancing factors are ultimately predictive of the outcome.

This study aims to use data from previous school-based cluster randomised trials to: (a) estimate the intra-cluster (intra-school) correlation coefficients for pupil mental health outcomes and (b) identify the school (cluster) level characteristics that are most strongly predictive of pupil mental health outcomes.

Please explain the aims of the project and what you intend to achieve

This study will conduct secondary analyses using data from several school-based cluster randomised trials of interventions to improve pupil mental health outcomes to:

- (i) Collate estimates of the between-cluster and within-cluster components of variance and the ICCs from school-based CRTs in mental health.
Specifically, we will:
 - a. Estimate components of variance at the school, year group, classroom and pupil levels
 - b. Compare the size of the ICC for the same pupil outcome between pupil-report, parent report and teacher-report

- c. Compare ICC values for the same outcome between baseline and follow-up
- d. Compare ICC values between trial arms
- e. Compare ICC values across children with different demographic characteristics (e.g., based on age, gender, ethnicity, SES)
- (ii) Describe the strength of school-level characteristics for predicting pupil mental health outcomes in UK school-based CRTs. Specifically, we will:
 - a. Examine the strength of relationship of the school (cluster) level variables used to balance the randomisation with the pupil outcomes.
 - b. Examine the strength of relationship of other cluster-level characteristics that were not used to balance the randomisation with the pupil health outcomes.
 - c. Examine the strength of relationship of cluster-level summaries of baseline measurements of the outcome with the pupil health outcomes at follow-up.
 - d. Describe the extent to which the cluster-level characteristics account for the size of the intra-cluster correlation coefficient.

The outcomes for this project will include a peer-reviewed publication, presentations at conferences related to the topic area (such as the annual *Current Developments in Cluster Randomised Trials and Stepped Wedge Designs* meeting), and further dissemination through social media. The research will also form part of Kitty Parker's PhD and therefore, will be included in her final thesis (due for submission May 2023).

Please describe how the research will be conducted in a way that ensures its quality and integrity

This study has been designed to address a gap in the methodological literature and to aid researchers in the design and analysis of future school-based CRTs. The study objectives have been informed by a thorough literature review of the area [116, 117]. The work has also been discussed with other academics within the College who have confirmed the unique and important contribution this work will make. To ensure quality and integrity, we will develop an in-depth protocol detailing the background, methods, and data analysis plan for this proposal. This project will follow the protocol rigorously and the protocol will be made available through the

Open Science Framework to allow for transparency in the research process. We also plan to keep detailed information on data preparation and analysis used in this study (see Data Management section). Plans to ensure the data management is robust and rigorous can be found in the Data Management section of this application. We plan to disseminate this research through different forms to relevant audiences, including dissemination through a peer-reviewed publication, presentations at conferences, and via social media and the National Institute for Health Research platform. This project will also form part of Kitty Parker's PhD thesis (due for submission May 2023).

Methodology

Please provide a summary of the research methodology below. For each method, please describe how it has been selected and how the data will be analysed.

Method

Data from five completed [145, 161, 202, 203, 298] and two ongoing school-based CRTs measuring mental health, behaviour and well-being outcomes will be used in this study. Use has been granted by the principle investigator from each of these studies. Consent for use of this data was obtained during the original study. All cluster-level and individual-level data are fully anonymised.

Ethical approval was granted for the original studies for which the data were collected, and participant data were fully anonymised.

Description of Participants

Participants include school pupils in full-time education who took part in one of seven school-based CRTs in the United Kingdom. Written consent was obtained at the cluster/school level in order for schools to participate in each of the included studies. Written consent was also obtained for class teachers if they were involved in the researcher study. Parental/guardian written consent was obtained for their child's/children's participation and data collection. Either pupil's written consent, or verbal assent was obtained from children for participation and data collection in this research.

A description of the specific participant information for each of the seven CRTs are listed below:

1. MYRIAD [298] – 85 schools participated in this study and written consent was obtained at the school level from headteachers. Consent from 739 teachers was also obtained. Parental/caregiver passive ‘opt-out’ consent, and child assent was obtained at each data collection point. 26,885 school students (ages 12–14 years) provided consent and participated in this study.
2. STARS [161] – 80 schools participated in this study and written consent was obtained from the headteacher for the school’s participation and from the class teacher for their involvement after nomination by the headteacher, including for reporting outcomes on pupils. 2075 school pupils aged 4-9 years provided consent for participation and at each data collection point. Parents/guardians could provide passive ‘opt-out’ consent for their child, and verbal assent was obtained from children each time they were asked to complete a questionnaire.
3. KiVa [145] – Headteachers in 22 schools provided written consent for their school to participate in the trial and also consent to allow the research team to collect data to use in analyses. Parents/guardians of 3214 pupils, aged 7-11 years, provided ‘opt-out’ (passive) consent. Pupils provided active consent to complete the KiVa pupil online questionnaire and at each data collection point.
4. PACES [203] – Headteachers of 41 schools provided written consent for participation. Parental/guardian passive ‘op-out’ consent and signed assent was obtained from 1362 school pupils (aged 9 – 10 years) for participation and providing data at each follow-up.
5. PROMISE [202] – Headteachers of 8 schools provided written consent for participation. Parental/guardian passive ‘op-out’ consent and signed assent was obtained from 1064 school pupils (aged 12-16 years) for participation and providing data at each follow-up.
6. iCATS – This is an ongoing study, therefore the total number of schools and pupils who will be participating in this research is not known. Children will be aged 5-11 years. Informed written consent will be sought from schools, school staff and parents/carers. Assent will be obtained from children.

7. MyCATS – This is an ongoing study, therefore the total number of schools and pupils who will be participating in this research is not known. Children will be aged 4- 6 years. Informed written consent will be sought from schools, school staff and parents/carers. Assent will be obtained from children.

Due to the nature of the research questions, it is necessary to involve vulnerable populations (i.e., children) in the research study.

Why methods were selected?

Secondary data analysis is used to address our research questions because it would be expensive and inefficient to conduct primary research solely to estimate the parameters of interest. The existence of data from relevant previously conducted trials provides the opportunity to meet our objectives efficiently. A secondary data analysis of these mental health datasets will be undertaken as: 1) the researchers have access to these datasets through co-authorship/supervisors/professional links/affiliations; 2) the studies provide data for a wide range of mental health outcomes; 3) the data have been previously assessed and ‘cleaned’, making them good candidates for secondary data analysis.

Data Analysis

The data will be analysed quantitatively and will involve the following:

1. Descriptive statistics will be used to summarise characteristics and outcomes of the participating schools and pupils.
2. The ICC and components of variance will be estimated by fitting mixed effects (“multilevel”) regression models to the outcomes. Estimates will be reported with 95% confidence intervals.
3. Mixed effects (“multilevel”) models will be fitted to examine the relationships between the cluster level predictors and the outcomes. P-values will be reported for each predictor.
4. We will report the size of the reduction in the intra-cluster (intra-school) correlation coefficient and the reduction in the total outcome variance that results from using the cluster level factors as predictors in the mixed effects model.

Where will the project be undertaken?

The project will be undertaken on a University of Exeter, password protected, Bitlocker encrypted laptop (see Data Management section for details on how data will be stored). The main applicant (Kitty Parker) will undertake this work on the laptop at home. The laptop will be locked and password protected when unattended.

Data management

What data will be collected and used during the project?

No data will be collected for the current project as the project involves a secondary data analysis of pre-existing fully anonymised data from completed school-based CRTs.

The specific data we plan to use will be:

- Cluster-level (school-level) demographic information (e.g., socioeconomic status, percentage free school meals, percentage of white individuals) and pupil demographic information (e.g., age)
- Number of clusters and pupils providing outcome data
- Pupil mental health related outcomes report by teachers, parents and the pupils themselves.

Data were collected during each of the school-based CRTs which involved the completion of questionnaires.

Is there an access control process or a gatekeeper for access to data e.g. secondary data?

Yes. Access has been granted by each of the principle investigators from each school-based CRT. They have been fully informed on how we intend to use the data.

Where and how will data be stored during the project?

Data will be downloaded using a secure drop box from each of the principle investigators by the main applicant (Kitty Parker) using her University of Exeter encrypted laptop which requires personal authentication to access. The data will be held electronically and securely on the main applicant's University One Drive account which is only accessible through use of secure username and password. Access to this will be restricted to the applicants of the current proposal (i.e., Kitty Parker and Obi Ukoumunne). All applicants will have to use a secure username and password to access the folder.

The data for the current project are not generated by the applicants but are already

collected as part of previous school-based CRTs. In keeping with the UK Data Services End User License, data will be destroyed upon completion of the project. Completion of the project will be defined as the point at which all analyses have been conducted and any papers related to the project have been accepted for publication (in recognition of the fact peer reviewers may recommend changes/additional analyses). Data will be destroyed using a secure erasure programme (see more details below). Detailed information around data preparation and analytic code used for the study will be retained.

How long will the data be retained after the project is complete?

12 months (until 31st May 2024)

Will any of the data be used in future research and/or made available to other research projects?

Data will not be retained for future research and/or made available to other research projects by the applicants.

How will data be destroyed when it is no longer needed?

No personal data will be collected or stored as part of this project, but anonymised data will be destroyed upon completion of the project using a secure erasure programme by the main applicant (Kitty Parker). In the event that the main applicant is unable to do this, one of the other applicants (Obi Ukoumunne) will undertake this task.

How will access to the data be controlled?

Only the applicants of the current project will be provided with access to the data. As previously discussed, this will be done through the University of Exeter One Drive which allows folders to be shared between specific members of staff (both internally and externally) using password controls.

Will participant data be treated as confidential?

Yes, data are fully anonymised on download from the Principle Investigator and will be stored according to GDPR guidelines. Confidentiality will be preserved at all times by not attempting to identify individuals, teachers or schools in the data.

Will participant data be anonymous?

Yes, data are fully anonymised on download from the Principle Investigator of each study and, therefore, no further procedures will be needed or undertaken to anonymise the data.

Appendix 12 - Chapter 6: 95% confidence intervals (CIs) for school-level intra-cluster correlation coefficient (ICC) estimates from the 5 datasets

STARS study school-level ICCs with 95% CIs

Outcome (reporter)	Measurement time (months)¹	N	ICC	95% CI
Total difficulties score (SDQ) (Teacher)	0	2074	0.120	(0.083 to 0.170)
	9	2001	0.180	(0.130 to 0.243)
	18	1848	0.179	(0.129 to 0.243)
	30	1756	0.121	(0.082 to 0.175)
Emotion subscale (SDQ) (Teacher)	0	2074	0.101	(0.068 to 0.146)
	9	2001	0.202	(0.149 to 0.269)
	18	1848	0.179	(0.129 to 0.243)
	30	1756	0.090	(0.059 to 0.136)
Conduct subscale (SDQ) (Teacher)	0	2074	0.062	(0.038 to 0.100)
	9	2001	0.092	(0.061 to 0.138)
	18	1848	0.117	(0.079 to 0.169)
	30	1756	0.104	(0.069 to 0.155)
Hyperactivity subscale (SDQ) (Teacher)	0	2074	0.053	(0.032 to 0.088)
	9	2001	0.090	(0.059 to 0.135)
	18	1848	0.091	(0.059 to 0.137)
	30	1756	0.072	(0.044 to 0.114)
Peer problems subscale (SDQ) (Teacher)	0	2074	0.152	(0.101 to 0.210)
	9	2001	0.119	(0.081 to 0.170)
	18	1848	0.131	(0.091 to 0.186)
	30	1756	0.098	(0.064 to 0.146)

Outcome (reporter)	Measurement time (months)¹	N	ICC	95% CI
Prosocial behaviour subscale (SDQ) (Teacher)	0	2074	0.234	(0.176 to .304)
	9	2001	0.251	(0.189 to .324)
	18	1848	0.204	(0.150 to .273)
	30	1756	0.164	(0.116 to 0.226)
Total difficulties score (SDQ) (Parent)	0	1466	0.026	(0.009 to 0.070)
	9	1285	0.046	(0.022 to 0.095)
	18	1225	0.031	(0.011 to 0.083)
	30	1125	0.034	(0.012 to 0.089)
Emotion subscale (SDQ) (Parent)	0	1467	0.025	(0.009 to 0.066)
	9	1286	0.031	(0.012 to 0.078)
	18	1227	0.014	(0.003 to 0.077)
	30	1126	0.019	(0.004 to 0.082)
Conduct subscale (SDQ) (Parent)	0	1467	0.013	(0.002 to 0.072)
	9	1287	0.016	(0.004 to 0.072)
	18	1228	0.030	(0.010 to 0.087)
	30	1127	0.001	(0 to 1)
Hyperactivity subscale (SDQ) (Parent)	0	1466	0.004	(0 to 0.431)
	9	1287	0.012	(0.002 to 0.079)
	18	1227	0.009	(0.001 to 0.113)
	30	1127	0.010	(0.001 to 0.123)
Peer problems subscale (SDQ) (Parent)	0	1466	0.021	(0.007 to 0.066)
	9	1286	0.049	(0.023 to 0.099)
	18	1227	0.027	(0.009 to 0.078)
	30	1126	0.045	(0.020 to 0.099)

Outcome (reporter)	Measurement time (months)¹	N	ICC	95% CI
Prosocial behaviour subscale (SDQ) (Parent)	0	1467	0	
	9	1287	0	
	18	1228	0.007	(<0.001 to 0.180)
	30	1127	0	
Pupil Behaviour Questionnaire (teacher)	0	2053	0.064	(0.040 to 0.101)
	9	1986	0.086	(0.056 to 0.130)
	18	1886	0.082	(0.052 to 0.127)
	30	1760	0.081	(0.051 to 0.126)
'How I Feel About My School measure' questionnaire (pupil)	0	2074	0.052	(0.031 to 0.087)
	9	2001	0.077	(0.050 to 0.119)
	18	1848	0.106	(0.071 to 0.155)
	30	1756	0.111	(0.074 to 0.162)

¹ Time points at 9, 18, 30 months adjusted for trial arm status

KiVa study school-level ICCs with 95% CIs

Outcome (reporter)	Measurement time (months)¹	N	ICC	95% CI
Total difficulties score	0	2832	0.037	(0.012 to 0.111)
(SDQ) (teacher)	12	2652	0.075	(0.032 to 0.164)
Emotional subscale	0	2832	0.033	(0.009 to 0.109)
(SDQ) (teacher)	12	2652	0.092	(0.042 to 0.190)
Conduct subscale	0	2832	0.042	(0.016 to 0.108)
(SDQ) (teacher)	12	2652	0.055	(0.024 to 0.121)
Hyperactivity subscale	0	2832	0.005	(<0.001 to 0.118)
(SDQ) (teacher)	12	2652	0.030	(0.010 to 0.086)
Peer problems subscale	0	2832	0.025	(.008 to 0.080)
(SDQ) (teacher)	12	2652	0.041	(0.015 to 0.104)
Prosocial behaviour subscale	0	2832	0.010	(0.001 to 0.163)
(SDQ) (teacher)	12	2652	0.017	(0.003 to 0.091)
Bullying victimisation	0	2876	0.012	(0.003 to 0.043)
(Olweus Bully/Victim Questionnaire) (Pupil)	12	2581	0.019	(0.006 to 0.058)
Told school about being bullied	0	2876	0.013	(0.004 to 0.041)
(Olweus Bully/Victim Questionnaire) (Pupil)	12	2581	0.009	(0.002 to 0.046)
Did not tell school about being bullied	0	2876	0.010	(0.002 to 0.041)
(Olweus Bully/Victim Questionnaire) (Pupil)	12	2581	0.018	(0.006 to 0.052)
Told home about being bullied	0	2876	0.006	(0.001 to 0.042)
(Olweus Bully/Victim Questionnaire) (Pupil)	12	2581	0.017	(0.006 to 0.047)
Bully perpetration	0	2876	0.010	(0.003 to 0.038)
(Olweus Bully/Victim Questionnaire) (Pupil)	12	2581	0.010	(0.002 to 0.040)

¹ Time point at 12 months adjusted for trial arm status

PACES study school-level ICCs with 95% CIs

Outcome (measure) (reporter)	Measurement time (months)¹	N	ICC	95% CI
Total difficulties score (SDQ) (parent)	0	547	0.002	(0 to 1)
	6	460	0	
	12	425	0.050	(0.012 to 0.192)
Emotional subscale (SDQ) (parent)	0	566	0	
	6	475	0	
	12	439	0.032	(0.005 to 0.169)
Conduct subscale (SDQ) (parent)	0	563	0	
	6	473	0	
	12	441	0.006	(0 to 0.936)
Hyperactivity subscale (SDQ) (parent)	0	566	0	
	6	475	0	
	12	437	0.010	(0 to 0.601)
Peer problems subscale (SDQ) (parent)	0	561	0.028	(0.004 to 0.162)
	6	475	0.059	(0.017 to 0.182)
	12	438	0.017	(<0.001 to 0.483)
Prosocial behaviour subscale (SDQ) (parent)	0	561	0	
	6	471	0	
	12	440	0	
Total anxiety score (RCADS-30) (Parent)	0	482	0	
	6	426	0	
	12	406	0.016	(<0.001 to 0.616)
Depression subscale (RCADS-30) (Parent)	0	560	0.017	(0.001 to 0.182)
	6	477	0	
	12	445	0	
Separation Anxiety Disorder subscale (RCADS-30) (Parent)	0	519	0	
	6	448	0.021	(0.001 to 0.299)

Outcome (measure) (reporter)	Measurement time (months)¹	N	ICC	95% CI
	12	432	0.016	(0 to 0.751)
Social phobia subscale (RCADS-30) (Parent)	0	558	0.007	(0 to 0.636)
	6	479	0	
	12	441	0.023	(0.003 to 0.180)
Generalised Anxiety Disorder subscale (RCADS-30) (Parent)	0	557	0	
	6	477	0.011	(<0.001 to 0.388)
	12	444	0.022	(.002 to 0.216)
Panic disorder subscale (RCADS-30) (Parent)	0	550	0.004	(0 to 0.998)
	6	473	0	
	12	443	0.007	(0 to 0.786)
Obsessive-compulsive Disorder subscale (RCADS-30) (Parent)	0	559	0	
	6	478	0.004	(0 to 0.998)
	12	444	0.006	(0 to 0.927)
Total anxiety score (RCADS-30) (Pupil)	0	1281	0	
	6	1274	0.002	(0 to 0.988)
	12	1203	0.010	(0.001 to 0.148)
Depression subscale (RCADS-30) (Pupil)	0	1332	0.008	(<0.001 to 0.167)
	6	1305	0	
	12	1250	0.013	(0.001 to 0.163)
Separation Anxiety Disorder subscale (RCADS-30) (Pupil)	0	1330	0.010	(0.001 to 0.140)
	6	1308	0.023	(0.005 to 0.093)
	12	1247	0.026	(0.007 to 0.092)
Social phobia subscale (RCADS-30) (Pupil)	0	1328	0	
	6	1307	0.014	(0.002 to 0.080)
	12	1244	0.006	(0 to 0.350)
Generalised Anxiety Disorder subscale (RCADS-30) (Pupil)	0	1328	0	
	6	1305	0	
	12	1242	0.002	(0 to 1)
Panic disorder subscale	0	1326	0	

