# Handling **C**omplexity in **E**vidence from systematic reviews and meta-analyses of **P**ublic **H**ealth **I**nterventions (CEPHI project)

Final report to NIHR

Dylan Kneale*, Alison O’Mara-Eves*, Bridget Candy, Katy Sutcliffe, Lizzie Cain, Sandy Oliver, Niccola Hutchinson-Pascal, James Thomas

*Joint leads

## Contents

## Executive Summary

**Background:** The development of evidence-based strategies to tackle complex public health issues has been widely recommended. Nevertheless, methodologically robust sources of evidence may not necessarily be perceived as useful to decision-making in local settings. Despite their high regard, the ability to utilise evidence from meta-analyses and systematic reviews is hampered by the lack of explicit connection between the contexts in which interventions were evaluated and the context in which the evidence is to be applied. In this research we seek to develop approaches for exploring and enhancing the generalisability of meta-analysis through additional synthesis.

**Approach:** Mixed methods design underpinned by co-production. This involved co-producing a systems-based logic model and conducting secondary syntheses of existing systematic review evidence to develop new analysis approaches.

**Methods:** This research focusses on children's health as a case example and the identification of school-based interventions to help improve children's health. The research included three work packages (WPs):

*WP1*: Through a series of workshops with stakeholders, we co-produced a systems-based logic model that helps to identify contextual features of interest and how they interact to influence children's health. This co-production was transformative for the research and helped to reframe children's health away from a stigmatizing and individualistic focus on obesity to a broader focus on children's health. The systems-based logic model is an output in its own right, and was also be used to guide later stages.

*WP2*: This work package developed and refined new approaches in examining the generalisability of evidence, drawing on WP1. First, we assessed how using existing observational data and employing statistical approaches (namely reweighting of effect sizes, binary logistic regressions, and cluster analysis) in novel ways can help to create an overall measure of effect from meta-analysis that is more applicable to a defined population and/or more interpretable for decision-making. Next, we explored the utility of Qualitative Comparative Analysis (QCA) in examining the influence of context. Drawing on set-theory, we attempted to use QCA to examine configurations of different contextual features that align with more successful interventions.

*WP3*: Given that the primary motivation of this project is that end-users are not utilising review evidence because of its disconnect with their particular local circumstances, we explored the utility of the proposed enhancements to address this issue.

**Discussion:** This project has identified (i) a participatory approach for conceptualising health systems; (ii) refined approaches and developed new methods for exploring the influence of context in meta-analytic evidence; and (iii) demonstrated the important of co-production in challenging researchers' assumptions about how systems and factors influence health (in this case children's health). The methodological advancements developed in this research for examining context in meta-analysis provide useful adjunct evidence to decision-makers, alongside existing meta-analytic evidence**.** Here we have focused on four approaches, where two in particular appear to provide a clearer message around the likely impact of contextual and population factors for decision-making. While caveats surround all four approaches, we believe that all four show further potential. However, understanding the potential of these approaches was hindered by an absence of contextual data reported within studies. The results of this research highlight the gulf between the deep and nuanced way in which diverse groups of stakeholders understand the factors that influence children's health, and the sparse treatment of context by researchers within trials and systematic reviews.

**Keywords:** context, generalisability, co-production, evidence, systematic review, secondary data analysis, QCA, meta-analysis, recalibration, logic model.

## Acknowledgements

Thank you to all those who contributed their time, expertise and experience to co-create the logic model with us, without whom this project would look very different.

## Conflict of interest and funding statement

This research is conducted by researchers at the EPPI-Centre, UCL in conjunction with Co-Production Collective. All project members declare no conflicts of interest.

The research is funded by National Institute of Health Research under the Public Health Research stream (PHR Project: NIHR133736 - Handling complexity in evidence from systematic reviews and meta-analyses of public health intervention).

# Background

## Introduction

Achieving population-level change in health outcomes is difficult. Childhood obesity is one such example, where levels of childhood obesity had broadly stabilised pre-pandemic with no apparent declines in prevalence despite substantial investment (van Jaarsveld and Gulliford 2015, NHS Digital 2020). The COVID-19 pandemic has heightened concern around the health of children who have experienced substantial and prolonged disruption to their daily lives.