Outcome (measure) (reporter)	Measurement time (months)¹	N	ICC	95% CI	
(RCADS-30) (Pupil)	6	1305	0	(<0.001 to 0.151)	
	12	1247	0.006		
Obsessive-compulsive Disorder subscale (RCADS-30) (Pupil)	0	1325	0		
	6	1307	0		
	12	1245	0		
Victimisation (Olweus Bully/Victim Questionnaire) (Pupil)	0	1338	0.015	(0.002 to 0.083)	
	6	1316	0.031	(0.010 to 0.089)	
	12	1254	0.005	(0 to 0.732)	
Worry (Penn Worry Scale) (Pupil)	0	1310	0		
	6	1298	0		
	12	1230	0.010		(0.001 to 0.124)
Self-esteem (Rosenberg Self-Esteem Scale) (Pupil)	0	1295	0		
	6	1285	0.023		(0.006 to 0.084)
	12	1224	0.012		(<0.001 to 0.243)
Total life satisfaction (CHU9D) (Pupil)	0	1333	0.009	(0.001 to 0.122)	
	6	1302	0.003	(0 to 0.956)	
	12	1241	0.027	(0.007 to 0.106)	

¹ Time points at 6 and 12 months adjusted for trial arm status

PROMISE study school-level ICCs with 95% CIs

Outcome (measure) (reporter)	Measurement time (months)¹	N	ICC	95% CI
Total anxiety score (RCADS-30) (Pupil)	0	4588	0.007	(0.001 to 0.085)
	6	4395	0	
	12	3948	0.006	(<0.001 to 0.121)
Depression subscale (RCADS-30) (Pupil)	0	4607	0.010	(0.002 to 0.060)
	6	4416	0	
	12	3954	0.005	(<0.001 to 0.093)
Panic disorder subscale (RCADS-30) (Pupil)	0	4612	0.009	(0.002 to 0.051)
	6	4422	0	
	12	3957	0.005	(<0.001 to .0061)
Social phobia subscale (RCADS-30) (Pupil)	0	4612	0	
	6	4420	0.002	(0 to 0.999)
	12	3956	0.006	(<0.001 to 0.284)
Generalised Anxiety Disorder subscale (RCADS-30) (Pupil)	0	4616	0	
	6	4427	0.004	(<0.001 to 0.132)
	12	3958	0.008	(0.001 to 0.066)
Separation Anxiety Disorder subscale (RCADS-30) (Pupil)	0	4616	0.014	(0.004 to 0.050)
	6	4426	0.002	(<0.001 to 0.038)
	12	3958	0.007	(0.001 to 0.044)
Self-esteem (Rosenberg self-esteem scale) (Pupil)	0	4576	.004	(<.001 to 128)
	6	4392	0	
	12	3944	0	
Negative thinking (Personal failure subscale - CATS) (Pupil)	0	4596	0.009	(0.002 to 0.047)
	6	4401	<0.001	(0 to 1)
	12	3945	0.001	(0 to 0.995)
School connectedness (PSSM scale) (Pupil)	0	4567	0.007	(0.001 to 0.056)
	6	4367	0.016	(0.004 to 0.067)
	12	3913	0.016	(0.004 to 0.070)

Outcome (measure) (reporter)	Measurement time (months)¹	N	ICC	95% CI
Symptoms of low mood (SMFQ) (Pupil)	0	4784	0.010	(0.002 to 0.058)
	6	4480	0.001	(0 to 1)
	12	4140	0.005	(<0.001 to 0.090)

¹ Time points at 6 and 12 months adjusted for trial arm status

MYRIAD study school-level ICCs with 95% CIs

Outcome (measure) (reporter)	Measurement time (months)¹	N	ICC	95% CI
Total difficulties score (SDQ) (Pupil)	0	8252	0.025	(0.016 to 0.040)
	12	8042	0.021	(0.013 to 0.035)
	19	7542	0.020	(0.011 to 0.034)
	24	7225	0.017	(0.010 to 0.030)
Emotional subscale (SDQ) (Pupil)	0	8254	0.018	(0.011 to 0.030)
	12	8042	0.022	(0.014 to 0.034)
	19	7542	0.022	(0.013 to 0.036)
	24	7226	0.020	(0.012 to 0.033)
Conduct subscale (SDQ) (Pupil)	0	8253	0.022	(0.013 to 0.035)
	12	8042	0.017	(0.010 to 0.029)
	19	7542	0.014	(0.007 to 0.027)
	24	7226	0.011	(0.005 to 0.023)
Hyperactivity subscale (SDQ) (Pupil)	0	8253	0.021	(0.013 to 0.034)
	12	8042	0.014	(0.008 to 0.025)
	19	7542	0.015	(0.008 to 0.028)
	24	7226	0.013	(0.007 to 0.025)
Peer problems subscale (SDQ) (Pupil)	0	8253	0.015	(0.009 to 0.025)
	12	8042	0.017	(0.010 to 0.028)
	19	7542	0.013	(0.007 to 0.024)
	24	7225	0.015	(0.009 to 0.027)
Prosocial behaviour subscale (SDQ) (Pupil)	0	8254	0.012	(0.006 to 0.025)
	12	8042	0.019	(0.011 to 0.032)
	19	7542	0.020	(0.012 to 0.034)
	24	7226	0.022	(0.013 to 0.036)
Symptoms of depression (CES-D) (Pupil)	0	8370	0.016	(0.009 to 0.027)
	12	8054	0.023	(0.015 to 0.037)
	19	7561	0.019	(0.011 to 0.033)

Outcome (measure) (reporter)	Measurement time (months)¹	N	ICC	95% CI
	24	7238	0.019	(0.011 to 0.032)
Well-being (WEMWEBS) (Pupil)	0	8333	0.015	(0.008 to 0.028)
	12	8058	0.019	(0.011 to 0.032)
	19	7572	0.017	(0.010 to 0.030)
	24	7244	0.016	(0.009 to 0.029)
Executive Function (BRIEF-2) (Pupil)	12	7121	0.090	(0.062 to 0.131)
	19	7022	0.065	(0.042 to 0.101)
	24	6878	0.058	(0.038 to 0.086)
Total anxiety score (RCADS-30) (Pupil)	12	7585	0.031	
	19	7175	0.028	(0.018 to 0.045)
	24	6987	0.028	(0.018 to 0.044)
Separation Anxiety Disorder subscale (RCADS-30) (Pupil)	12	7599	0.021	(0.013 to 0.034)
	19	7184	0.019	(0.011 to 0.032)
	24	6996	0.016	(0.010 to 0.027)
Generalised Anxiety Disorder subscale (RCADS-30) (Pupil)	12	7619	0.030	(0.019 to 0.045)
	19	7196	0.027	(0.017 to 0.043)
	24	7002	0.026	(0.016 to 0.041)
Panic Disorder subscale (RCADS-30) (Pupil)	12	7587	0.019	(0.012 to 0.031)
	19	7176	0.020	(0.012 to 0.034)
	24	6989	0.023	(0.014 to 0.037)
Social Anxiety subscale (RCADS-30) (Pupil)	12	7603	0.036	(0.024 to 0.054)
	19	7186	0.040	(0.026 to 0.060)
	24	6998	0.033	(0.021 to 0.050)
Obsessive Compulsive Disorder subscale (RCADS-30) (Pupil)	12	7606	0.017	(0.010 to 0.028)
	19	7191	0.018	(0.010 to 0.030)
	24	7001	0.017	(0.010 to 0.030)
Total score (SCCS) (Pupil)	12	7805	0.042	(0.027 to 0.064)
	19	7332	0.039	(0.025 to 0.059)
	24	7087	0.032	(0.020 to 0.050)

Outcome (measure) (reporter)	Measurement time (months)¹	N	ICC	95% CI
School leadership and student involvement subscale (SCCS) (Pupil)	12	7843	0.048	(0.031 to 0.073)
	19	7355	0.064	(0.044 to 0.092)
	24	7117	0.060	(0.041 to 0.087)
Respectful climate subscale (SCCS) (Pupil)	12	7838	0.050	(0.033 to 0.074)
	19	7346	0.045	(0.030 to 0.067)
	24	7109	0.031	(0.020 to 0.048)
Peer climate subscale (SCCS) (Pupil)	12	7826	0.060	(0.042 to 0.086)
	19	7343	0.059	(0.041 to 0.084)
	24	7104	0.049	(0.034 to 0.072)
Caring adults subscale (SCCS) (Pupil)	12	7812	0.035	(0.023 to 0.053)
	19	7337	0.032	(0.021 to 0.050)
	24	7094	0.029	(0.019 to 0.045)
Mindfulness (CAMM) (Pupil)	12	7924	0.019	(0.011 to 0.031)
	19	7472	0.024	(0.014 to 0.039)
	24	7171	0.020	(0.012 to 0.034)
Suicide ideation (Pupil)	12	6698	0.011	(0.005 to 0.023)
	19	6497	0.013	(0.007 to 0.026)
	24	6322	0.012	(0.006 to 0.025)
Self-harm (Pupil)	12	7232	0.006	(0.002 to 0.016)
	19	6820	0.011	(0.005 to 0.022)
	24	6598	0.005	(0.001 to 0.021)
Total difficulties score (SDQ) (Teacher)	12	5873	0.051	(0.026 to 0.097)
	19	5522	0.054	(0.026 to 0.110)
	24	4477	0.075	(0.041 to 0.132)
Emotional subscale (SDQ) (Teacher)	12	5873	0.059	(0.033 to 0.103)
	19	5522	0.043	(0.019 to 0.093)
	24	4477	0.051	(0.025 to 0.102)
Conduct subscale (SDQ) (Teacher)	12	5873	0.029	(0.014 to 0.061)
	19	5522	0.012	(0.002 to 0.062)

Outcome (measure) (reporter)	Measurement time (months)¹	N	ICC	95% CI
	24	4477	0.031	(0.014 to 0.066)
Hyperactivity subscale (SDQ) (Teacher)	12	5873	0.023	(0.009 to 0.057)
	19	5522	0.047	(0.025 to 0.086)
	24	4477	0.055	(0.030 to 0.099)
Peer subscale (SDQ) (Teacher)	12	5873	0.028	(0.012 to 0.064)
	19	5522	0.030	(0.011 to 0.078)
	24	4477	0.043	(0.019 to 0.092)
Prosocial subscale (SDQ) (Teacher)	12	5873	0.026	(0.009 to 0.074)
	19	5522	0.039	(0.017 to 0.086)
	24	4477	0.092	(0.055 to 0.152)
Executive Function (BRIEF-2) (Teacher)	12	5898	0.043	(0.019 to 0.092)
	19	5534	0.098	(0.057 to 0.162)
	24	4479	0.102	(0.060 to 0.168)

¹ Time points at 12, 19 and 24 months adjusted for trial arm status

References

1. Pocock SJ. *Clinical Trials - A Practical Approach*. John Wiley & Sons Ltd; 2013.
2. Campbell MJ, Walters S. *How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research*. Chichester: John Wiley and Sons; 2014.
3. Eldridge SM, Kerry S. *A Practical Guide to Cluster Randomised Trials in Health Services Research*. Chichester: John Wiley & Sons; 2012.
4. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. Chichester: Wiley; 2000.
5. Hayes R, Moulton L. *Cluster Randomised Trials*. Florida: CRC Press; 2009.
6. Murray DM. *Design and Analysis of Group-Randomized Trials*. New York: Oxford University Press; 1998.
7. Torgerson DJ. Contamination in trials: is cluster randomisation the answer? *Bmj*. 2001;322(7282):355-7.
8. Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology*. 2015;44(3):1051-67.
9. Campbell MK, Grimshaw JM. Cluster randomised trials: time for improvement. The implications of adopting a cluster design are still largely being ignored. *BMJ*. 1998;317(7167):1171-2.
10. Eldridge SM, Ukoumunne OC, Carlin JB. The intra-cluster correlation coefficient in cluster randomized trials: A review of definitions. *International Statistical Review*. 2009;77(3):378-94.
11. Kerry SM, Bland JM. The intracluster correlation coefficient in cluster randomisation. *BMJ*. 1998;316(7142):1455-60.
12. Cornfield J. Randomization by group: a formal analysis. *American Journal of Epidemiology*. 1978;108(2):100-2.
13. Gulliford MC, Adams G, Ukoumunne OC, Latinovic R, Chinn S, Campbell MJ. Intracluster correlation coefficient and outcome prevalence are associated in clustered binary data. *Journal of Clinical Epidemiology*. 2005;58(3):246-51.
14. Hayes R, Bennett S. Simple sample size calculation for cluster-randomized trials. *International journal of epidemiology*. 1999;28(2):319-26.
15. Spybrook J. Detecting Intervention Effects Across Context An Examination of the Precision of Cluster Randomized Trials. *The Journal of Experimental Education*. 2014;82(3):334-57.
16. Hemming K, Girling AJ, Sitch AJ, Marsh J, Lilford RJ. Sample size calculations for cluster randomised controlled trials with a fixed number of clusters. *BMC Medical Research Methodology*. 2011;11(1):102.
17. Donner A, Birkett N, Buck C. Randomization by cluster. Sample size requirements and analysis. *American Journal of Epidemiology*. 1981;114(6):906-14.
18. Ukoumunne O, Gulliford M, Chinn S, Sterne J, Burney P. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technology Assessment*. 1999;3(5).
19. Killip S, Mahfoud Z, Pearce K. What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *Annals of Family Medicine*. 2004;2(3):204-8.
20. Campbell MJ. Cluster randomized trials in general (family) practice research. *Statistical Methods in Medical Research*. 2000;9(2):81-94.
21. Kerry SM, Bland JM. Trials which randomize practices I: how should they be analysed? *Family Practice*. 1998;15(1):80-3.
22. Leyrat C, Morgan KE, Leurent B, Kahan BC. Cluster randomized trials with a small number of clusters: which analyses should be used? *International Journal of Epidemiology*. 2017;47(1):321-31.

23. Goldstein H, Browne W, Rasbash J. Multilevel modelling of medical data. *Statistics in medicine*. 2002;21(21):3291-315.
24. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13-22.
25. Ridout MS, Demétrio CG, Firth D. Estimating intraclass correlation for binary data. *Biometrics*. 1999;55(1):137-48.
26. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology*. 2004;57(8):785-94.
27. Wu S, Crespi CM, Wong WK. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contemp Clinical Trials*. 2012;33(5):869-80.
28. Giraudeau B, Ravaud P. Preventing bias in cluster randomised trials. *PLoS Med*. 2009;6(5):e1000065.
29. Eldridge S, Kerry S, Torgerson DJ. Bias in identifying and recruiting participants in cluster randomised trials: what can be done? *BMJ*. 2009;339:b4006.
30. Sim J, Dawson A. Informed consent and cluster-randomized trials. *Am J Public Health*. 2012;102(3):480-5.
31. Nix HP, Weijer C, Brehaut JC, Forster D, Goldstein CE, Taljaard M. Informed consent in cluster randomised trials: a guide for the perplexed. *BMJ Open*. 2021;11(9):e054213.
32. Gallo A, Weijer C, White A, Grimshaw JM, Boruch R, Brehaut JC, et al. What is the role and authority of gatekeepers in cluster randomized trials in health research? *Trials*. 2012;13(1):116.
33. Epstein DS, Enticott JC, Larson HJ, Barton C. Recruiting for research on sensitive topics in schools: an experience with Vaxcards, a collectable vaccine card game. *Trials*. 2021;22(1):320.
34. Eldridge SM, Ashby D, Feder GS. Informed patient consent to participation in cluster randomized trials: an empirical exploration of trials in primary care. *Clinical Trials*. 2005;2(2):91-8.
35. Taljaard M, Brehaut JC, Weijer C, Boruch R, Donner A, Eccles MP, et al. Variability in research ethics review of cluster randomized trials: a scenario-based survey in three countries. *Trials*. 2014;15(1):48.
36. Hewitt CE, Torgerson DJ. Is restricted randomisation necessary? *BMJ*. 2006;332(7556):1506-8.
37. Ivers NM, Halperin IJ, Barnsley J, Grimshaw JM, Shah BR, Tu K, et al. Allocation techniques for balance at baseline in cluster randomized trials: a methodological review. *Trials*. 2012;13(1):120.
38. Raab GM, Butcher I. Balance in cluster randomized trials. *Statistics in Medicine*. 2001;20(3):351-65.
39. Moulton LH. Covariate-based constrained randomization of group-randomized trials. *Clinical Trials*. 2004;1(3):297-305.
40. Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. Stratified randomization for clinical trials. *Journal of Clinical Epidemiology*. 1999;52(1):19-26.
41. Song JW, Chung KC. Observational studies: cohort and case-control studies. *Plastic and Reconstructive Surgery*. 2010;126(6):2234-42.
42. Ukoumunne OC, Thompson SG. Analysis of cluster randomized trials with repeated cross-sectional binary measurements. *Stat Med*. 2001;20(3):417-33.
43. Eldridge SM, Lancaster GA, Campbell MJ, Thabane L, Hopewell S, Coleman CL, et al. Defining Feasibility and Pilot Studies in Preparation for Randomised Controlled Trials: Development of a Conceptual Framework. *PLOS ONE*. 2016;11(3):e0150205.
44. Eldridge SM, Costelloe CE, Kahan BC, Lancaster GA, Kerry SM. How big should the pilot study for my cluster randomised trial be? *Statistical Methods in Medical Research*. 2016;25(3):1039-56.

45. United Nations. World Population Prospects 2022: Department of Economic and Social Affairs, Population Division; 2022 [Available from: <https://population.un.org/wpp/Download/Standard/Population/>].
46. HM Government. Childhood obesity: a plan for action 2019 [Available from: <https://www.gov.uk/government/publications/tackling-obesity-government-strategy>].
47. Costello EJ, Egger H, Angold A. 10-Year Research Update Review: The Epidemiology of Child and Adolescent Psychiatric Disorders: I. Methods and Public Health Burden. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2005;44(10):972-86.
48. Institute of Medicine and National Research Council. Investing in the Health and Well-Being of Young Adults. Washington (DC): The National Academies Press (US); 2015.
49. Joseph PD, Craig JC, Caldwell PH. Clinical trials in children. *British Journal of Clinical Pharmacology*. 2015;79(3):357-69.
50. Goesling B. A practical guide to cluster randomized trials in school health research. *Journal of School Health*. 2019;89(11):916-25.
51. Hemming K, Taljaard M, Moerbeek M, Forbes A. Contamination: How much can an individually randomized trial tolerate? *Statistics in Medicine*. 2021;40(14):3329-51.
52. Institute of Medicine (US) Committee on Clinical Research Involving Children. The Necessity and Challenges of Clinical Research Involving Children. Field MJ BR, editor. Washington (DC): National Academies Press (US); 2004.
53. Wallester S, Hill SR, Bero LA. Characteristics and quality of reporting of cluster randomized trials in children: reporting needs improvement. *Journal of Clinical Epidemiology*. 2011;64(12):1331-40.
54. Punch S. Research with Children: The Same or Different from Research with Adults? *Childhood*. 2002;9(3):321-41.
55. Langford R, Bonell CP, Jones HE, Poulidou T, Murphy SM, Waters E, et al. The WHO Health Promoting School framework for improving the health and well-being of students and their academic achievement. *Cochrane Database Syst Rev*. 2014(4):Cd008958.
56. Coyne I. Research with Children and Young People: The Issue of Parental (Proxy) Consent. *Children & Society*. 2010;24(3):227-37.
57. Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. *BMJ*. 2012;345:e5661.
58. Handlos LN, Chakraborty H, Sen PK. Evaluation of cluster-randomized trials on maternal and child health research in developing countries. *Tropical Medicine & International Health*. 2009;14(8):947-56.
59. Spybrook J, Zhang Q, Kelcey B, Dong N. Learning from cluster randomized trials in education: An assessment of the capacity of studies to determine what works, for whom, and under what conditions. *Educational Evaluation and Policy Analysis*. 2020;42(3):354-74.
60. Simons-Morton BG, Farhat T. Recent findings on peer group influences on adolescent smoking. *Journal of Prevention*. 2010;31(4):191-208.
61. Sellström E, Bremberg S. Is there a "school effect" on pupil outcomes? A review of multilevel studies. *Journal of epidemiology and community health*. 2006;60(2):149-55.
62. Markham WA, Aveyard P. A new theory of health promoting schools based on human functioning, school organisation and pedagogic practice. *Social Science & Medicine*. 2003;56(6):1209-20.
63. Macintyre S, Ellaway A, Cummins S. Place effects on health: how can we conceptualise, operationalise and measure them? *Social Science & Medicine*. 2002;55(1):125-39.
64. Lauder H, Kounali D, Robinson T, Goldstein H, Thrupp M, editors. *Social Class, Pupil Composition, Pupil Progress and School Performance: An Analysis of Primary Schools* 2008.
65. Harker R, Tymms P. The Effects of Student Composition on School Outcomes. *School Effectiveness and School Improvement*. 2004;15(2):177-99.
66. Cohen GL, Prinstein MJ. Peer contagion of aggression and health risk behavior among adolescent males: an experimental investigation of effects on public conduct and private attitudes. *Child Development*. 2006;77(4):967-83.

67. Moffitt TE. Adolescence-limited and life-course-persistent antisocial behavior: a developmental taxonomy. *Psychological Review*. 1993;100(4):674-701.
68. Bonell C, Jamal F, Harden A, Wells H, Parry W, Fletcher A. Systematic review of the effects of schools and school environment interventions on health: evidence mapping and synthesis. *Public Health Research*. 2013;1(1).
69. Dong N, Reinke WM, Herman KC, Bradshaw CP, Murray DW. Meaningful Effect Sizes, Intraclass Correlations, and Proportions of Variance Explained by Covariates for Planning Two- and Three-Level Cluster Randomized Trials of Social and Behavioral Outcomes. *Evaluation Review*. 2016;40(4):334-77.
70. Hale DR, Fitzgerald-Yau N, Viner RM. A systematic review of effective interventions for reducing multiple health risk behaviors in adolescence. *Am J Public Health*. 2014;104(5):e19-41.
71. Shackleton N, Hale D, Bonell C, Viner RM. Intraclass correlation values for adolescent health outcomes in secondary schools in 21 European countries. *SSM - Population Health*. 2016;2.
72. Stuart B, Becque T, Moore M, Little P. Clustering of continuous and binary outcomes at the general practice level in individually randomised studies in primary care - a review of 10 years of primary care trials. *BMC Medical Research Methodology*. 2020;20(1):83.
73. University of Aberdeen - Health Services Research Unit. Database of intra-correlation coefficients (ICCs) 2023 [Available from: <https://www.abdn.ac.uk/hsru/what-we-do/tools/>].
74. Bartlett R, Wright T, Olarinde T, Holmes T, Beamon ER, Wallace D. Schools as Sites for Recruiting Participants and Implementing Research. *Journal of Community Health Nursing*. 2017;34(2):80-8.
75. Lytle LA, Johnson CC, Bachman K, Wambsgans K, Perry CL, Stone EJ, et al. Successful recruitment strategies for school-based health promotion: experiences from CATCH. *Journal of School Health*. 1994;64(10):405-9.
76. Pound B, Riddell M, Byrnes G, Kelly H. Perception of social value predicts participation in school-based research. *Australian and New Zealand Journal of Public Health*. 2000;24(5):543-5.
77. Treweek S, Lockhart P, Pitkethly M, Cook JA, Kjeldstrøm M, Johansen M, et al. Methods to improve recruitment to randomised controlled trials: Cochrane systematic review and meta-analysis. *BMJ Open*. 2013;3(2):e002360.
78. Caldwell PHY, Hamilton S, Tan A, Craig JC. Strategies for Increasing Recruitment to Randomised Controlled Trials: Systematic Review. *PLOS Medicine*. 2010;7(11):e1000368.
79. Ross S, Grant A, Counsell C, Gillespie W, Russell I, Prescott R. Barriers to participation in randomised controlled trials: a systematic review. *Journal of Clinical Epidemiology*. 1999;52(12):1143-56.
80. Befort C, Lynch R, James RL, Carroll SL, Nollen N, Davis A. Perceived barriers and benefits to research participation among school administrators. *Journal of School Health*. 2008;78(11):581-6; quiz 615-7.
81. Aventin Á, Lohan M, Maguire L, Clarke M. Recruiting faith-and non-faith-based schools, adolescents and parents to a cluster randomised sexual-health trial: experiences, challenges and lessons from the mixed-methods Jack Feasibility Trial. *Trials*. 2016;17(1):1-13.
82. Felzmann H. Ethical issues in school-based research. *Research Ethics Review*. 2009;5(3):104-9.
83. Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials*. 2016;17:72-.
84. Torgerson CJ, Torgerson DJ. The Need for Randomised Controlled Trials in Educational Research. *British Journal of Educational Studies*. 2001;49(3):316-28.
85. Burnett C, Coldwell M. Randomised controlled trials and the interventionisation of education. *Oxford Review of Education*. 2021;47(4):423-38.
86. Li W, Konstantopoulos S. Power Analysis for Moderator Effects in Longitudinal Cluster Randomized Designs. *Educational and Psychological Measurement*. 2023;83(1):116-45.

87. Hedges LV, Schauer J. Randomised trials in education in the USA. *Educational Research*. 2018;60(3):265-75.
88. Hedges LV, Hedberg EC. Intraclass Correlation Values for Planning Group-Randomized Trials in Education. *Educational Evaluation and Policy Analysis*. 2007;29(1):60-87.
89. Stockford SM. Meta-analysis of intraclass correlation coefficients from multilevel models of educational achievement: Arizona State University; 2009.
90. Hedges LV, Hedberg EC. Intraclass Correlations and Covariate Outcome Correlations for Planning Two- and Three-Level Cluster-Randomized Experiments in Education. *Evaluation Review*. 2013;37(6):445-89.
91. Bosker RJ, Witziers B. A Meta Analytical Approach Regarding School Effectiveness: The True Size of School Effects and the Effect Size of Educational Leadership. U.S. Department of Education; 1995.
92. Zopluoglu C. A cross-national comparison of intra-class correlation coefficient in educational achievement outcomes. *Journal of Measurement and Evaluation in Education and Psychology*. 2012;3(1):242-78.
93. World Economic Forum. Global Gender Gap Report 2020 2020 [Available from: <https://www.weforum.org/reports/gender-gap-2020-report-100-years-pay-equality/infographics>].
94. Crocetti MT, Amin DD, Scherer R. Assessment of risk of bias among pediatric randomized controlled trials. *Pediatrics*. 2010;126(2):298-305.
95. Thomson D, Hartling L, Cohen E, Vandermeer B, Tjosvold L, Klassen TP. Controlled trials in children: quantity, methodological quality and descriptive characteristics of pediatric controlled trials published 1948-2006. *PLoS One*. 2010;5(9).
96. Hedberg EC. Academic and Behavioral Design Parameters for Cluster Randomized Trials in Kindergarten: An Analysis of the Early Childhood Longitudinal Study 2011 Kindergarten Cohort (ECLS-K 2011). *Evaluation Review*. 2016;40(4):279-313.
97. Murray DM, Hannan PJ. Planning for the appropriate analysis in school-based drug-use prevention studies. *Journal of Consulting and Clinical Psychology*. 1990;58(4):458-68.
98. Murray DM, Rooney BL, Hannan PJ, Peterson AV, Ary DV, Biglan A, et al. Intraclass correlation among common measures of adolescent smoking: estimates, correlates, and applications in smoking prevention studies. *American Journal of Epidemiology*. 1994;140(11):1038-50.
99. Murray DM, Short B. Intraclass correlation among measures related to alcohol use by young adults: estimates, correlates and applications in intervention studies. *Journal of Studies on Alcohol*. 1995;56(6):681-94.
100. Murray DM, Short BJ. Intraclass correlation among measures related to tobacco use by adolescents: estimates, correlates, and applications in intervention studies. *Addictive Behaviors*. 1997;22(1):1-12.
101. Siddiqui O, Hedeker D, Flay BR, Hu FB. Intraclass correlation estimates in a school-based smoking prevention study: outcome and mediating variables, by sex and ethnicity. *American Journal of Epidemiology*. 1996;144(4):425-33.
102. Murray DM, Clark M, Wagenaar AC. Intraclass correlations from a community-based alcohol prevention study: the effect of repeat observations on the same communities. *Journal of studies on alcohol*. 2000;61(6):881-90.
103. Ennett ST, Flewelling RL, Lindrooth RC, Norton EC. School and neighborhood characteristics associated with school rates of alcohol, cigarette, and marijuana use. *Journal of Health and Social Behavior*. 1997;38(1):55-71.
104. Resnicow K, Zhang N, Vaughan RD, Reddy SP, James S, Murray DM. When intraclass correlation coefficients go awry: a case study from a school-based smoking prevention study in South Africa. *Am J Public Health*. 2010;100(9):1714-8.
105. Murray DM, Phillips GA, Birnbaum AS, Lytle LA. Intraclass Correlation for Measures from a Middle School Nutrition Intervention Study: Estimates, Correlates, and Applications. *Health Education & Behavior*. 2001;28(6):666-79.

106. Juras R. Estimates of Intraclass Correlation Coefficients and Other Design Parameters for Studies of School-Based Nutritional Interventions. *Evaluation Review*. 2016;40(4):314-33.
107. Gray HL, Burgermaster M, Tipton E, Contento IR, Koch PA, Di Noia J. Intraclass Correlation Coefficients for Obesity Indicators and Energy Balance-Related Behaviors Among New York City Public Elementary Schools. *Health, Education and Behavior*. 2016;43(2):172-81.
108. Murray DM, Catellier DJ, Hannan PJ, Treuth MS, Stevens J, Schmitz KH, et al. School-level intraclass correlation for physical activity in adolescent girls. *Medicine & Science in Sports & Exercise*. 2004;36(5):876-82.
109. Murray DM, Stevens J, Hannan PJ, Catellier DJ, Schmitz KH, Dowda M, et al. School-level intraclass correlation for physical activity in sixth grade girls. *Medicine & Science in Sports & Exercise*. 2006;38(5):926-36.
110. Diaz-Ordaz K, Froud R, Sheehan B, Eldridge S. A systematic review of cluster randomised trials in residential facilities for older people suggests how to improve quality. *BMC medical research methodology*. 2013;13(1):1-10.
111. Eldridge S, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clinical Trials*. 2004;1(1):80-90.
112. Froud R, Eldridge S, Diaz Ordaz K, Marinho VCC, Donner A. Quality of cluster randomized controlled trials in oral health: a systematic review of reports published between 2005 and 2009. *Community Dentistry and Oral Epidemiology*. 2012;40:3-14.
113. Murray DM, Pals SL, George SM, Kuzmichev A, Lai GY, Lee JA, et al. Design and analysis of group-randomized trials in cancer: A review of current practices. *Prev Med*. 2018;111:241-7.
114. Lee YL, Lim YMF, Law KB, Sivasampu S. Intra-cluster correlation coefficients in primary care patients with type 2 diabetes and hypertension. *Trials*. 2020;21(1):530.
115. Bonell C, Humphrey N, Fletcher A, Moore L, Anderson R, Campbell R. Why schools should promote students' health and wellbeing. *BMJ*. 2014;348:g3078.
116. Parker K, Nunns MP, Xiao Z, Ford T, Ukoumunne OC. Characteristics and practices of school-based cluster randomised controlled trials for improving health outcomes in pupils in the UK: a systematic review protocol. *BMJ Open*. 2021;11(2):e044143.
117. Parker K, Nunns M, Xiao Z, Ford T, Ukoumunne OC. Characteristics and practices of school-based cluster randomised controlled trials for improving health outcomes in pupils in the United Kingdom: a methodological systematic review. *BMC Medical Research Methodology*. 2021;21(1):152.
118. Snyder H. Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*. 2019;104:333-9.
119. Clarke J. What is a systematic review? *Evidence Based Nursing*. 2011;14(3):64-.
120. Uman LS. Systematic reviews and meta-analyses. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*. 2011;20(1):57-9.
121. Turner RM, Thompson SG, Spiegelhalter DJ. Prior distributions for the intraclass correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clinical Trials*. 2005;2(2):108-18.
122. Parker K, Eddy S, Nunns M, Xiao Z, Ford T, Eldridge S, et al. Systematic review of the characteristics of school-based feasibility cluster randomised trials of interventions for improving the health of pupils in the UK. *Pilot and Feasibility Studies*. 2022;8(1):132.
123. Parker K, Nunns M, Xiao Z, Ford T, Ukoumunne OC. Intraclass correlation coefficients from school-based cluster randomized trials of interventions for improving health outcomes in pupils. *Journal of Clinical Epidemiology*. 2023;158:18-26.
124. Cohen J, Onunaku N, Clothier S, Poppe J. Helping young children succeed: Strategies to promote early childhood social and emotional development. 2005.
125. Moher D, Liberati A, Tetzlaff J, Altman DG, The PG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine*. 2009;6(7):e1000097.

126. Higgins JPT TJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA. Cochrane Handbook for Systematic Reviews of Interventions version 6.2 Cochrane; 2021 [Available from: www.training.cochrane.org/handbook].
127. Taljaard M, McGowan J, Grimshaw J, Brehaut J, McRae A, Eccles M, et al. Electronic search strategies to identify reports of cluster randomized trials in MEDLINE: Low precision will improve with adherence to reporting standards. *BMC Medical Research Methodology*. 2010;10:15.
128. Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. *Systematic Reviews*. 2017;6(1):245.
129. Davies KS. Formulating the evidence based practice question: a review of the frameworks. *Evidence Based Library and Information Practice*. 2011;6(2):75-80.
130. Department of Education. Early education and childcare - Statutory guidance for local authorities 2018 [Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/718179/Early_education_and_childcare-statutory_guidance.pdf].
131. Dodge KA. Annual Research Review: Universal and targeted strategies for assigning interventions to achieve population impact. *Journal of Child Psychology and Psychiatry*. 2020;61(3):255-67.
132. Campbell MJ, Lancaster GA, Eldridge SM. A randomised controlled trial is not a pilot trial simply because it uses a surrogate endpoint. *Pilot and Feasibility Studies*. 2018;4(1):130.
133. The EndNote Team. EndNote. EndNote X9 version ed. Philadelphia, PA: Clarivate; 2013.
134. HM Government. Types of School n.d. [Available from: <https://www.gov.uk/types-of-school>].
135. RAF Association. UK Comparison Table of School Year Groups across the UK (April 2020) 2022 [Available from: <https://www.raf-ff.org.uk/wp-content/uploads/2020/06/UK-school-year-comparison-table-2020-plus-devolved-state-edu-comparison-table.pdf>].
136. Cuijpers P. Examining the effects of prevention programs on the incidence of new cases of mental disorders: the lack of statistical power. *American Journal of Psychiatry*. 2003;160(8):1385-91.
137. Smith PG, Morrow RH, Ross DA. Wellcome Trust–Funded Monographs and Book Chapters. In: Smith PG, Morrow RH, Ross DA, editors. *Field Trials of Health Interventions: A Toolbox*. Oxford (UK): OUP Oxford London School of Hygiene and Tropical Medicine 2015.; 2015.
138. Dale A, Marsh C. *The 1991 Census User's Guide*. London: HM Stationery Office; 1993.
139. HM Government. *The English Indices of Deprivation 2019*. 2019 [Available from: <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019>].
140. Roach KE. Measurement of Health Outcomes: Reliability, Validity and Responsiveness. *JPO: Journal of Prosthetics and Orthotics*. 2006;18(6):P8-P12.
141. Range L, Embry T, MacLeod T. Active and passive consent: a comparison of actual research with children. *Ethical Human Sciences and Services*. 2001;3(1):23-31.
142. StataCorp. *Stata. Stata Statistical Software: Release 16 ed*. College Station, TX: StataCorp LLC; 2019.
143. Adab P, Pallan MJ, Lancashire ER, Hemming K, Frew E, Barrett T, et al. Effectiveness of a childhood obesity prevention programme delivered through schools, targeting 6 and 7 year olds: cluster randomised controlled trial (WAVES study). *BMJ*. 2018;360:k211.
144. Aveyard P, Cheng K, Almond J, Sherratt E, Lancashire R, Lawrence T, et al. Cluster randomised controlled trial of expert system based on the transtheoretical (“stages of change”) model for smoking prevention and cessation in schools. *BMJ*. 1999;319(7215):948-53.
145. Axford N, Bjornstad G, Clarkson S, Ukoumunne OC, Wrigley Z, Matthews J, et al. The Effectiveness of the KiVa Bullying Prevention Program in Wales, UK: Results from a Pragmatic Cluster Randomized Controlled Trial. *Prevention Science*. 2020;21(5):615-26.