The development of evidence-based strategies has been widely recommended as a means of improving public health outcomes (Brownson, Fielding et al. 2009). Evidence from systematic reviews and meta-analyses is considered one of the most robust sources for public health decision-making, particularly in assessing the effectiveness of interventions (Berlin and Golub 2014). The opportunity to utilise evidence from meta-analyses and systematic reviews, however, is often hampered by the lack of connection between the contexts in which interventions were evaluated and the context in which the evidence is to be applied. Contextual factors influence the design and delivery of interventions in terms of governance structures, delivery bodies, epidemiological factors, and populations at risk, among others, which can influence the acceptability, reach, and adherence to interventions. Where an intervention works for one population or setting, there is no guarantee that it will work for others, which has implications for rolling out interventions to reduce health inequalities that have not been evaluated with the target population.

Researchers have emphasised the importance of contextual salience of evidence in determining its perceived usefulness among decision-makers (Kneale, Rojas-García et al. 2017, Oliver, Roche et al. 2018, Oliver, Langer et al. 2021). However, systematic reviews often struggle to provide contextually-salient evidence, and consequently local public health decision-makers typically base their decisions on smaller local evaluation studies that are not always methodologically robust, and on anecdotal input (Kneale, Rojas-García et al. 2017). The systematic process of evidence synthesis is thereby undermined by idiosyncratic patterns of evidence use (Ioannidis, Greenland et al. 2014).

This report describes the findings of an NIHR-funded project aiming to enhance the utility of systematic review evidence through developing new approaches for 'Handling **C**omplexity in **E**vidence from systematic reviews and meta-analyses of **P**ublic **H**ealth **I**nterventions (CEPHI project)'.

## How is contextual generalisability currently treated in meta-analysis?

Extrapolating meta-analytic findings to a specific situation is dependent on having a well-defined target (inference) population (Glass 2000, Hedges 2013, O'Muircheartaigh and Hedges 2014). However, systematic reviews assemble evidence from all eligible studies, which differ from a proposed inference population with respect to a range of contextual factors, thus making judgements about applicability less direct. Concern has been raised regarding the under-representation or under-reporting of evidence relating to disadvantaged groups, which undermines the potential of systematic reviews to inform approaches to health inequalities (e.g., incentives for obesity prevention (Paul-Ebhohimhen and Avenell 2008)).

The generalisability of meta-analytic findings is also related to the external validity of the included studies. Most meta-analysts consider the external validity of primary studies individually, using an array of different checklists and frameworks composed of items that may have little methodological justification (Ahmad, Boutron et al. 2010, Burchett, Blanchard et al. 2018). Similarly, there may be a wide gulf between the factors considered important in assessing generalisability by the systematic reviewer, and factors perceived as important among local stakeholders. Furthermore, there is a need to recognise generalisability as a multidimensional construct encompassing both the applicability of evidence (reflecting whether an intervention is feasible) and its transferability (whether the intervention would have the same sort of impact) (Wang, Moss et al. 2006).

Having attempted to assess the external validity of the individual studies, the meta-analyst then needs to consider generalisability at the review level. Many meta-analysts attempt to assess generalisability statistically, by trying to identify and explain any heterogeneity (i.e., variation between studies) through sub-group or regression analyses. Decisions regarding these analyses, however, are just as likely to be data-driven as they are to be driven by concerns about applicability or transferability (Petticrew, Tugwell et al. 2011), thereby increasing the risk that findings are spurious. Additionally, these analyses may not consider multiple factors necessary for the assessment of generalisability and are often conducted without consideration of a specific situation in which the evidence is intended to be applied. The present research set out to explore how study level data describing context, and data describing context for settings where evidence is to be applied, could be synthesized to aid public health decision-making. Systematic reviews are produced often without a clear context of use and particular concern of this research is to examine what can be done to further interpret systematic review evidence where there is a clear context for use.

## Overall approach and research questions

This work was underpinned by the principles of co-production, which is an approach to working together in equal partnership for equal benefit underpinned by the core values of being human, inclusive, transparent and challenging. The knowledge was co-created in a way that was context-specific to the participants, pluralistic (incorporating the perspectives of multiple stakeholders), goal-orientated, participatory, and interactive (see (Norström, Cvitanovic et al. 2020)). School-based interventions related to childhood health was chosen as the topic to develop our approaches.

The overarching aim of the project was to develop methods of exploring and enhancing the generalisability of meta-analysis. The research questions, across three work packages (WP) are:

1. *WP1. Assessing contextual generalisability in a defined setting:* Can local knowledge of contextual features (in terms of people, intervention, usual care conditions, or other

features) be harnessed through co-production of a logic model, and applied in generalisability analyses?
2. *WP2. Adjusting for generalisability and examining the influence of context:* Which statistical approaches (recalibration, binomial logistic regression, enhanced subgroup analysis) can adjust for context to enhance the generalisability of a completed meta-analysis? How might Qualitative Comparative Analysis (QCA) help identify contextual features that trigger more successful interventions, or hinder positive effects? Do the approaches offer a robust way to consider health inequalities in the systematic review?
3. *WP3. Evaluating and disseminating:* How should the approaches be used in review production and decision-making? How are the findings viewed by stakeholders?

### Ethics

This research followed the Economic and Social Research Council's research ethics framework. Ethical approval was gained from the UCL Institute of Education Research Ethics Committee (REC 1498).

### Project advisory group

Early in the project, we recruited an advisory group that reflected different sets of expertise and perspectives (teachers, parents, citizens with lived experience, public health practitioners, clinicians, etc.). This was facilitated by Co-Production Collective who helped to identify members, run meetings, and liaise with the group throughout the project. Further details of this advisory group can be found in Appendix 1. During recruitment, it became clear that there was strong interest among teachers regarding interventions to improve children's health, which inspired the focus on school-based interventions.

# Methods and results

## WP1: Assessing contextual generalisability in a defined setting

This work package addressed the question, "Can local knowledge of contextual features (in terms of people, intervention, usual care conditions, or other features) be harnessed through co-production of a logic model, and be applied in innovative generalisability analyses?" Here, we describe the development of the logic model and discuss the potential applications of the model; the second part of the question, regarding the application to innovative analytical approaches, is more thoroughly addressed in the section for work package 2.

### Background

The features of an area that determine how well an intervention to improve child health may 'fit' is a form of local knowledge held by stakeholders. This valuable experience-based information about which features may be important and why could be formalised in a way that facilitates common understanding and is actionable using a systems-based logic model. Through the model, such 'local experience-based' knowledge can be contrasted with 'local data' about an area (see later sections) to provide more of a macro-view. A systems-based logic model sets out to theorise aspects of complexity around relationships between intervention and broader context and how these interact. Essentially, it shows a 'theory' about diverse processes that might lead to an outcome.

For this methods development work, we focused on child health related to overweight/obesity. This topic was chosen in part because it has a complex set of factors and outcomes that would benefit from a systems-based approach (PHE 2019); it is a much-researched topic and likely to have usable reviews for our methods development work; and the research team have collective experience of researching this and related topics.

### Development

The co-production workshops were designed to gain insights of local factors relating to child health. There were two initial workshops to develop the model, then a third workshop with a broader audience to check and challenge the model. The meetings were conducted virtually and hosted on Zoom.