146. Bonell C, Allen E, Warren E, McGowan J, Bevilacqua L, Jamal F, et al. Effects of the Learning Together intervention on bullying and aggression in English secondary schools (INCLUSIVE): a cluster randomised controlled trial. *The Lancet*. 2018;392(10163):2452-64.
147. Breheny K, Passmore S, Adab P, Martin J, Hemming K, Lancashire ER, et al. Effectiveness and cost-effectiveness of The Daily Mile on childhood weight outcomes and wellbeing: a cluster randomised controlled trial. *International Journal of Obesity*. 2020;44(4):812-22.
148. Breslin G, Shannon S, Rafferty R, Fitzpatrick B, Belton S, O'Brien W, et al. The effect of sport for LIFE: all island in children from low socio-economic status: a clustered randomized controlled trial. *Health and Quality of Life Outcomes*. 2019;17(1):1-12.
149. Campbell R, Starkey F, Holliday J, Audrey S, Bloor M, Parry-Langdon N, et al. An informal school-based peer-led intervention for smoking prevention in adolescence (ASSIST): a cluster randomised trial. *The Lancet*. 2008;371(9624):1595-602.
150. Chisholm K, Patterson P, Torgerson C, Turner E, Jenkinson D, Birchwood M. Impact of contact on adolescents' mental health literacy and stigma: the SchoolSpace cluster randomised controlled trial. *BMJ Open*. 2016;6(2):e009435.
151. Christian MS, Evans CE, Nykjaer C, Hancock N, Cade JE. Evaluation of the impact of a school gardening intervention on children's fruit and vegetable intake: a randomised controlled trial. *International Journal of Behavioral Nutrition and Physical Activity*. 2014;11(1):1-15.
152. Conner M, Grogan S, West R, Simms-Ellis R, Scholtens K, Sykes-Muskett B, et al. Effectiveness and cost-effectiveness of repeated implementation intention formation on adolescent smoking initiation: A cluster randomized controlled trial. *Journal of Consulting and Clinical Psychology*. 2019;87(5):422.
153. Connolly P, Miller S, Kee F, Sloan S, Gildea A, McIntosh E, et al. A cluster randomised controlled trial and evaluation and cost-effectiveness analysis of the Roots of Empathy schools-based programme for improving social and emotional well-being outcomes among 8-to 9-year-olds in Northern Ireland. *Public Health Research*. 2018;6(4).
154. Conrod PJ, O'Leary-Barrett M, Newton N, Topper L, Castellanos-Ryan N, Mackie C, et al. Effectiveness of a selective, personality-targeted prevention program for adolescent alcohol use and misuse: a cluster randomized controlled trial. *JAMA Psychiatry*. 2013;70(3):334-42.
155. Croker H, Lucas R, Wardle J. Cluster-randomised trial to evaluate the 'Change for Life' mass media/social marketing campaign in the UK. *BMC Public Health*. 2012;12(1):1-14.
156. Cunningham CJ, Elton R, Topping GV. A randomised control trial of the effectiveness of personalised letters sent subsequent to school dental inspections in increasing registration in unregistered children. *BMC Oral Health*. 2009;9(1):1-8.
157. Diedrichs PC, Atkinson MJ, Steer RJ, Garbett KM, Rumsey N, Halliwell E. Effectiveness of a brief school-based body image intervention 'Dove Confident Me: Single Session' when delivered by teachers and researchers: Results from a cluster randomised controlled trial. *Behaviour Research and Therapy*. 2015;74:94-104.
158. Evans C, Greenwood DC, Thomas JD, Cleghorn CL, Kitchen MS, Cade JE. SMART lunch box intervention to improve the food and nutrient content of children's packed lunches: UK wide cluster randomised controlled trial. *Journal of Epidemiology & Community Health*. 2010;64(11):970-6.
159. Evans CE, Ransley JK, Christian MS, Greenwood DC, Thomas JD, Cade JE. A cluster-randomised controlled trial of a school-based fruit and vegetable intervention: Project Tomato. *Public Health Nutrition*. 2013;16(6):1073-81.
160. Fairclough SJ, Hackett AF, Davies IG, Gobbi R, Mackintosh KA, Warburton GL, et al. Promoting healthy weight in primary school children through physical activity and nutrition education: a pragmatic evaluation of the CHANGE! randomised intervention study. *BMC Public Health*. 2013;13(1):1-14.
161. Ford T, Hayes R, Byford S, Edwards V, Fletcher M, Logan S, et al. The effectiveness and cost-effectiveness of the Incredible Years® Teacher Classroom Management programme in

- primary school children: results of the STARS cluster randomised controlled trial. *Psychological Medicine*. 2019;49(5):828-42.
162. Foulkes J, Knowles Z, Fairclough S, Stratton G, O'Dwyer M, Ridgers N, et al. Effect of a 6-week active play intervention on fundamental movement skill competence of preschool children: a cluster randomized controlled trial. *Perceptual and Motor Skills*. 2017;124(2):393-412.
163. Giles M, McClenahan C, Armour C, Millar S, Rae G, Mallett J, et al. Evaluation of a theory of planned behaviour-based breastfeeding intervention in Northern Irish schools using a randomized cluster design. *British Journal of Health Psychology*. 2014;19(1):16-35.
164. Graham A, Moore L, Sharp D, Diamond I. Improving teenagers' knowledge of emergency contraception: cluster randomised controlled trial of a teacher led intervention. *BMJ*. 2002;324(7347):1179.
165. Griffin TL, Jackson DM, McNeill G, Aucott LS, MacDiarmid JI. A brief educational intervention increases knowledge of the sugar content of foods and drinks but does not decrease intakes in scottish children aged 10–12 years. *Journal of Nutrition Education and Behavior*. 2015;47(4):367-73.
166. Hardman M, Davies G, Duxbury J, Davies R. A cluster randomised controlled trial to evaluate the effectiveness of fluoride varnish as a public health measure to reduce caries in children. *Caries Research*. 2007;41(5):371-6.
167. Harrington DM, Davies MJ, Bodicoat DH, Charles JM, Chudasama YV, Gorely T, et al. Effectiveness of the 'Girls Active' school-based physical activity programme: A cluster randomised controlled trial. *International Journal of Behavioral Nutrition and Physical Activity*. 2018;15(1):1-18.
168. Henderson M, Wight D, Raab G, Abraham C, Parkes A, Scott S, et al. Impact of a theoretically based sex education programme (SHARE) delivered by teachers on NHS registered conceptions and terminations: final results of cluster randomised trial. *BMJ*. 2007;334(7585):133.
169. Hislop MD, Stokes KA, Williams S, McKay CD, England ME, Kemp SP, et al. Reducing musculoskeletal injury and concussion risk in schoolboy rugby players with a pre-activity movement control exercise programme: a cluster randomised controlled trial. *British Journal of Sports Medicine*. 2017;51(15):1140-6.
170. Hodgkinson A, Abbott J, Hurley MA, Lowe N, Qualter P. An educational intervention to prevent overweight in pre-school years: a cluster randomised trial with a focus on disadvantaged families. *BMC Public Health*. 2019;19(1):1-13.
171. Howlin P, Gordon RK, Pasco G, Wade A, Charman T. The effectiveness of Picture Exchange Communication System (PECS) training for teachers of children with autism: a pragmatic, group randomised controlled trial. *Journal of Child Psychology and Psychiatry*. 2007;48(5):473-81.
172. Hubbard G, Stoddart I, Forbat L, Neal RD, O'Carroll RE, Haw S, et al. School - based brief psycho - educational intervention to raise adolescent cancer awareness and address barriers to medical help - seeking about cancer: a cluster randomised controlled trial. *Psycho - Oncology*. 2016;25(7):760-71.
173. Humphrey N, Barlow A, Wigelsworth M, Lendrum A, Pert K, Joyce C, et al. A cluster randomized controlled trial of the Promoting Alternative Thinking Strategies (PATHS) curriculum. *Journal of School Psychology*. 2016;58:73-89.
174. Jago R, Edwards MJ, Sebire SJ, Tomkinson K, Bird EL, Banfield K, et al. Effect and cost of an after-school dance programme on the physical activity of 11–12 year old girls: The Bristol Girls Dance Project, a school-based cluster randomised controlled trial. *International Journal of Behavioral Nutrition and Physical Activity*. 2015;12(1):1-15.
175. James J, Thomas P, Cavan D, Kerr D. Preventing childhood obesity by reducing consumption of carbonated drinks: cluster randomised controlled trial. *BMJ*. 2004;328(7450):1237.

176. Kendrick D, Groom L, Stewart J, Watson M, Mulvaney C, Casterton R. "Risk Watch": Cluster randomised controlled trial evaluating an injury prevention program. *Injury Prevention*. 2007;13(2):93-9.
177. Kendrick D, Royal S. Cycle helmet ownership and use; a cluster randomised controlled trial in primary school children in deprived areas. *Archives of Disease in Childhood*. 2004;89(4):330-5.
178. Kipping RR, Howe LD, Jago R, Campbell R, Wells S, Chittleborough CR, et al. Effect of intervention aimed at increasing physical activity, reducing sedentary behaviour, and increasing fruit and vegetable consumption in children: active for Life Year 5 (AFLY5) school based cluster randomised controlled trial. *BMJ*. 2014;348:g3256.
179. Lakshman RR, Sharp SJ, Ong KK, Forouhi NG. A novel school-based intervention to improve nutrition knowledge in children: cluster randomised controlled trial. *BMC Public Health*. 2010;10:123.
180. Lloyd J, Creanor S, Logan S, Green C, Dean SG, Hillsdon M, et al. Effectiveness of the Healthy Lifestyles Programme (HeLP) to prevent obesity in UK primary-school children: a cluster randomised controlled trial. *The Lancet Child & Adolescent Health*. 2018;2(1):35-45.
181. Marcano-Olivier M, Pearson R, Ruparell A, Horne PJ, Viktor S, Erjavec M. A low-cost Behavioural Nudge and choice architecture intervention targeting school lunches increases children's consumption of fruit: a cluster randomised trial. *International Journal of Behavioral Nutrition and Physical Activity*. 2019;16(1):1-9.
182. Markham WA, Bridle C, Grimshaw G, Stanton A, Aveyard P. Trial protocol and preliminary results for a cluster randomised trial of behavioural support versus brief advice for smoking cessation in adolescents. *BMC Research Notes*. 2010;3(1):1-10.
183. McKay M, Agus A, Cole J, Doherty P, Foxcroft D, Harvey S, et al. Steps Towards Alcohol Misuse Prevention Programme (STAMPP): a school-based and community-based cluster randomised controlled trial. *BMJ Open*. 2018;8(3):e019722.
184. Milsom K, Blinkhorn A, Walsh T, Worthington H, Kearney-Mitchell P, Whitehead H, et al. A cluster-randomized controlled trial: fluoride varnish in school children. *Journal of Dental Research*. 2011;90(11):1306-11.
185. Milsom K, Blinkhorn A, Worthington H, Threlfall A, Buchanan K, Kearney-Mitchell P, et al. The effectiveness of school dental screening: a cluster-randomized control trial. *Journal of Dental Research*. 2006;85(10):924-8.
186. Moore L, Tapper K. The impact of school fruit tuck shops and school food policies on children's fruit consumption: a cluster randomised trial of schools in deprived areas. *Journal of Epidemiology & Community Health*. 2008;62(10):926-31.
187. Mulvaney CA, Kendrick D, Watson MC, Coupland CA. Increasing child pedestrian and cyclist visibility: cluster randomised controlled trial. *Journal of Epidemiology & Community Health*. 2006;60(4):311-5.
188. Murphy S, Moore G, Tapper K, Lynch R, Clarke R, Raisanen L, et al. Free healthy breakfasts in primary schools: a cluster randomised controlled trial of a policy intervention in Wales, UK. *Public Health Nutrition*. 2011;14(2):219-26.
189. Norris E, Dunsmuir S, Duke-Williams O, Stamatakis E, Shelton N. Physically active lessons improve lesson activity and on-task behavior: A cluster-randomized controlled trial of the "Virtual Traveller" Intervention. *Health Education & Behavior*. 2018;45(6):945-56.
190. Nutbeam D, Macaskill P, Smith C, Simpson JM, Catford J. Evaluation of two school smoking education programmes under normal classroom conditions. *BMJ*. 1993;306(6870):102-7.
191. Obsuth I, Sutherland A, Cope A, Pilbeam L, Murray AL, Eisner M. London Education and Inclusion Project (LEIP): Results from a cluster-randomized controlled trial of an intervention to reduce school exclusion and antisocial behavior. *Journal of Youth and Adolescence*. 2017;46(3):538-57.
192. Patterson E, Brennan M, Linskey K, Webb D, Shields M, Patterson C. A cluster randomised intervention trial of asthma clubs to improve quality of life in primary school

- children: the School Care and Asthma Management Project (SCAMP). *Archives of Disease in Childhood*. 2005;90(8):786-91.
193. Pine C, McGoldrick P, Burnside G, Curnow M, Chesters R, Nicholson J, et al. An intervention programme to establish regular toothbrushing: understanding parents' beliefs and motivating children. *International Dental Journal*. 2000;50(6):312-23.
194. Redmond CA, Blinkhorn FA, Kay EJ, Davies RM, Worthington HV, Blinkhorn AS. A cluster randomized controlled trial testing the effectiveness of a school - based dental health education program for adolescents. *Journal of Public Health Dentistry*. 1999;59(1):12-7.
195. Rees G, Bakhshi S, Surujlal-Harry A, Stasinopoulos M, Baker A. A computerised tailored intervention for increasing intakes of fruit, vegetables, brown bread and wholegrain cereals in adolescent girls. *Public Health Nutrition*. 2010;13(8):1271-8.
196. Rowland D, DiGiuseppi C, Gross M, Afolabi E, Roberts I. Randomised controlled trial of site specific advice on school travel patterns. *Archives of Disease in Childhood*. 2003;88(1):8-11.
197. Sahota P, Rudolf MC, Dixey R, Hill AJ, Barth JH, Cade J. Randomised controlled trial of primary school based intervention to reduce risk factors for obesity. *BMJ*. 2001;323(7320):1029.
198. Sayal K, Taylor JA, Valentine A, Guo B, Sampson CJ, Sellman E, et al. Effectiveness and cost - effectiveness of a brief school - based group programme for parents of children at risk of ADHD: a cluster randomised controlled trial. *Child: Care, Health and Development*. 2016;42(4):521-33.
199. Scott S, O' Connor TG, Futh A, Matias C, Price J, Doolan M. Impact of a parenting program in a high - risk, multi - ethnic community: The PALS trial. *Journal of Child Psychology and Psychiatry*. 2010;51(12):1331-41.
200. Sharpe H, Patalay P, Vostanis P, Belsky J, Humphrey N, Wolpert M. Use, acceptability and impact of booklets designed to support mental health self-management and help seeking in schools: results of a large randomised controlled trial in England. *European Child & Adolescent Psychiatry*. 2017;26(3):315-24.
201. Shemilt I, Harvey I, Shepstone L, Swift L, Reading R, Mugford M, et al. A national evaluation of school breakfast clubs: evidence from a cluster randomized controlled trial and an observational analysis. *Child: Care, Health and Development*. 2004;30(5):413-27.
202. Stallard P, Sayal K, Phillips R, Taylor JA, Spears M, Anderson R, et al. Classroom based cognitive behavioural therapy in reducing symptoms of depression in high risk adolescents: pragmatic cluster randomised controlled trial. *BMJ*. 2012;345:e6058.
203. Stallard P, Skryabina E, Taylor G, Phillips R, Daniels H, Anderson R, et al. Classroom-based cognitive behaviour therapy (FRIENDS): a cluster randomised controlled trial to Prevent Anxiety in Children through Education in Schools (PACES). *The Lancet Psychiatry*. 2014;1(3):185-92.
204. Stephenson J, Strange V, Forrest S, Oakley A, Copas A, Allen E, et al. Pupil-led sex education in England (RIPPLE study): cluster-randomised intervention trial. *The Lancet*. 2004;364(9431):338-46.
205. Tymms PB, Curtis SE, Routen AC, Thomson KH, Bolden DS, Bock S, et al. Clustered randomised controlled trial of two education interventions designed to increase physical activity and well-being of secondary school students: the MOVE Project. *BMJ Open*. 2016;6(1).
206. Worthington HV, Hill KB, Mooney J, Hamilton FA, Blinkhorn AS. A cluster randomized controlled trial of a dental health education program for 10 - year - old children. *Journal of Public Health Dentistry*. 2001;61(1):22-7.
207. Bolzern J, Mnyama N, Bosanquet K, Torgerson DJ. A review of cluster randomized trials found statistical evidence of selection bias. *Journal of Clinical Epidemiology*. 2018;99:106-12.
208. Story M, Kaphingst KM, French S. The Role of Schools in Obesity Prevention. *The Future of Children*. 2006;16(1):109-42.

209. Rutterford C, Taljaard M, Dixon S, Copas A, Eldridge S. Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review. *Journal of Clinical Epidemiology*. 2015;68(6):716-23.
210. Díaz-Ordaz K, Kenward MG, Cohen A, Coleman CL, Eldridge S. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. *Clinical Trials*. 2014;11(5):590-600.
211. Kelcey B, Shen Z, Spybrook J. Intraclass Correlation Coefficients for Designing Cluster-Randomized Trials in Sub-Saharan Africa Education. *Evaluation Review*. 2016;40(6):500-25.
212. HM Government. Schools, pupils and their characteristics: January 2019. 2019 [Available from: <https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2019>].
213. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. *Am J Public Health*. 2004;94(3):423-32.
214. Ivers N, Taljaard M, Dixon S, Bennett C, McRae A, Taleban J, et al. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. *BMJ*. 2011;343.
215. HM Government. Get information about schools n.d. [Available from: <https://www.gov.uk/government/organisations/department-for-education>].
216. de Hoop E, Teerenstra S, van Gaal BG, Moerbeek M, Borm GF. The “best balance” allocation led to optimal balance in cluster-controlled trials. *Journal of Clinical Epidemiology*. 2012;65(2):132-7.
217. Lancaster GA, Thabane L. Guidelines for reporting non-randomised pilot and feasibility studies. *Pilot and Feasibility Studies*. 2019;5(1):114.
218. Fazzari MJ, Kim MY, Heo M. Sample size determination for three-level randomized clinical trials with randomization at the first or second level. *Journal of Biopharmaceutical Statistics*. 2014;24(3):579-99.
219. Billingham SA, Whitehead AL, Julious SA. An audit of sample sizes for pilot and feasibility trials being undertaken in the United Kingdom registered in the United Kingdom Clinical Research Network database. *BMC Medical Research Methodology*. 2013;13(1):1-6.
220. Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, et al. A tutorial on pilot studies: the what, why and how. *BMC Medical Research Methodology*. 2010;10(1):1-10.
221. Chan CL, Leyrat C, Eldridge SM. Quality of reporting of pilot and feasibility cluster randomised trials: a systematic review. *BMJ Open*. 2017;7(11):e016970.
222. Kristunas CA, Hemming K, Eborall H, Eldridge S, Gray LJ. The current use of feasibility studies in the assessment of feasibility for stepped-wedge cluster randomised trials: a systematic review. *BMC Medical Research Methodology*. 2019;19(1):12.
223. Avery KNL, Williamson PR, Gamble C, O’Connell Francischetto E, Metcalfe C, Davidson P, et al. Informing efficient randomised controlled trials: exploration of challenges in developing progression criteria for internal pilot studies. *BMJ Open*. 2017;7(2):e013537.
224. Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. *Journal of Evaluation in Clinical Practice*. 2004;10(2):307-12.
225. BioMed Central. ISRCTN Registry: Springer Nature; 2022 [Available from: https://www.isrctn.com/?gclid=EAlaIqObChMIwv7i-rj19wIVg7LVCh3Ncgl3EAAYASAAEgJmjPD_BwE].
226. Thabane L, Hopewell S, Lancaster GA, Bond CM, Coleman CL, Campbell MJ, et al. Methods and processes for development of a CONSORT extension for reporting pilot randomized controlled trials. *Pilot Feasibility Studies*. 2016;2:25.
227. StataCorp. Stata. Release 17. College Station, TX: StataCorp LLC; 2021.
228. Barber SE, Jackson C, Hewitt C, Ainsworth HR, Buckley H, Akhtar S, et al. Assessing the feasibility of evaluating and delivering a physical activity intervention for pre-school children: a pilot randomised controlled trial. *Pilot and Feasibility Studies*. 2016;2(1):12.