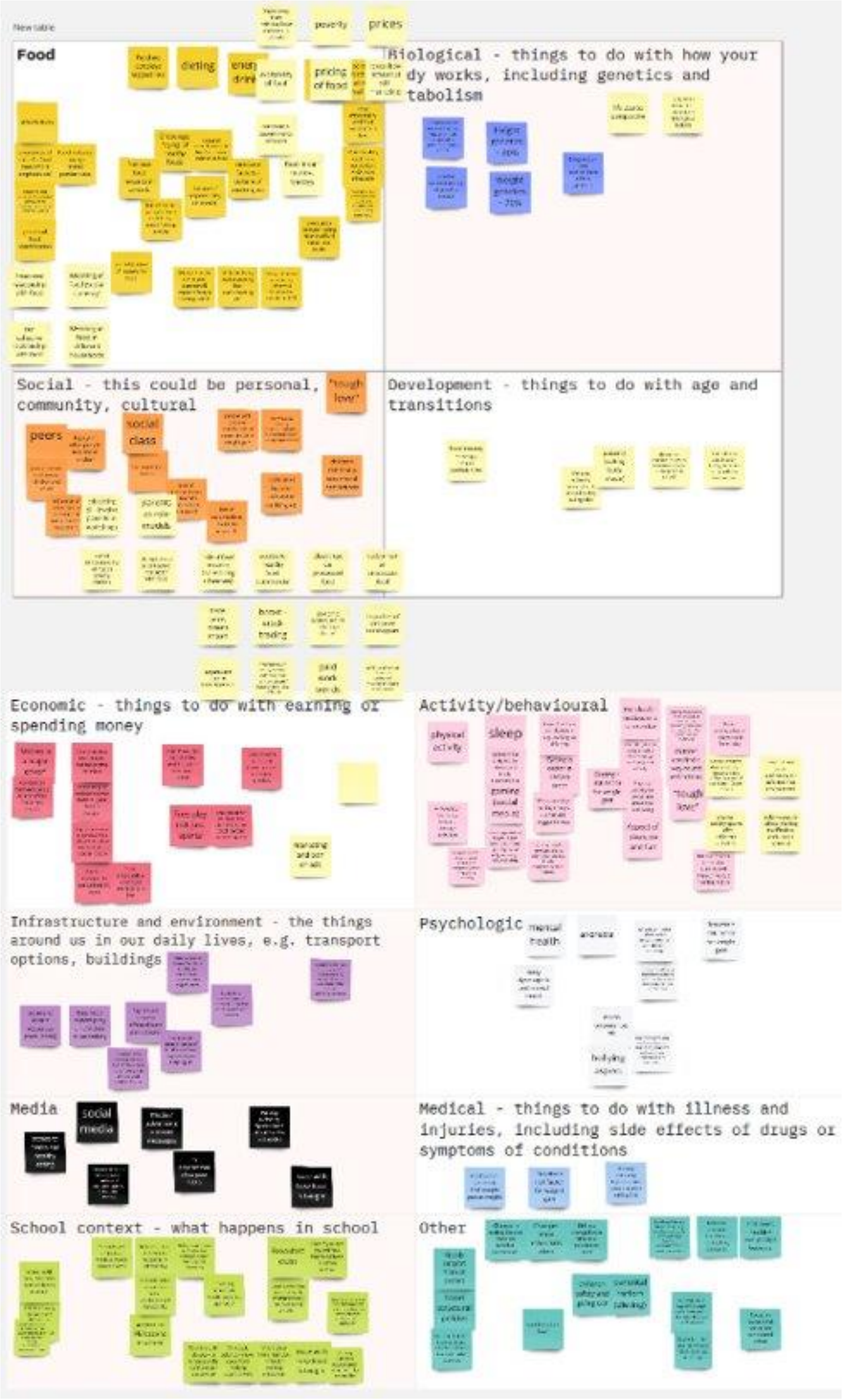
The first two workshops followed a similar pattern. Workshops 1 and 2 started with a brief introduction to the project and its goals, followed by discussions in breakout groups. A facilitator from the project team in each of the breakout groups ensured the ground rules were maintained. The discussions were guided by starter questions. A second project team member posted notes of the points raised on a live, virtual whiteboard (in Miro) (see screenshot after workshop 1 in Figure 1).

In Workshop 1, in addition to the project team, there were 11 attendees, who described their relevant expertise primarily as: GP and advocacy (1); teaching (4); lived experience (4); and research (2).

In Workshop 2, we invited the same people who were invited to Workshop 1.  There were 13 attendees, who described their relevant expertise primarily as: GP and advocacy (1); teaching (3); lived experience (5); research (2); occupational therapy (1); and nutrition and advocacy (1).

A member of Co-Production Collective facilitated both workshops and liaised with participants before and after the events.

*Figure 1: Screenshot of points raised during workshop 1*



New table

poverty

prices

Food

dieting

pricing of food

Biological - things to do with how your body works, including genetics and metabolism

Social - this could be personal, community, cultural

"tough love"

peers

social class

Development - things to do with age and transitions

Economic - things to do with earning or spending money

Activity/behavioural

physical activity

sleep

"tough love"

Infrastructure and environment - the things around us in our daily lives, e.g. transport options, buildings

Psychologic

mental health

bullying

Media

social media

Medical - things to do with illness and injuries, including side effects of drugs or symptoms of conditions

School context - what happens in school

Other

## Draft final model

The workshops' discussions shifted the emphasis of the whole model from childhood overweight and obesity, to children's health and wellbeing with a focus on healthy eating, physical activity, and mental health. The model therefore had these three outcomes at the centre of the model.

Prior to the first workshop, we had identified broad domains of concepts to guide discussions (see Box 1). In Workshop 1 we asked people to consider factors across these different domains. After the first workshop, we recognised that the different factors emerging operated at different socioecological levels with some cross-cutting themes (see Box 2). Both the broad domains and the levels were discussed and refined in workshop 2.

After workshop 2, we had over 1,000 pieces of text from the virtual whiteboard post-it notes to sort out. Our first task was to combine duplicates and simplify similar concepts. Two team members then organised the text into domains and levels using Excel before it was reviewed by a third team member.

**Box 1. Broad domains**

1. Food
2. Biological
3. Social
4. Developmental
5. Economic
6. Activity/ behavioural
7. Infrastructure/ environment
8. Psychological
9. Media

Once the themes/concepts were organised, we needed a visualisation method that: preserved the levels and concepts; could represent concepts 'hierarchically' (from more to less detailed); could represent some 'logical' relationships; and could preserve the original sentiments for reference. We used Miro for the visual representation.

The draft final model is available to view here[1], with a video showing how to navigate the model here.

## Checking and challenging the model

We held a third workshop in which we invited the original participants back, plus a new group of policymakers and practitioners. Workshop 3 had 12 attendees: 7 'returners' and 5 new attendees. The aim of workshop 3 was to check and challenge the logic model, as well as introduce new methodological developments. Participants were provided with a link to the interactive model and a short instructional video prior to the workshop. During the 2-hour workshop, we presented the project progress and then used breakout groups to seek feedback.

The findings from this workshop are presented in the section on Work package 3, and will inform future development and dissemination of the model.

**Box 2. Levels**

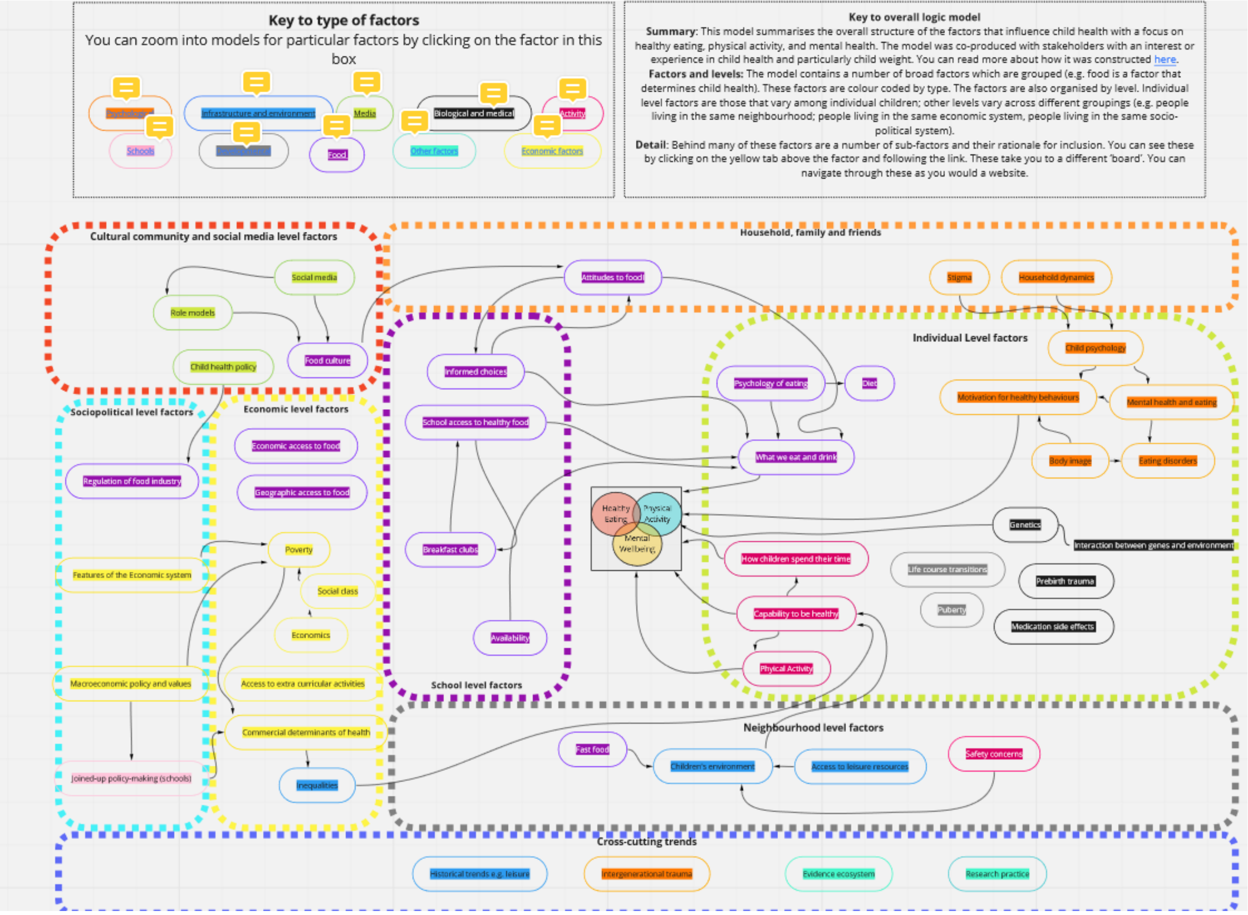
- Individual
- Household, family, and friends
- School
- Neighbourhood (place-based)
- Cultural community (incl. social media)
- Economic systems
- Socio-political, infrastructure, national policy, media
- *Cross-cutting factors (time, history, etc)*