229. Bonell C, Fletcher A, Fitzgerald-Yau N, Hale D, Allen E, Elbourne D, et al. Initiating change locally in bullying and aggression through the school environment (INCLUSIVE): a pilot randomised controlled trial. *Health Technol Assessment*. 2015;19(53).
230. Carlin A, Murphy MH, Nevill A, Gallagher AM. Effects of a peer-led Walking In Schools intervention (the WISH study) on physical activity levels of adolescent girls: a cluster randomised pilot study. *Trials*. 2018;19(1):31.
231. Clemes SA, Bingham DD, Pearson N, Chen Y-L, Edwardson CL, McEachan RRC, et al. Stand Out in Class: restructuring the classroom environment to reduce sitting time – findings from a pilot cluster randomised controlled trial. *International Journal of Behavioral Nutrition and Physical Activity*. 2020;17(1):55.
232. Corder K, Brown HE, Schiff A, van Sluijs EMF. Feasibility study and pilot cluster-randomised controlled trial of the GoActive intervention aiming to promote physical activity among adolescents: outcomes and lessons learnt. *BMJ Open*. 2016;6(11):e012335.
233. Corepal R, Best P, O’Neill R, Kee F, Badham J, Dunne L, et al. A feasibility study of ‘The StepSmart Challenge’ to promote physical activity in adolescents. *Pilot and Feasibility Studies*. 2019;5(1):132.
234. Forster AS, Cornelius V, Rockliffe L, Marlow LA, Bedford H, Waller J. A cluster randomised feasibility study of an adolescent incentive intervention to increase uptake of HPV vaccination. *British Journal of Cancer*. 2017;117(8):1121-7.
235. Gammon C, Morton K, Atkin A, Corder K, Daly-Smith A, Quarmby T, et al. Introducing physically active lessons in UK secondary schools: feasibility study and pilot cluster-randomised controlled trial. *BMJ Open*. 2019;9(5):e025080.
236. Ginja S, Arnott B, Araujo-Soares V, Namdeo A, McColl E. Feasibility of an incentive scheme to promote active travel to school: a pilot cluster randomised trial. *Pilot and Feasibility Studies*. 2017;3(1):57.
237. Jago R, Sebire SJ, Cooper AR, Haase AM, Powell J, Davis L, et al. Bristol Girls Dance Project Feasibility Trial: outcome and process evaluation results. *International Journal of Behavioral Nutrition and Physical Activity*. 2012;9(1):83.
238. Jago R, Sebire SJ, Davies B, Wood L, Edwards MJ, Banfield K, et al. Randomised feasibility trial of a teaching assistant led extracurricular physical activity intervention for 9 to 11 year olds: Action 3:30. *International Journal of Behavioral Nutrition and Physical Activity*. 2014;11:114.
239. Johnstone A, Hughes AR, Bonnar L, Booth JN, Reilly JJ. An active play intervention to improve physical activity and fundamental movement skills in children of low socio-economic status: feasibility cluster randomised controlled trial. *Pilot and Feasibility Studies*. 2019;5(1):45.
240. Kipping RR, Payne C, Lawlor DA. Randomised controlled trial adapting US school obesity prevention to England. *Archives of Disease in Childhood*. 2008;93(6):469-73.
241. Lloyd JJ, Wyatt KM, Creanor S. Behavioural and weight status outcomes from an exploratory trial of the Healthy Lifestyles Programme (HeLP): a novel school-based obesity prevention programme. *BMJ Open*. 2012;2(3):e000390.
242. Lohan M, Aventin Á, Clarke M, Curran RM, McDowell C, Agus A, et al. Can Teenage Men Be Targeted to Prevent Teenage Pregnancy? A Feasibility Cluster Randomised Controlled Intervention Trial in Schools. *Prevention Science*. 2018;19(8):1079-90.
243. McSweeney L, Araújo-Soares V, Rapley T, Adamson A. A feasibility study with process evaluation of a preschool intervention to improve child and family lifestyle behaviours. *BMC Public Health*. 2017;17(1):248.
244. Meiksin R, Crichton J, Dodd M, Morgan GS, Williams P, Willmott M, et al. A school intervention for 13- to 15-year-olds to prevent dating and relationship violence: the Project Respect pilot cluster RCT. *Public Health Research*. 2020;8(5).
245. Newbury-Birch D, Scott S, O’Donnell A, Coulton S, Howel D, McColl E, et al. A pilot feasibility cluster randomised controlled trial of screening and brief alcohol intervention to prevent hazardous drinking in young people aged 14–15 years in a high school setting (SIPS JR-HIGH). *Public Health Research*. 2014;2(6).

246. Sahota P, Christian M, Day R, Cocks K. The feasibility and acceptability of a primary school-based programme targeting diet and physical activity: the PhunkyFoods Programme. *Pilot and Feasibility Studies*. 2019;5(1):152.
247. Sebire SJ, Jago R, Banfield K, Edwards MJ, Campbell R, Kipping R, et al. Results of a feasibility cluster randomised controlled trial of a peer-led school-based intervention to increase the physical activity of adolescent girls (PLAN-A). *International Journal of Behavioral Nutrition and Physical Activity*. 2018;15(1):50.
248. Segrott J, Rothwell H, Hewitt G, Playle R, Huang C, Murphy S, et al. Preventing alcohol misuse in young people: an exploratory cluster randomised controlled trial of the Kids, Adults Together (KAT) programme. *Public Health Research*. 2015;3(15).
249. Sharpe H, Schober I, Treasure J, Schmidt U. Feasibility, acceptability and efficacy of a school-based prevention programme for eating disorders: cluster randomised controlled trial. *British Journal of Psychiatry*. 2013;203(6):428-35.
250. White J, Hawkins J, Madden K, Grant A, Er V, Angel L, et al. Adapting the ASSIST model of informal peer-led intervention delivery to the Talk to FRANK drug prevention programme in UK secondary schools (ASSIST + FRANK): intervention development, refinement and a pilot cluster randomised controlled trial. *Public Health Research*. 2017;5(7).
251. Wright B, Marshall D, Adamson J, Ainsworth H, Ali S, Allgar V, et al. Social Stories™ to alleviate challenging behaviour and social difficulties exhibited by children with autism spectrum disorder in mainstream schools: design of a manualised training toolkit and feasibility study for a cluster randomised controlled trial with nested qualitative and cost-effectiveness components. *Health Technol Assessment*. 2016;20(6).
252. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ*. 2008;337:a1655.
253. Guthold R, Stevens GA, Riley LM, Bull FC. Global trends in insufficient physical activity among adolescents: a pooled analysis of 298 population-based surveys with 1·6 million participants. *The Lancet Child & Adolescent Health*. 2020;4(1):23-35.
254. Sadler K, Vizard T, Ford T, Marcheselli F, Pearce N, Mandalia D, et al. Mental health of children and young people in England, 2017. Leeds, UK: NHS Digital; 2018.
255. Arain M, Campbell MJ, Cooper CL, Lancaster GA. What is a pilot or feasibility study? A review of current practice and editorial policy. *BMC Medical Research Methodology*. 2010;10(1):67.
256. Henderson M, Wight D, Nixon C, Hart G. Retaining young people in a longitudinal sexual health survey: a trial of strategies to maintain participation. *BMC Medical Research Methodology*. 2010;10(1):9.
257. Pound B, Riddell M, Byrnes G, Kelly H. Perception of social value predicts participation in school-based research. *Australian and New Zealand Journal of Public Health*. 2000;24(5):543-5.
258. Steenholt CB, Pisinger VSC, Danquah IH, Tolstrup JS. School and class-level variations and patterns of physical activity: a multilevel analysis of Danish high school students. *BMC Public Health*. 2018;18(1):255.
259. Hale DR, Patalay P, Fitzgerald-Yau N, Hargreaves DS, Bond L, Görzig A, et al. School-level variation in health outcomes in adolescence: analysis of three longitudinal studies in England. *Prevention Science*. 2014;15(4):600-10.
260. Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. *Clinical Trials*. 2005;2(2):99-107.
261. Gulliford MC, Ukoumunne OC, Chinn S. Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: data from the Health Survey for England 1994. *American Journal of Epidemiology*. 1999;149(9):876-83.
262. Murray DM, & Blitstein, J. L. . Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review*. 2003;27:79-103.

263. UNESCO. International Standard Classification of Education: ISCED 2011. Montreal: UIS; 2012. p. 85.
264. The EndNote Team. EndNote. EndNote 20 version ed. Philadelphia, PA: Clarivate; 2022.
265. Goldstein H, Browne W, Rasbash J. Multilevel modelling of medical data. *Statistics in Medicine*. 2002;21(21):3291-315.
266. Britannica E. Maps of the World: Encyclopædia Britannica, Inc.; 2023 [Available from: <https://www.britannica.com/topic/Maps-of-the-World-1788586>].
267. Lopata C, Thomeer ML, Rodgers JD, Donnelly JP, McDonald CA, Volker MA, et al. Cluster Randomized Trial of a School Intervention for Children with Autism Spectrum Disorder. *Journal of Clinical Child and Adolescent Psychology*. 2019;48(6):922-33.
268. Nykänen M, Sund R, Vuori J. Enhancing safety competencies of young adults: A randomized field trial (RCT). *Journal of Safety Research*. 2018;67:45-56.
269. Stebbins S, Cummings DA, Stark JH, Vukotich C, Mitruka K, Thompson W, et al. Reduction in the incidence of influenza A but not influenza B associated with use of hand sanitizer and cough hygiene in schools: a randomized controlled trial. *The Pediatric Infectious Disease Journal*. 2011;30(11):921-6.
270. Ochoa-Avilés A, Verstraeten R, Huybregts L, Andrade S, Van Camp J, Donoso S, et al. A school-based intervention improved dietary intake outcomes and reduced waist circumference in adolescents: a cluster randomized controlled trial. *Nutrition Journal*. 2017;16(1):79.
271. Stephenson J, Strange V, Allen E, Copas A, Johnson A, Bonell C, et al. The long-term effects of a peer-led sex education programme (RIPPLE): a cluster randomised trial in schools in England. *PLoS Med*. 2008;5(11):e224; discussion e.
272. Watson-Jones D, Baisley K, Ponsiano R, Lemme F, Remes P, Ross D, et al. Human papillomavirus vaccination in Tanzanian schoolgirls: cluster-randomized trial comparing 2 vaccine-delivery strategies. *The Journal of Infectious Diseases*. 2012;206(5):678-86.
273. Stallard P, Sayal K, Phillips R, Taylor JA, Spears M, Anderson R, et al. Classroom based cognitive behavioural therapy in reducing symptoms of depression in high risk adolescents: pragmatic cluster randomised controlled trial. *British Medical Journal*. 2012;345:e6058.
274. Emery CA, Rose MS, McAllister JR, Meeuwisse WH. A prevention strategy to reduce the incidence of injury in high school basketball: a cluster randomized controlled trial. *Clinical Journal of Sport Medicine*. 2007;17(1):17-24.
275. Feng Y, Yin W, Hu D, Zhang YP, Ellwood RP, Pretty IA. Assessment of autofluorescence to detect the remineralization capabilities of sodium fluoride, monofluorophosphate and non-fluoride dentifrices. A single-blind cluster randomized trial. *Caries Research*. 2007;41(5):358-64.
276. Brinker TJ, Faria BL, de Faria OM, Klode J, Schadendorf D, Utikal JS, et al. Effect of a Face-Aging Mobile App–Based Intervention on Skin Cancer Protection Behavior in Secondary Schools in Brazil: A Cluster-Randomized Clinical Trial. *JAMA Dermatology*. 2020;156(7):737-45.
277. Cooper PJ, Chico ME, Vaca MG, Moncayo AL, Bland JM, Mafla E, et al. Effect of albendazole treatments on the prevalence of atopy in children living in communities endemic for geohelminth parasites: a cluster-randomised trial. *Lancet*. 2006;367(9522):1598-603.
278. Shah S, Peat JK, Mazurski EJ, Wang H, Sindhusake D, Bruce C, et al. Effect of peer led programme for asthma education in adolescents: cluster randomised controlled trial. *BMJ*. 2001;322(7286):583-5.
279. Kirk HE, Spencer-Smith M, Wiley JF, Cornish KM. Gamified Attention Training in the Primary School Classroom: A Cluster-Randomized Controlled Trial. *Journal of Attention Disorders*. 2021;25(8):1146-59.
280. Teesson M, Newton NC, Slade T, Carragher N, Barrett EL, Champion KE, et al. Combined universal and selective prevention for adolescent alcohol use: a cluster randomized controlled trial. *Psychological Medicine*. 2017;47(10):1761-70.

281. Sánchez-Jiménez V, Muñoz-Fernández N, Ortega-Rivera J. Efficacy evaluation of "Dat-e Adolescence": A dating violence prevention program in Spain. *PLoS One*. 2018;13(10):e0205802.
282. Newton NC, Conrod PJ, Slade T, Carragher N, Champion KE, Barrett EL, et al. The long-term effectiveness of a selective, personality-targeted prevention program in reducing alcohol use and related harms: a cluster randomized controlled trial. *Journal of Child Psychology and Psychiatry*. 2016;57(9):1056-65.
283. Wen X, Chen W, Gans KM, Colby SM, Lu C, Liang C, et al. Two-year effects of a school-based prevention programme on adolescent cigarette smoking in Guangzhou, China: a cluster randomized trial. *International Journal of Epidemiology*. 2010;39(3):860-76.
284. Cunha SS, Alexander N, Barreto ML, Pereira ES, Dourado I, de Fátima Maroja M, et al. BCG Revaccination Does Not Protect Against Leprosy in the Brazilian Amazon: A Cluster Randomised Trial. *PLOS Neglected Tropical Diseases*. 2008;2(2):e167.
285. Halliday KE, Okello G, Turner EL, Njagi K, McHaro C, Kengo J, et al. Impact of Intermittent Screening and Treatment for Malaria among School Children in Kenya: A Cluster Randomised Trial. *PLOS Medicine*. 2014;11(1):e1001594.
286. Korevaar E, Kasza J, Taljaard M, Hemming K, Haines T, Turner EL, et al. Intra-cluster correlations from the CLustered OUtcome Dataset bank to inform the design of longitudinal cluster trials. *Clinical Trials*. 2021;18(5):529-40.
287. Turner RM, Omar RZ, Thompson SG. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Statistics in Medicine*. 2001;20(3):453-72.
288. Jones BG, Streeter AJ, Baker A, Moyeed R, Creanor S. Bayesian statistics in the design and analysis of cluster randomised controlled trials and their reporting quality: a methodological systematic review. *Systematic Reviews*. 2021;10(1):91.
289. Li F, Hughes JP, Hemming K, Taljaard M, Melnick ER, Heagerty PJ. Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: An overview. *Statistical Methods in Medical Research*. 2021;30(2):612-39.
290. Department of Education. Promoting and supporting mental health and wellbeing in schools and colleges 2021 [Available from: <https://www.gov.uk/guidance/mental-health-and-wellbeing-support-in-schools-and-colleges>].
291. Reinke WM, Stormont M, Herman KC, Puri R, Goel N. Supporting children's mental health in schools: Teacher perceptions of needs, roles, and barriers. *School Psychology Quarterly*. 2011;26:1-13.
292. Banerjee R, McLaughlin C, Cotney J, Roberts L, Peereboom C. Promoting emotional health, well-being and resilience in primary schools. *Public Policy Institute of Wales*; 2016.
293. Kessler RC, Amminger GP, Aguilar-Gaxiola S, Alonso J, Lee S, Ustün TB. Age of onset of mental disorders: a review of recent literature. *Current Opinion in Psychiatry*. 2007;20(4):359-64.
294. Kidger J, Turner N, Hollingworth W, Evans R, Bell S, Brockman R, et al. An intervention to improve teacher well-being support and training to support students in UK high schools (the WISE study): A cluster randomised controlled trial. *PLOS Medicine*. 2021;18(11):e1003847.
295. Jessiman P, Kidger J, Spencer L, Geijer-Simpson E, Kaluzeviciute G, Burn AM, et al. School culture and student mental health: a qualitative study in UK secondary schools. *BMC Public Health*. 2022;22(1):619.
296. Kidger J, Araya R, Donovan J, Gunnell D. The effect of the school environment on the emotional health of adolescents: a systematic review. *Pediatrics*. 2012;129(5):925-49.
297. Hemming K, Eldridge S, Forbes G, Weijer C, Taljaard M. How to design efficient cluster randomised trials. *BMJ*. 2017;358:j3064.
298. Kuyken W, Ball S, Crane C, Ganguli P, Jones B, Montero-Marin J, et al. Effectiveness and cost-effectiveness of universal school-based mindfulness training compared with normal school provision in reducing risk of mental health problems and promoting well-being in adolescence: the MYRIAD cluster randomised controlled trial. *BMJ Mental Health*. 2022;25(3):99-109.

299. Webster-Stratton C, Reid MJ. The Incredible Years parents, teachers, and children training series: A multifaceted treatment approach for young children with conduct problems. Evidence-based psychotherapies for children and adolescents, 3rd ed. New York, NY, US: The Guilford Press; 2018. p. 122-41.
300. Goodman R. Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry.* 2001;40(11):1337-45.
301. Allwood M, Allen K, Price A, Hayes R, Edwards V, Ball S, et al. The reliability and validity of the pupil behaviour questionnaire: a child classroom behaviour assessment tool. *Emotional and Behavioural Difficulties.* 2018;23(4):361-71.
302. Allen K, Marlow R, Edwards V, Parker C, Rodgers L, Ukoumunne OC, et al. 'How I Feel About My School': The construction and validation of a measure of wellbeing at school for primary school children. *Clinical Child Psychology and Psychiatry.* 2018;23(1):25-41.
303. Salmivalli C, Kärnä A, Poskiparta E. Counteracting bullying in Finland: The KiVa program and its effects on different forms of being bullied. *International Journal of Behavioral Development.* 2011;35(5):405-11.
304. Olweus D. The Revised Olweus Bully / Victim Questionnaire. Bergen: Research Center for Health Promotion (HEMIL Center); University of Bergen; 1996.
305. Kärnä A, Voeten M, Little TD, Poskiparta E, Alanen E, Salmivalli C. Going to scale: a nonrandomized nationwide trial of the KiVa antibullying program for grades 1-9. *Journal of Consulting and Clinical Psychology.* 2011;79(6):796-805.
306. Barrett P. Friends for Life - Group leaders' manual for children. Bowen Hills: Australian Academic Press; 2004.
307. Sandín B, Chorot P, Valiente RM, Chorpita BF. Development of a 30-item version of the Revised Child Anxiety and Depression Scale. 2010.
308. Chorpita BF, Tracey SA, Brown TA, Collica TJ, Barlow DH. Assessment of worry in children and adolescents: an adaptation of the Penn State Worry Questionnaire. *Behaviour Research and Therapy.* 1997;35(6):569-81.
309. Rosenberg M. Society and the adolescent self-image: Princeton university press; 2015.
310. Furber G, Segal L. The validity of the Child Health Utility instrument (CHU9D) as a routine outcome measure for use in child and adolescent mental health services. *Health and Quality of Life Outcomes.* 2015;13(1):22.
311. Universal School-based Approaches to Preventing Adolescent Depression: Past Findings and Future Directions of the Resourceful Adolescent Program [press release]. United Kingdom: Clifford Beers Foundation 2004.
312. Angold A, Costello, E. J., Messer, S. C., Pickles, A., Winder, F., & Silver, D. The development of a short questionnaire for use in epidemiological studies of depression in children and adolescents. *International Journal of Methods in Psychiatric Research.* 1995;5:237 – 49.
313. Schniering CA, Rapee RM. Development and validation of a measure of children's automatic thoughts: the children's automatic thoughts scale. *Behaviour Research and Therapy.* 2002;40(9):1091-109.
314. Goodenow C. The psychological sense of school membership among adolescents: Scale development and educational correlates. *Psychology in the Schools.* 1993;30(1):79-90.
315. Radloff LS. The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement.* 1977;1(3):385-401.
316. Tennant R, Hiller L, Fishwick R, Platt S, Joseph S, Weich S, et al. The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. *Health and Quality of Life Outcomes.* 2007;5(1):63.
317. Gioia GA, Isquith PK, Guy SC, Kenworthy L. BRIEF-2: Behavior rating inventory of executive function: Psychological Assessment Resources Lutz, FL; 2015.
318. Spier E. Alaska school climate and connectedness survey: 2016 statewide report. 2016.

319. Greco LA, Baer RA, Smith GT. Assessing mindfulness in children and adolescents: development and validation of the Child and Adolescent Mindfulness Measure (CAMM). *Psychol Assess.* 2011;23(3):606-14.
320. Stone LL, Otten R, Engels RC, Vermulst AA, Janssens JM. Psychometric properties of the parent and teacher versions of the strengths and difficulties questionnaire for 4- to 12-year-olds: a review. *Clinical Child and Family Psychology Review.* 2010;13(3):254-74.
321. Collishaw S, Goodman R, Ford T, Rabe-Hesketh S, Pickles A. How far are associations between child, family and community factors and child psychopathology informant-specific and informant-general? *Journal of Child Psychology and Psychiatry.* 2009;50(5):571-80.
322. van den Heuvel M, Jansen D, Stewart RE, Smits-Engelsman BCM, Reijneveld SA, Flapper BCT. How reliable and valid is the teacher version of the Strengths and Difficulties Questionnaire in primary school children? *PLoS One.* 2017;12(4):e0176605.
323. Department of Education. Behaviour in schools - Advice for headteachers and school staff 2022 [Available from: <https://www.gov.uk/government/publications/behaviour-in-schools--2>].
324. Bonell C, Parry W, Wells H, Jamal F, Fletcher A, Harden A, et al. The effects of the school environment on student health: a systematic review of multi-level studies. *Health Place.* 2013;21:180-91.
325. Bradshaw CP, Waasdorp TE, Debnam KJ, Johnson SL. Measuring school climate in high schools: a focus on safety, engagement, and the environment. *Journal of School Health.* 2014;84(9):593-604.
326. Smart D, Sanson A. Social Competence in Young Adulthood, Its Nature and Antecedents. *Family Matters.* 2003(64):4-9.
327. Department of Education. Academic year 2021/22 - School capacity 2022 [Available from: <https://explore-education-statistics.service.gov.uk/find-statistics/school-capacity>].
328. Liu J. Childhood externalizing behavior: theory and implications. *J Child Adolesc Psychiatr Nurs.* 2004;17(3):93-103.
329. National Institute for Health and Care Research. PROSPERO: International prospective register of systematic reviews: University of York: Centre for Reviews and Dissemination 2023 [Available from: <https://www.crd.york.ac.uk/PROSPERO/>].
330. Kuyken W, Weare K, Ukoumunne OC, Vicary R, Motton N, Burnett R, et al. Effectiveness of the Mindfulness in Schools Programme: non-randomised controlled feasibility study. *British Journal of Psychiatry.* 2013;203(2):126-31.
331. Reardon T, Ball S, Breen M, Brown P, Day E, Ford T, et al. Identifying Child Anxiety Through Schools-identification to intervention (iCATS-i2i): protocol for single-arm feasibility trial. *Pilot and Feasibility Studies.* 2022;8(1):176.
332. Evans-Lacko S, Rose D, Little K, Flach C, Rhydderch D, Henderson C, et al. Development and psychometric properties of the reported and intended behaviour scale (RIBS): a stigma-related behaviour measure. *Epidemiol Psychiatr Sci.* 2011;20(3):263-71.
333. Lewis J, Julious SA. Sample sizes for cluster-randomised trials with continuous outcomes: Accounting for uncertainty in a single intra-cluster correlation estimate. *Statistical Methods in Medical Research.* 2021;30(11):2459-70.
334. Cunha DB, de Souza Bda S, Pereira RA, Sichieri R. Effectiveness of a randomized school-based intervention involving families and teachers to prevent excessive weight gain among adolescents in Brazil. *PLoS ONE.* 2013;8(2):e57498.
335. Leme AC, Lubans DR, Guerra PH, Dewar D, Toassa EC, Philippi ST. Preventing obesity among Brazilian adolescent girls: Six-month outcomes of the Healthy Habits, Healthy Girls-Brazil school-based randomized controlled trial. *Prev Med.* 2016;86:77-83.
336. Liu Z, Li Q, Maddison R, Ni Mhurchu C, Jiang Y, Wei DM, et al. A School-Based Comprehensive Intervention for Childhood Obesity in China: A Cluster Randomized Controlled Trial. *Childhood Obesity.* 2019;15(2):105-15.

337. Grydeland M, Bjelland M, Anderssen SA, Klepp KI, Bergh IH, Andersen LF, et al. Effects of a 20-month cluster randomised controlled school-based intervention trial on BMI of school-aged boys and girls: the HEIA study. *British Journal of Sports Medicine*. 2014;48(9):768-73.
338. Fitzgibbon ML, Stolley MR, Schiffer L, Van Horn L, KauferChristoffel K, Dyer A. Hip-Hop to Health Jr. for Latino preschool children. *Obesity*. 2006;14(9):1616-25.
339. Gray HL, Burgermaster M, Tipton E, Contento IR, Koch PA, Di Noia J. Intraclass Correlation Coefficients for Obesity Indicators and Energy Balance-Related Behaviors Among New York City Public Elementary Schools. *Health Education & Behavior*. 2016;43(2):172-81.
340. Sichiari R, Paula Trotte A, de Souza RA, Veiga GV. School randomised trial on prevention of excessive weight gain by discouraging students from drinking sodas. *Public Health Nutrition*. 2009;12(2):197-202.
341. Waters E, Gibbs L, Tadic M, Ukoumunne OC, Magarey A, Okely AD, et al. Cluster randomised trial of a school-community child health promotion and obesity prevention intervention: findings from the evaluation of fun 'n healthy in Moreland! *BioMed Central Public Health*. 2018;18(1):92.
342. Pena S, Carranza M, Cuadrado C, Parra DC, Villalobos Dintrans P, Castillo C, et al. Effectiveness of a Gamification Strategy to Prevent Childhood Obesity in Schools: A Cluster Controlled Trial. *Obesity*. 2021;17:17.
343. Li B, Pallan M, Liu WJ, Hemming K, Frew E, Lin R, et al. The CHIRPY DRAGON intervention in preventing obesity in Chinese primary-school-aged children: A cluster-randomised controlled trial. *PLoS Medicine / Public Library of Science*. 2019;16(11):e1002971.
344. Viggiano A, Viggiano E, Di Costanzo A, Viggiano A, Andreozi E, Romano V, et al. Kaledo, a board game for nutrition education of children and adolescents at school: cluster randomized controlled trial of healthy lifestyle promotion. *European Journal of Pediatrics*. 2015;174(2):217-28.
345. Robbins LB, Ling J, Wen F. Attending After-School Physical Activity Club 2 Days a Week Attenuated an Increase in Percentage Body Fat and a Decrease in Fitness Among Adolescent Girls at Risk for Obesity. *American Journal of Health Promotion*. 2020;34(5):500-4.
346. Lubans DR, Morgan PJ, Dewar D, Collins CE, Plotnikoff RC, Okely AD, et al. The Nutrition and Enjoyable Activity for Teen Girls (NEAT girls) randomized controlled trial for adolescent girls from disadvantaged secondary schools: rationale, study protocol, and baseline results. *BioMed Central Public Health*. 2010;10:652.
347. Daly RM, Ducher G, Hill B, Telford RM, Eser P, Naughton G, et al. Effects of a Specialist-Led, School Physical Education Program on Bone Mass, Structure, and Strength in Primary School Children: A 4-Year Cluster Randomized Controlled Trial. *Journal of Bone & Mineral Research*. 2016;31(2):289-98.
348. Martinez-Vizcaino V, Pozuelo-Carrascosa DP, Garcia-Prieto JC, Cavero-Redondo I, Solera-Martinez M, Garrido-Miguel M, et al. Effectiveness of a school-based physical activity intervention on adiposity, fitness and blood pressure: MOVI-KIDS study. *British Journal of Sports Medicine*. 2020;54(5):279-85.
349. Ten Hoor GA, Rutten GM, Van Breukelen GJP, Kok G, Ruiters RAC, Meijer K, et al. Strength exercises during physical education classes in secondary schools improve body composition: a cluster randomized controlled trial. *International Journal of Behavioral Nutrition & Physical Activity*. 2018;15(1):92.
350. Bayer O, von Kries R, Strauss A, Mitschek C, Toschke AM, Hose A, et al. Short- and mid-term effects of a setting based prevention program to reduce obesity risk factors in children: a cluster-randomized trial. *Clinical Nutrition*. 2009;28(2):122-8.
351. Muckelbauer R, Libuda L, Clausen K, Toschke AM, Reinehr T, Kersting M. Promotion and provision of drinking water in schools for overweight prevention: randomized, controlled cluster trial. *Pediatrics*. 2009;123(4):e661-7.
352. Kriemler S, Zahner L, Schindler C, Meyer U, Hartmann T, Hebestreit H, et al. Effect of school based physical activity programme (KISS) on fitness and adiposity in primary schoolchildren: cluster randomised controlled trial. *British Medical Journal*. 2010;340:c785.

353. Tarp J, Domazet SL, Froberg K, Hillman CH, Andersen LB, Bugge A. Effectiveness of a School-Based Physical Activity Intervention on Cognitive Performance in Danish Adolescents: LCoMotion-Learning, Cognition and Motion - A Cluster Randomized Controlled Trial. *PLoS ONE*. 2016;11(6):e0158087.
354. Stavnsbo M, Aadland E, Anderssen SA, Chinapaw M, Steene-Johannessen J, Andersen LB, et al. Effects of the Active Smarter Kids (ASK) physical activity intervention on cardiometabolic risk factors in children: A cluster-randomized controlled trial. *Prev Med*. 2020;130:105868.
355. Davis J, Nikah K, Asigbee FM, Landry MJ, Vandyousefi S, Ghaddar R, et al. Design and participant characteristics of TX sprouts: A school-based cluster randomized gardening, nutrition, and cooking intervention. *Contemporary Clinical Trials*. 2019;85:105834.
356. Champion KE, Newton NC, Stapinski L, Slade T, Barrett EL, Teesson M. A cross-validation trial of an Internet-based prevention program for alcohol and cannabis: Preliminary results from a cluster randomised controlled trial. *Australian & New Zealand Journal of Psychiatry*. 2016;50(1):64-73.
357. Tael-Oeren M, Naughton F, Sutton S. A parent-oriented alcohol prevention program "Effekt" had no impact on adolescents' alcohol use: Findings from a cluster-randomized controlled trial in Estonia. *Drug & Alcohol Dependence*. 2019;194:279-87.
358. Martinez-Montilla JM, Mercken L, de Vries H, Candel M, Lima-Rodriguez JS, Lima-Serrano M. A Web-Based, Computer-Tailored Intervention to Reduce Alcohol Consumption and Binge Drinking Among Spanish Adolescents: Cluster Randomized Controlled Trial. *Journal of Medical Internet Research*. 2020;22(1):e15438.
359. Bodin MC, Strandberg AK. The Orebro prevention programme revisited: a cluster-randomized effectiveness trial of programme effects on youth drinking. *Addiction*. 2011;106(12):2134-43.
360. Sumnall H, Agus A, Cole J, Doherty P, Foxcroft D, Harvey S, et al. Steps Towards Alcohol Misuse Prevention Programme (STAMPP): a school- and community- based cluster randomised controlled trial. *NIHR Journals Library Public Health Research*. 2017;04:04.
361. Koning IM, Vollebergh WA, Smit F, Verdurmen JE, Van Den Eijnden RJ, Ter Bogt TF, et al. Preventing heavy alcohol use in adolescents (PAS): cluster randomized trial of a parent and student intervention offered separately and simultaneously. *Addiction*. 2009;104(10):1669-78.
362. D'Amico EJ, Tucker JS, Miles JN, Zhou AJ, Shih RA, Green HD, Jr. Preventing alcohol use with a voluntary after-school program for middle school students: results from a cluster randomized controlled trial of CHOICE. *Prevention Science*. 2012;13(4):415-25.
363. Vallentin-Holbech L, Rasmussen BM, Stock C. Effects of the social norms intervention The GOOD Life on norm perceptions, binge drinking and alcohol-related harms: A cluster-randomised controlled trial. *Preventive Medicine Reports*. 2018;12:304-11.
364. Haug S, Paz Castro R, Kowatsch T, Filler A, Dey M, Schaub MP. Efficacy of a web- and text messaging-based intervention to reduce problem drinking in adolescents: Results of a cluster-randomized controlled trial. *Journal of Consulting & Clinical Psychology*. 2017;85(2):147-59.
365. Palacios AM, Freeland-Graves JH, Dulience SJ, Delnatus JR, Iannotti LL. Differences in factors associated with anemia in Haitian children from urban and rural areas. *PLoS ONE*. 2021;16(4):e0247975.
366. Miller G, Luo R, Zhang L, Sylvia S, Shi Y, Foo P, et al. Effectiveness of provider incentives for anaemia reduction in rural China: a cluster randomised trial. *British Medical Journal*. 2012;345:e4809.
367. Makris KC, Konstantinou C, Andrianou XD, Charisiadis P, Kyriacou A, Gribble MO, et al. A cluster-randomized crossover trial of organic diet impact on biomarkers of exposure to pesticides and biomarkers of oxidative stress/inflammation in primary school children. *PLoS ONE*. 2019;14(9):e0219420.
368. Azam MT, Bush HM, Coker AL, Westgate PM. Effect sizes and intra-cluster correlation coefficients measured from the Green Dot High School study for guiding sample size

- calculations when designing future violence prevention cluster randomized trials in school settings. *Contemporary Clinical Trials Communications*. 2021;23:100831.
369. Pakpour AH, Gholami M, Gellert P, Yekaninejad MS, Dombrowski SU, Webb TL. The Effects of Two Planning Interventions on the Oral Health Behavior of Iranian Adolescents: A Cluster Randomized Controlled Trial. *Annals of Behavioral Medicine*. 2016;50(3):409-18.
370. Young C, Wong KY, Cheung LK. Effectiveness of educational poster on knowledge of emergency management of dental trauma--part 2: cluster randomised controlled trial for secondary school students. *PLoS ONE*. 2014;9(8):e101972.
371. Rodriguez G, Ruiz B, Faleiros S, Vistoso A, Marro ML, Sanchez J, et al. Probiotic Compared with Standard Milk for High-caries Children: A Cluster Randomized Trial. *Journal of Dental Research*. 2016;95(4):402-7.
372. Haleem A, Siddiqui MI, Khan AA. School-based strategies for oral health education of adolescents--a cluster randomized controlled trial. *BioMed Central Oral Health*. 2012;12:54.
373. Nammontri O, Robinson PG, Baker SR. Enhancing oral health via sense of coherence: a cluster-randomized trial. *Journal of Dental Research*. 2013;92(1):26-31.
374. Pakpour AH, Yekaninejad MS, Sniehotta FF, Updegraff JA, Dombrowski SU. The effectiveness of gain-versus loss-framed health messages in improving oral health in Iranian secondary schools: a cluster-randomized controlled trial. *Annals of Behavioral Medicine*. 2014;47(3):376-87.
375. Martiniuk AL, Speechley KN, Secco M, Campbell MK, Donner A. Evaluation of an epilepsy education program for Grade 5 students: a cluster randomized trial. *Epilepsy & Behavior*. 2007;10(4):604-10.
376. Rossetto A, Morgan AJ, Hart LM, Kelly CM, Jorm AF. Frequency and quality of first aid offered by older adolescents: a cluster randomised crossover trial of school-based first aid courses. *PeerJ*. 2020;8:e9782.
377. Tahlil T, Woodman RJ, Coveney J, Ward PR. Six-months follow-up of a cluster randomized trial of school-based smoking prevention education programs in Aceh, Indonesia. *BioMed Central Public Health*. 2015;15:1088.
378. Lassander M, Hintsanen M, Suominen S, Mullola S, Vahlberg T, Volanen SM. Effects of school-based mindfulness intervention on health-related quality of life: moderating effect of gender, grade, and independent practice in cluster randomized controlled trial. *Quality of Life Research*. 2021;24:24.
379. Denbaek AM, Andersen A, Bonnesen CT, Laursen B, Ersboll AK, Due P, et al. Effect Evaluation of a Randomized Trial to Reduce Infectious Illness and Illness-related Absenteeism Among Schoolchildren: The Hi Five Study. *Pediatric Infectious Disease Journal*. 2018;37(1):16-21.
380. Nsangi A, Semakula D, Oxman AD, Austvoll-Dahlgren A, Oxman M, Rosenbaum S, et al. Effects of the Informed Health Choices primary school intervention on the ability of children in Uganda to assess the reliability of claims about treatment effects: a cluster-randomised controlled trial. *Lancet*. 2017;390(10092):374-88.
381. Priest P, McKenzie JE, Audas R, Poore M, Brunton C, Reeves L. Hand sanitiser provision for reducing illness absences in primary school children: a cluster randomised trial. *PLoS Medicine / Public Library of Science*. 2014;11(8):e1001700.
382. Kesztyus D, Lauer R, Traub M, Kesztyus T, Steinacker JM. Effects of statewide health promotion in primary schools on children's sick days, visits to a physician and parental absence from work: a cluster-randomized trial. *BioMed Central Public Health*. 2016;16(1):1244.
383. Rosen L, Manor O, Engelhard D, Brody D, Rosen B, Peleg H, et al. Can a handwashing intervention make a difference? Results from a randomized controlled trial in Jerusalem preschools. *Prev Med*. 2006;42(1):27-32.
384. Phillips-Howard PA, Nyothach E, Ter Kuile FO, Omoto J, Wang D, Zeh C, et al. Menstrual cups and sanitary pads to reduce school attrition, and sexually transmitted and reproductive tract infections: a cluster randomised controlled feasibility study in rural Western Kenya. *British Medical Journal Open*. 2016;6(11):e013229.

385. Suss-Havemann C, Kosan J, Seibold T, Dibbern NM, Daubmann A, Kubitz JC, et al. Implementation of Basic Life Support training in schools: a randomised controlled trial evaluating self-regulated learning as alternative training concept. *BioMed Central Public Health*. 2020;20(1):50.
386. Ssewamala FM, Shu-Huah Wang J, Brathwaite R, Sun S, Mayo-Wilson LJ, Neilands TB, et al. Impact of a Family Economic Intervention (Bridges) on Health Functioning of Adolescents Orphaned by HIV/AIDS: A 5-Year (2012-2017) Cluster Randomized Controlled Trial in Uganda. *Am J Public Health*. 2021;111(3):504-13.
387. Woods-Townsend K, Hardy-Johnson P, Bagust L, Barker M, Davey H, Griffiths J, et al. A cluster-randomised controlled trial of the LifeLab education intervention to improve health literacy in adolescents. *PLoS ONE*. 2021;16(5):e0250545.
388. Berg RL, Pickett W, Fitz-Randolph M, Broste SK, Knobloch MJ, Wood DJ, et al. Hearing conservation program for agricultural students: short-term outcomes from a cluster-randomized trial with planned long-term follow-up. *Prev Med*. 2009;49(6):546-52.
389. Marlenga B, Linneman JG, Pickett W, Wood DJ, Kirkhorn SR, Broste SK, et al. Randomized trial of a hearing conservation intervention for rural students: long-term outcomes. *Pediatrics*. 2011;128(5):e1139-46.
390. Karki P, Uranw S, Bastola S, Mahato R, Shrestha NR, Sherpa K, et al. Effectiveness of Systematic Echocardiographic Screening for Rheumatic Heart Disease in Nepalese Schoolchildren: A Cluster Randomized Clinical Trial. *Journal of the American Medical Association Cardiology*. 2021;6(4):420-6.
391. Freeman MC, Clasen T, Brooker SJ, Akoko DO, Rheingans R. The impact of a school-based hygiene, water quality and sanitation intervention on soil-transmitted helminth reinfection: a cluster-randomized trial. *American Journal of Tropical Medicine & Hygiene*. 2013;89(5):875-83.
392. Gyorkos TW, Maheu-Giroux M, Blouin B, Casapia M. Impact of health education on soil-transmitted helminth infections in schoolchildren of the Peruvian Amazon: a cluster-randomized controlled trial. *PLoS Neglected Tropical Diseases*. 2013;7(9):e2397.
393. Whelan J, Marshall H, Sullivan TR. Intracluster correlation coefficients in a large cluster randomized vaccine trial in schools: Transmission and impact of shared characteristics. *PLoS ONE*. 2021;16(10):e0254330.
394. Dreifelbis R, Freeman MC, Greene LE, Saboori S, Rheingans R. The impact of school water, sanitation, and hygiene interventions on the health of younger siblings of pupils: a cluster-randomized trial in Kenya. *Am J Public Health*. 2014;104(1):e91-7.
395. Liu X, Hou W, Zhao Z, Cheng J, van Beeck EF, Peng X, et al. A hand hygiene intervention to decrease hand, foot and mouth disease and absence due to sickness among kindergarteners in China: A cluster-randomized controlled trial. *Journal of Infection*. 2019;78(1):19-26.
396. Joachim A, Dewald F, Suarez I, Zemlin M, Lang I, Stutz R, et al. Pooled RT-qPCR testing for SARS-CoV-2 surveillance in schools - a cluster randomised trial. *EClinicalMedicine*. 2021;39:101082.
397. Karanja DMS, Awino EK, Wiegand RE, Okoth E, Abudho BO, Mwinzi PNM, et al. Cluster randomized trial comparing school-based mass drug administration schedules in areas of western Kenya with moderate initial prevalence of *Schistosoma mansoni* infections. *PLoS Neglected Tropical Diseases*. 2017;11(10):e0006033.
398. Kovacs F, Oliver-Frontera M, Plana MN, Royuela A, Muriel A, Gestoso M, et al. Improving schoolchildren's knowledge of methods for the prevention and management of low back pain: a cluster randomized controlled trial. *Spine*. 2011;36(8):E505-12.
399. Iserbyt P, Theys L, Ward P, Charlier N. The effect of a specialized content knowledge workshop on teaching and learning Basic Life Support in elementary school: A cluster randomized controlled trial. *Resuscitation*. 2017;112:17-21.
400. Glang AE, Koester MC, Chesnutt JC, Gioia GA, McAvoy K, Marshall S, et al. The effectiveness of a web-based resource in improving postconcussion management in high schools. *Journal of Adolescent Health*. 2015;56(1):91-7.

401. Nauta J, Knol DL, Adriaensens L, Klein Wolt K, van Mechelen W, Verhagen EA. Prevention of fall-related injuries in 7-year-old to 12-year-old children: a cluster randomised controlled trial. *British Journal of Sports Medicine*. 2013;47(14):909-13.
402. Slauterbeck JR, Choquette R, Tourville TW, Krug M, Mandelbaum BR, Vacek P, et al. Implementation of the FIFA 11+ Injury Prevention Program by High School Athletic Teams Did Not Reduce Lower Extremity Injuries: A Cluster Randomized Controlled Trial. *American Journal of Sports Medicine*. 2019;47(12):2844-52.
403. Emery CA, Cassidy JD, Klassen TP, Rosychuk RJ, Rowe BH. Effectiveness of a home-based balance-training program in reducing sports-related injuries among healthy adolescents: a cluster randomized controlled trial. *Canadian Medical Association Journal*. 2005;172(6):749-54.
404. De Bock F, Breitenstein L, Fischer JE. Positive impact of a pre-school-based nutritional intervention on children's fruit and vegetable intake: results of a cluster-randomized trial. *Public Health Nutrition*. 2012;15(3):466-75.
405. Wyse R, Delaney T, Stacey F, Zoetemeyer R, Lecathelinais C, Lamont H, et al. Effectiveness of a Multistrategy Behavioral Intervention to Increase the Nutritional Quality of Primary School Students' Web-Based Canteen Lunch Orders (Click & Crunch): Cluster Randomized Controlled Trial. *Journal of Medical Internet Research*. 2021;23(9):e26054.
406. Amaro S, Viggiano A, Di Costanzo A, Madeo I, Viggiano A, Baccari ME, et al. Kaledo, a new educational board-game, gives nutritional rudiments and encourages healthy eating in children: a pilot cluster randomized trial. *European Journal of Pediatrics*. 2006;165(9):630-5.
407. Kaufman-Shriqui V, Fraser D, Friger M, Geva D, Bilenko N, Vardi H, et al. Effect of a School-Based Intervention on Nutritional Knowledge and Habits of Low-Socioeconomic School Children in Israel: A Cluster-Randomized Controlled Trial. *Nutrients*. 2016;8(4):234.
408. Ezendam NP, Brug J, Oenema A. Evaluation of the Web-based computer-tailored FATaintPHAT intervention to promote energy balance among adolescents: results from a school cluster randomized trial. *Archives of Pediatrics & Adolescent Medicine*. 2012;166(3):248-55.
409. He M, Xiang F, Zeng Y, Mai J, Chen Q, Zhang J, et al. Effect of Time Spent Outdoors at School on the Development of Myopia Among Children in China: A Randomized Clinical Trial. *Journal of the American Medical Association*. 2015;314(11):1142-8.
410. Steenaert E, Crutzen R, Candel M, de Vries NK. The effectiveness of an interactive organ donation education intervention for Dutch lower-educated students: a cluster randomized controlled trial. *Trials*. 2019;20(1):643.
411. Hill JJ, Keating JL. Daily exercises and education for preventing low back pain in children: cluster randomized controlled trial. *Physical Therapy*. 2015;95(4):507-16.
412. Shaygan M, Jahandide Z, Zarifsanaiy N. An investigation of the effect of smartphone-based pain management application on pain intensity and the quality-of-life dimensions in adolescents with chronic pain: a cluster randomized parallel-controlled trial. *Quality of Life Research*. 2021;31:31.
413. Rathleff MS, Roos EM, Olesen JL, Rasmussen S. Exercise during school hours when added to patient education improves outcome for 2 years in adolescent patellofemoral pain: a cluster randomised trial. *British Journal of Sports Medicine*. 2015;49(6):406-12.
414. Lubans DR, Smith JJ, Eather N, Leahy AA, Morgan PJ, Lonsdale C, et al. Time-efficient intervention to improve older adolescents' cardiorespiratory fitness: findings from the 'Burn 2 Learn' cluster randomised controlled trial. *British Journal of Sports Medicine*. 2020;21:21.
415. Andrade S, Lachat C, Ochoa-Aviles A, Verstraeten R, Huybregts L, Roberfroid D, et al. A school-based intervention improves physical fitness in Ecuadorian adolescents: a cluster-randomized controlled trial. *International Journal of Behavioral Nutrition & Physical Activity*. 2014;11:153.
416. Harris N, Warbrick I, Atkins D, Vandal A, Plank L, Lubans DR. Feasibility and Provisional Efficacy of Embedding High-Intensity Interval Training Into Physical Education Lessons: A Pilot Cluster-Randomized Controlled Trial. *Pediatric Exercise Science*. 2021:1-10.

417. Muller I, Schindler C, Adams L, Endes K, Gall S, Gerber M, et al. Effect of a Multidimensional Physical Activity Intervention on Body Mass Index, Skinfolds and Fitness in South African Children: Results from a Cluster-Randomised Controlled Trial. *International Journal of Environmental Research & Public Health*. 2019;16(2):15.
418. Puder JJ, Marques-Vidal P, Schindler C, Zahner L, Niederer I, Burgi F, et al. Effect of multidimensional lifestyle intervention on fitness and adiposity in predominantly migrant preschool children (Ballabeina): cluster randomised controlled trial. *British Medical Journal*. 2011;343:d6195.
419. Cardon G, Labarque V, Smits D, De Bourdeaudhuij I. Promoting physical activity at the pre-school playground: the effects of providing markings and play equipment. *Prev Med*. 2009;48(4):335-40.
420. Kolle E, Solberg RB, Safvenbom R, Dyrstad SM, Berntsen S, Resaland GK, et al. The effect of a school-based intervention on physical activity, cardiorespiratory fitness and muscle strength: the School in Motion cluster randomized trial. *International Journal of Behavioral Nutrition & Physical Activity*. 2020;17(1):154.
421. McNeil DA, Wilson BN, Siever JE, Ronca M, Mah JK. Connecting children to recreational activities: results of a cluster randomized trial. *American Journal of Health Promotion*. 2009;23(6):376-87.
422. Schneider J, Polet J, Hassandra M, Lintunen T, Laukkanen A, Hankonen N, et al. Testing a physical education-delivered autonomy supportive intervention to promote leisure-time physical activity in lower secondary school students: the PETALS trial. *BioMed Central Public Health*. 2020;20(1):1438.
423. De Bock F, Genser B, Raat H, Fischer JE, Renz-Polster H. A participatory physical activity intervention in preschools: a cluster randomized controlled trial. *American Journal of Preventive Medicine*. 2013;45(1):64-74.
424. Suchert V, Isensee B, Sargent J, Weisser B, Hanewinkel R, lauft. Study G. Prospective effects of pedometer use and class competitions on physical activity in youth: A cluster-randomized controlled trial. *Prev Med*. 2015;81:399-404.
425. Sutherland RL, Campbell EM, Lubans DR, Morgan PJ, Nathan NK, Wolfenden L, et al. The Physical Activity 4 Everyone Cluster Randomized Trial: 2-Year Outcomes of a School Physical Activity Intervention Among Adolescents. *American Journal of Preventive Medicine*. 2016;51(2):195-205.
426. Jago R, Tibbitts B, Porter A, Sanderson E, Bird E, Powell JE, et al. A revised teaching assistant-led extracurricular physical activity programme for 8- to 10-year-olds: the Action 3:30R feasibility cluster RCT. . *NIHR Journals Library Public Health Research*. 2019;12:12.
427. Robbins LB, Ling J, Sharma DB, Dalimonte-Merckling DM, Voskuil VR, Resnicow K, et al. Intervention Effects of "Girls on the Move" on Increasing Physical Activity: A Group Randomized Trial. *Annals of Behavioral Medicine*. 2019;53(5):493-500.
428. Toftager M, Christiansen LB, Ersboll AK, Kristensen PL, Due P, Troelsen J. Intervention effects on adolescent physical activity in the multicomponent SPACE study: a cluster randomized controlled trial. *PLoS ONE*. 2014;9(6):e99369.
429. Lonsdale C, Rosenkranz RR, Sanders T, Peralta LR, Bennie A, Jackson B, et al. A cluster randomized controlled trial of strategies to increase adolescents' physical activity and motivation in physical education: results of the Motivating Active Learning in Physical Education (MALP) trial. *Prev Med*. 2013;57(5):696-702.
430. Mendoza JA, Watson K, Baranowski T, Nicklas TA, Uscanga DK, Hanfling MJ. The walking school bus and children's physical activity: a pilot cluster randomized controlled trial. *Pediatrics*. 2011;128(3):e537-44.
431. Mendoza JA, Haaland W, Jacobs M, Abbey-Lambertz M, Miller J, Salls D, et al. Bicycle Trains, Cycling, and Physical Activity: A Pilot Cluster RCT. *American Journal of Preventive Medicine*. 2017;53(4):481-9.
432. Lonsdale C, Lester A, Owen KB, White RL, Peralta L, Kirwan M, et al. An internet-supported school physical activity intervention in low socioeconomic status communities:

- results from the Activity and Motivation in Physical Education (AMPED) cluster randomised controlled trial. *British Journal of Sports Medicine*. 2019;53(6):341-7.
433. Cramer N, Haviland MJ, Zhou C, Mendoza JA. Impact of Walking School Bus Programs on Self-Efficacy and Outcome Expectations. *Journal of Physical Activity & Health*. 2021;18(7):858-62.
434. Nettlefold L, Naylor PJ, Macdonald HM, McKay HA. Scaling up Action Schools! BC: How Does Voltage Drop at Scale Affect Student Level Outcomes? A Cluster Randomized Controlled Trial. *International Journal of Environmental Research & Public Health*. 2021;18(10):13.
435. Naylor PJ, Macdonald HM, Warburton DE, Reed KE, McKay HA. An active school model to promote physical activity in elementary schools: action schools! BC. *British Journal of Sports Medicine*. 2008;42(5):338-43.
436. Mendoza JA, Baranowski T, Jaramillo S, Fesinmeyer MD, Haaland W, Thompson D, et al. Fit 5 Kids TV Reduction Program for Latino Preschoolers: A Cluster Randomized Controlled Trial. *American Journal of Preventive Medicine*. 2016;50(5):584-92.
437. Bjelland M, Soenens B, Bere E, Kovacs E, Lien N, Maes L, et al. Associations between parental rules, style of communication and children's screen time. *BioMed Central Public Health*. 2015;15:1002.
438. Cohen KE, Morgan PJ, Plotnikoff RC, Callister R, Lubans DR. Physical activity and skills intervention: SCORES cluster randomized controlled trial. *Medicine & Science in Sports & Exercise*. 2015;47(4):765-74.
439. Whittemore R, Jeon S, Grey M. An internet obesity prevention program for adolescents. *Journal of Adolescent Health*. 2013;52(4):439-47.
440. Dzielska A, Mazur J, Nalecz H, Oblacinska A, Fijalkowska A. Importance of Self-Efficacy in Eating Behavior and Physical Activity Change of Overweight and Non-Overweight Adolescent Girls Participating in Healthy Me: A Lifestyle Intervention with Mobile Technology. *Nutrients*. 2020;12(7):17.
441. Bavarian N, Lewis KM, Acock A, DuBois DL, Yan Z, Vuchinich S, et al. Effects of a School-Based Social-Emotional and Character Development Program on Health Behaviors: A Matched-Pair, Cluster-Randomized Controlled Trial. *Journal of Primary Prevention*. 2016;37(1):87-105.
442. Piotrowski ZH, Hedeker D. Evaluation of the Be the Exception Sixth-Grade Program in Rural Communities to Delay the Onset of Sexual Behavior. *Am J Public Health*. 2016;106(S1):S132-S9.
443. Potter SC, Coyle KK, Glassman JR, Kershner S, Prince MS. It's Your Game...Keep It Real in South Carolina: A Group Randomized Trial Evaluating the Replication of an Evidence-Based Adolescent Pregnancy and Sexually Transmitted Infection Prevention Program. *Am J Public Health*. 2016;106(S1):S60-S9.
444. Constantine NA, Jerman P, Berglas NF, Angulo-Olaiz F, Chou CP, Rohrbach LA. Short-term effects of a rights-based sexuality education curriculum for high-school students: a cluster-randomized trial. *BioMed Central Public Health*. 2015;15:293.
445. Rohrbach LA, Berglas NF, Jerman P, Angulo-Olaiz F, Chou CP, Constantine NA. A Rights-Based Sexuality Education Curriculum for Adolescents: 1-Year Outcomes From a Cluster-Randomized Trial. *Journal of Adolescent Health*. 2015;57(4):399-406.
446. Mathews C, Eggers SM, Townsend L, Aaro LE, de Vries PJ, Mason-Jones AJ, et al. Effects of PREPARE, a Multi-component, School-Based HIV and Intimate Partner Violence (IPV) Prevention Programme on Adolescent Sexual Risk Behaviour and IPV: Cluster Randomised Controlled Trial. *AIDS & Behavior*. 2016;20(9):1821-40.
447. Martiniuk AL, O'Connor KS, King WD. A cluster randomized trial of a sex education programme in Belize, Central America. *International Journal of Epidemiology*. 2003;32(1):131-6.
448. Jemmott JB, 3rd, Jemmott LS, O'Leary A, Ngwane Z, Icard LD, Bellamy SL, et al. School-based randomized controlled trial of an HIV/STD risk-reduction intervention for South African adolescents. *Archives of Pediatrics & Adolescent Medicine*. 2010;164(10):923-9.

449. Aarestrup C, Bonnesen CT, Thygesen LC, Krarup AF, Waagstein AB, Jensen PD, et al. The effect of a school-based intervention on sunbed use in Danish pupils at continuation schools: a cluster-randomized controlled trial. *Journal of Adolescent Health*. 2014;54(2):214-20.
450. Hunter S, Love-Jackson K, Abdulla R, Zhu W, Lee JH, Wells KJ, et al. Sun protection at elementary schools: a cluster randomized trial. *Journal of the National Cancer Institute*. 2010;102(7):484-92.
451. Roetzheim RG, Love-Jackson KM, Hunter SG, Lee JH, Chen R, Abdulla R, et al. A cluster randomized trial of sun protection at elementary schools. Results from year 2. *American Journal of Preventive Medicine*. 2011;41(6):615-8.
452. Buller DB, Reynolds KD, Yaroch A, Cutter GR, Hines JM, Geno CR, et al. Effects of the Sunny Days, Healthy Ways curriculum on students in grades 6 to 8. *American Journal of Preventive Medicine*. 2006;30(1):13-22.
453. Onrust SA, van der Heijden A, Zschamisch AL, Speetjens PAM. Effectiveness of Fresh Start: A Randomized Study of a School-Based Program to Retain a Negative Attitude Toward Substance Use in Secondary School Freshmen. *Substance Use & Misuse*. 2018;53(6):921-30.
454. Andersen A, Krolner R, Bast LS, Thygesen LC, Due P. Effects of the X:IT smoking intervention: a school-based cluster randomized trial. *International Journal of Epidemiology*. 2015;44(6):1900-8.
455. Kiewik M, VanDerNagel JE, Kemna LE, Engels RC, DeJong CA. Substance use prevention program for adolescents with intellectual disabilities on special education schools: a cluster randomised control trial. *Journal of Intellectual Disability Research*. 2016;60(3):191-200.
456. Caria MP, Faggiano F, Bellocco R, Galanti MR, Group EU-DS. Effects of a school-based prevention program on European adolescents' patterns of alcohol use. *Journal of Adolescent Health*. 2011;48(2):182-8.
457. Hansen J, Hanewinkel R, Maruska K, Isensee B. The 'Eigenständig werden' prevention trial: a cluster randomised controlled study on a school-based life skills programme to prevent substance use onset. *British Medical Journal Open*. 2011;1(2):e000352.
458. Isensee B, Morgenstern M, Stoolmiller M, Maruska K, Sargent JD, Hanewinkel R. Effects of Smokefree Class Competition 1 year after the end of intervention: a cluster randomised controlled trial. *Journal of Epidemiology & Community Health*. 2012;66(4):334-41.
459. Ringwalt CL, Clark HK, Hanley S, Shamblen SR, Flewelling RL. Project ALERT: a cluster randomized trial. *Archives of Pediatrics & Adolescent Medicine*. 2009;163(7):625-32.
460. Krist L, Lotz F, Burger C, Strobele-Benschop N, Roll S, Rieckmann N, et al. Long-term effectiveness of a combined student-parent and a student-only smoking prevention intervention among 7th grade school children in Berlin, Germany. *Addiction*. 2016;111(12):2219-29.
461. Siddiqi K, Huque R, Kanaan M, Ahmed F, Ferdous T, Shah S, et al. Children Learning About Secondhand Smoke (CLASS II): A Pilot Cluster Randomized Controlled Trial. *Nicotine & Tobacco Research*. 2019;21(5):670-7.
462. Huque R, Dogar O, Cameron I, Thomson H, Amos A, Siddiqi K. Children Learning About Second-Hand Smoking: A Feasibility Cluster Randomized Controlled Trial. *Nicotine & Tobacco Research*. 2015;17(12):1465-72.
463. Valdivieso Lopez E, Rey-Renones C, Rodriguez-Blanco T, Ferre Grau C, Arija V, Barrera Uriarte ML, et al. Efficacy of a smoking prevention programme in Catalan secondary schools: a cluster-randomized controlled trial in Spain. *Addiction*. 2015;110(5):852-60.
464. Allara E, Angelini P, Gorini G, Bosi S, Carreras G, Gozzi C, et al. A prevention program for multiple health-compromising behaviors in adolescence: baseline results from a cluster randomized controlled trial. *Prev Med*. 2015;71:20-6.
465. Haug S, Paz Castro R, Kowatsch T, Filler A, Schaub MP. Efficacy of a technology-based, integrated smoking cessation and alcohol intervention for smoking cessation in adolescents: Results of a cluster-randomised controlled trial. *Journal of Substance Abuse Treatment*. 2017;82:55-66.

466. Hiemstra M, Ringlever L, Otten R, van Schayck OC, Jackson C, Engels RC. Long-term effects of a home-based smoking prevention program on smoking initiation: a cluster randomized controlled trial. *Prev Med*. 2014;60:65-70.
467. Sashegyi AI, Brown KS, Farrell PJ. Application of a generalized random effects regression model for cluster-correlated longitudinal data to a school-based smoking prevention trial. *American Journal of Epidemiology*. 2000;152(12):1192-200.
468. Gordon J, Biglan A, Smolkowski K. The impact on tobacco use of branded youth anti-tobacco activities and family communications about tobacco. *Prevention Science*. 2008;9(2):73-87.
469. Hodder RK, Freund M, Bowman J, Wolfenden L, Campbell E, Dray J, et al. Effectiveness of a pragmatic school-based universal resilience intervention in reducing tobacco, alcohol and illicit substance use in a population of adolescents: cluster-randomised controlled trial. *British Medical Journal Open*. 2017;7(8):e016060.
470. Tokolahi E, Vandal AC, Kersten P, Pearson J, Hocking C. Cluster-randomised controlled trial of an occupational therapy intervention for children aged 11-13 years, designed to increase participation to prevent symptoms of mental illness. *Child & Adolescent Mental Health*. 2018;23(4):313-27.
471. Makover H, Adrian M, Wilks C, Read K, Stoep AV, McCauley E. Indicated Prevention for Depression at the Transition to High School: Outcomes for Depression and Anxiety. *Prevention Science*. 2019;20(4):499-509.
472. Guo JL, Lee TC, Liao JY, Huang CM. Prevention of illicit drug use through a school-based program: results of a longitudinal, cluster-randomized controlled trial. *Journal of Adolescent Health*. 2015;56(3):314-22.
473. McCoy DC, Hanno EC, Ponczek V, Pinto C, Fonseca G, Marchi N. Um Compasso Para Aprender: A Randomized Trial of a Social-Emotional Learning Program in Homicide-Affected Communities in Brazil. *Child Development*. 2021;92(5):1951-68.
474. Jenson JM, Dieterich WA. Effects of a skills-based prevention program on bullying and bully victimization among elementary school children. *Prevention Science*. 2007;8(4):285-96.
475. Agle J, Jun M, Eldridge L, Agle DL, Xiao Y, Sussman S, et al. Effects of ACT Out! Social Issue Theater on Social-Emotional Competence and Bullying in Youth and Adolescents: Cluster Randomized Controlled Trial. *Journal of Medical Internet Research Mental Health*. 2021;8(1):e25860.
476. Baker-Henningham H, Scott Y, Francis T, Walker SP. Effects of a Teacher-Training Violence Prevention Program in Jamaican Preschools on Child Behavior, Academic Achievement, and School Attendance in Grade One of Primary School: Follow up of a Cluster Randomized Trial. *Frontiers in Psychology*. 2021;12:652050.
477. Baker-Henningham H, Scott Y, Bowers M, Francis T. Evaluation of a Violence-Prevention Programme with Jamaican Primary School Teachers: A Cluster Randomised Trial. *International Journal of Environmental Research & Public Health*. 2019;16(15):06.
478. Weisleder A, Mazzuchelli DSR, Lopez AS, Neto WD, Cates CB, Goncalves HA, et al. Reading Aloud and Child Development: A Cluster-Randomized Trial in Brazil. *Pediatrics*. 2018;141(1):01.
479. Williford A, Elledge LC, Boulton AJ, DePaolis KJ, Little TD, Salmivalli C. Effects of the KiVa antibullying program on cyberbullying and cybervictimization frequency among Finnish youth. *Journal of Clinical Child & Adolescent Psychology*. 2013;42(6):820-33.
480. Tak YR, Lichtwarck-Aschoff A, Gillham JE, Van Zundert RM, Engels RC. Universal School-Based Depression Prevention 'Op Volle Kracht': a Longitudinal Cluster Randomized Controlled Trial. *Journal of Abnormal Child Psychology*. 2016;44(5):949-61.
481. Perry Y, Werner-Seidler A, Calear A, Mackinnon A, King C, Scott J, et al. Preventing Depression in Final Year Secondary Students: School-Based Randomized Controlled Trial. *Journal of Medical Internet Research*. 2017;19(11):e369.
482. Bradshaw CP, Waasdorp TE, Leaf PJ. Effects of school-wide positive behavioral interventions and supports on child behavior problems. *Pediatrics*. 2012;130(5):e1136-45.

483. Edridge C, Wolpert M, Deighton J, Edbrooke-Childs J. An mHealth Intervention (ReZone) to Help Young People Self-Manage Overwhelming Feelings: Cluster-Randomized Controlled Trial. *Journal of Medical Internet Research*. 2020;22(7):e14223.
484. Willoughby MT, Piper B, King KM, Nduku T, Henny C, Zimmermann S. Testing the Efficacy of the Red-Light Purple-Light Games in Preprimary Classrooms in Kenya. *Frontiers in Psychology*. 2021;12:633049.
485. Lubman DI, Cheetham A, Sandral E, Wolfe R, Martin C, Blee F, et al. Twelve-month outcomes of MAKINGtheLINK: A cluster randomized controlled trial of a school-based program to facilitate help-seeking for substance use and mental health problems. *EClinicalMedicine*. 2020;18:100225.
486. O'Dea B, Subotic-Kerry M, King C, Mackinnon AJ, Achilles MR, Anderson M, et al. A cluster randomised controlled trial of a web-based youth mental health service in Australian schools. *The Lancet Regional Health Western Pacific*. 2021;12:100178.
487. Morgan L, Hooker JL, Sparapani N, Reinhardt VP, Schatschneider C, Wetherby AM. Cluster randomized trial of the classroom SCERTS intervention for elementary students with autism spectrum disorder. *Journal of Consulting & Clinical Psychology*. 2018;86(7):631-44.
488. Link BG, DuPont-Reyes MJ, Barkin K, Villatoro AP, Phelan JC, Painter K. A School-Based Intervention for Mental Illness Stigma: A Cluster Randomized Trial. *Pediatrics*. 2020;145(6):06.
489. Baker-Henningham H, Scott S, Jones K, Walker S. Reducing child conduct problems and promoting social skills in a middle-income country: cluster randomised controlled trial. *British Journal of Psychiatry*. 2012;201:101-8.
490. Boyd BA, Watson LR, Reszka SS, Sideris J, Alessandri M, Baranek GT, et al. Efficacy of the ASAP Intervention for Preschoolers with ASD: A Cluster Randomized Controlled Trial. *Journal of Autism & Developmental Disorders*. 2018;48(9):3144-62.
491. Lewis KM, DuBois DL, Bavarian N, Acock A, Silverthorn N, Day J, et al. Effects of Positive Action on the emotional health of urban youth: a cluster-randomized trial. *Journal of Adolescent Health*. 2013;53(6):706-11.
492. Tol WA, Komproe IH, Jordans MJ, Ndayisaba A, Ntamutumba P, Sipsma H, et al. School-based mental health intervention for children in war-affected Burundi: a cluster randomized trial. *BioMed Central Medicine*. 2014;12:56.
493. Kliewer W, Lepore SJ, Farrell AD, Allison KW, Meyer AL, Sullivan TN, et al. A school-based expressive writing intervention for at-risk urban adolescents' aggressive behavior and emotional lability. *Journal of Clinical Child & Adolescent Psychology*. 2011;40(5):693-705.
494. Ford T, Degli Esposti M, Crane C, Taylor L, Montero-Marin J, Blakemore SJ, et al. The Role of Schools in Early Adolescents' Mental Health: Findings From the MYRIAD Study. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2021;04:04.
495. Hart LM, Morgan AJ, Rossetto A, Kelly CM, Mackinnon A, Jorm AF. Helping adolescents to better support their peers with a mental health problem: A cluster-randomised crossover trial of teen Mental Health First Aid. *Australian & New Zealand Journal of Psychiatry*. 2018;52(7):638-51.
496. Volanen SM, Lassander M, Hankonen N, Santalahti P, Hintsanen M, Simonsen N, et al. Healthy learning mind - Effectiveness of a mindfulness program on mental health compared to a relaxation program and teaching as usual in schools: A cluster-randomised controlled trial. *Journal of Affective Disorders*. 2020;260:660-9.
497. Mazzoli E, Salmon J, Teo WP, Pesce C, He J, Ben-Soussan TD, et al. Breaking up classroom sitting time with cognitively engaging physical activity: Behavioural and brain responses. *PLoS ONE*. 2021;16(7):e0253733.
498. Shinde S, Weiss HA, Varghese B, Khandeparkar P, Pereira B, Sharma A, et al. Promoting school climate and health outcomes with the SEHER multi-component secondary school intervention in Bihar, India: a cluster-randomised controlled trial. *Lancet*. 2018;392(10163):2465-77.

499. Valente JY, Sanchez ZM. Short-Term Secondary Effects of a School-Based Drug Prevention Program: Cluster-Randomized Controlled Trial of the Brazilian Version of DARE's Keepin' it REAL. *Prevention Science*. 2021;05:05.
500. Howard SJ, Vasseleu E, Batterham M, Neilsen-Hewett C. Everyday Practices and Activities to Improve Pre-school Self-Regulation: Cluster RCT Evaluation of the PRSIST Program. *Frontiers in Psychology*. 2020;11:137.
501. Katz J, Knight V, Mercer SH, Skinner SY. Effects of a Universal School-Based Mental Health Program on the Self-concept, Coping Skills, and Perceptions of Social Support of Students with Developmental Disabilities. *Journal of Autism & Developmental Disorders*. 2020;50(11):4069-84.
502. Tirlea L, Truby H, Haines TP. Pragmatic, Randomized Controlled Trials of the Girls on the Go! Program to Improve Self-Esteem in Girls. *American Journal of Health Promotion*. 2016;30(4):231-41.
503. Golan M, Ahmad WA. School-based versus after-school delivery of a universal wellness programme - A randomized controlled multi-arm trial. *Eating Behaviors*. 2018;31:41-7.
504. DiPerna JC, Lei P, Bellinger J, Cheng W. Efficacy of the Social Skills Improvement System Classwide Intervention Program (SSIS-CIP) primary version. *School Psychology Quarterly*. 2015;30(1):123-41.
505. Watanabe J, Watanabe M, Yamaoka K, Adachi M, Nemoto A, Tango T. Effect of School-Based Home-Collaborative Lifestyle Education on Reducing Subjective Psychosomatic Symptoms in Adolescents: A Cluster Randomised Controlled Trial. *PLoS ONE*. 2016;11(10):e0165285.
506. Wasserman D, Hoven CW, Wasserman C, Wall M, Eisenberg R, Hadlaczky G, et al. School-based suicide prevention programmes: the SEYLE cluster-randomised, controlled trial. *Lancet*. 2015;385(9977):1536-44.
507. Bartholomew JB, Golaszewski NM, Jowers E, Korinek E, Roberts G, Fall A, et al. Active learning improves on-task behaviors in 4th grade children. *Prev Med*. 2018;111:49-54.
508. Dray J, Bowman J, Campbell E, Freund M, Hodder R, Wolfenden L, et al. Effectiveness of a pragmatic school-based universal intervention targeting student resilience protective factors in reducing mental health problems in adolescents. *Journal of Adolescence*. 2017;57:74-89.
509. Streimann K, Selart A, Trummal A. Effectiveness of a Universal, Classroom-Based Preventive Intervention (PAX GBG) in Estonia: a Cluster-Randomized Controlled Trial. *Prevention Science*. 2020;21(2):234-44.
510. Calear AL, Christensen H, Mackinnon A, Griffiths KM, O'Kearney R. The YouthMood Project: a cluster randomized controlled trial of an online cognitive behavioral program with adolescents. *Journal of Consulting & Clinical Psychology*. 2009;77(6):1021-32.
511. Newton NC, Andrews G, Champion KE, Teesson M. Universal Internet-based prevention for alcohol and cannabis use reduces truancy, psychological distress and moral disengagement: a cluster randomised controlled trial. *Prev Med*. 2014;65:109-15.
512. Gold C, Saarikallio S, Crooke AHD, McFerran KS. Group Music Therapy as a Preventive Intervention for Young People at Risk: Cluster-Randomized Trial. *Journal of Music Therapy*. 2017;54(2):133-60.
513. Mallick RB, Thabane L, Borhan ASM, Kathard H. A pilot study to determine the feasibility of a cluster randomised controlled trial of an intervention to change peer attitudes towards children who stutter. *South African Journal of Communication Disorders*. 2018;65(1):e1-e8.
514. Champion KE, Newton NC, Stapinski LA, Teesson M. Effectiveness of a universal internet-based prevention program for ecstasy and new psychoactive substances: a cluster randomized controlled trial. *Addiction*. 2016;111(8):1396-405.
515. Miller E, Tancredi DJ, McCauley HL, Decker MR, Virata MC, Anderson HA, et al. "Coaching boys into men": a cluster-randomized controlled trial of a dating violence prevention program. *Journal of Adolescent Health*. 2012;51(5):431-8.

516. Fabbri C, Rodrigues K, Leurent B, Allen E, Qiu M, Zuakulu M, et al. The EmpaTeach intervention for reducing physical violence from teachers to students in Nyarugusu Refugee Camp: A cluster-randomised controlled trial. *PLoS Medicine / Public Library of Science*. 2021;18(10):e1003808.
517. Temple JR, Baumler E, Wood L, Thiel M, Peskin M, Torres E. A Dating Violence Prevention Program for Middle School Youth: A Cluster Randomized Trial. *Pediatrics*. 2021;06:06.
518. Wolfe DA, Crooks C, Jaffe P, Chiodo D, Hughes R, Ellis W, et al. A school-based program to prevent adolescent dating violence: a cluster randomized trial. *Archives of Pediatrics & Adolescent Medicine*. 2009;163(8):692-9.
519. Miller E, Jones KA, Ripper L, Paglisotti T, Mulbah P, Abebe KZ. An Athletic Coach-Delivered Middle School Gender Violence Prevention Program: A Cluster Randomized Clinical Trial. *Journal of the American Medical Association Pediatrics*. 2020;174(3):241-9.
520. Devries KM, Knight L, Child JC, Mirembe A, Nakuti J, Jones R, et al. The Good School Toolkit for reducing physical violence from school staff to primary school students: a cluster-randomised controlled trial in Uganda. *The Lancet Global Health*. 2015;3(7):e378-86.
521. Beets MW, Flay BR, Vuchinich S, Snyder FJ, Acock A, Li KK, et al. Use of a social and character development program to prevent substance use, violent behaviors, and sexual activity among elementary-school students in Hawaii. *Am J Public Health*. 2009;99(8):1438-45.