[1] https://miro.com/app/board/uXjVOZrPSC0=/

## Summary of WP1

The key output and contribution of this package is a systems-based logic model that depicts which factors are viewed by stakeholders as important local influencers of child health and which may influence generalisability (Figure 2). Refocusing to child health from child obesity was a key contribution of this workshop that emanated from the co-production activities.

*Figure 2: Snapshot of logic model (full model on website)*



**Key to type of factors**

You can zoom into models for particular factors by clicking on the factor in this box

**Key to overall logic model**

**Summary**: This model summarises the overall structure of the factors that influence child health with a focus on healthy eating, physical activity, and mental health. The model was co-produced with stakeholders with an interest or experience in child health and particularly child weight. You can read more about how it was constructed here.

**Factors and levels:** The model contains a number of broad factors which are grouped (e.g. food is a factor that determines child health). These factors are colour coded by type. The factors are also organised by level. Individual level factors are those that vary among individual children; other levels vary across different groupings (e.g. people living in the same neighbourhood; people living in the same economic system, people living in the same socio-political system).

**Detail**: Behind many of these factors are a number of sub-factors and their rationale for inclusion. You can see these by clicking on the yellow tab above the factor and following the link. These take you to a different 'board'. You can navigate through these as you would a website.

## WP2: Adjusting for generalisability and examining the influence of context (10 pages)

This work package sought to address the following research questions:

- Which statistical approaches (recalibration, binomial logistic regression, enhanced subgroup analysis) can adjust for context to enhance the generalisability of a completed meta-analysis?
- How might Qualitative Comparative Analysis (QCA) help identify contextual features that trigger more successful interventions, or hinder positive effects?
- Do the approaches offer a robust way to consider health inequalities in the systematic review? (we consider the merits and caveats of the different approaches proposed throughout)

### Selecting systematic review test cases and Local Authority areas (districts)

#### *Selection of the reviews*

We searched for reviews of interventions conducted in schools to improve child health. We appraised candidate reviews based on the following criteria (see details in Appendix 2):

- Design: systematic review
- Intervention: school-based
- Contextual characteristics: broad range
- Outcomes: broader than Body Mass Index
- Analysis: presence of meta-analysis
- Heterogeneity: variation in impacts (statistical heterogeneity) and in contexts from which studies drawn (contextual heterogeneity)

A longlist of thirteen studies was identified (see Appendix 2) and two reviews were selected. The two reviews were selected for being well-conducted and best meeting the above criteria. Langford, Bonell et al. (2014), explored the effectiveness of the World Health Organization's (WHO's) Health Promoting Schools (HPS) framework. Andermo, Hallgren et al. (2020), explored the impact of school-based physical activity interventions. Langford et al., is a smaller review and in line with a typical public health systematic review so was the first choice for testing the approaches; Andermo et al. was used for two of the approaches where Langford did not have sufficient variation. This decision in itself is informative regarding the possible review scenarios in which each approach could be deployed.

#### *Characteristics of the Langford Review*

Langford, Bonell et al. (2014) synthesised data from trials where members of the school body had provided input to the intervention; had implemented changes to the school's ethos and/or environment; and where there was engagement with families or communities. A subset of seventeen studies explored the impact of HPS in reducing dietary fat intake, and the reanalysis focused on a group of ten physical activity and nutrition intervention studies. Overall, Langford, Bonell et al. (2014) found that the ten studies were ineffective in reducing dietary fat intake (SMD: -0.04; 95% CI: -0.20-0.12).

### *Characteristics of the Andermo Review*

The review by Andermo, Hallgren et al. (2020) included 30 different interventions within the main review and assessed eleven potential outcomes resulting from school-based physical activity interventions. We selected one of these outcomes - positive mental health. In the original review, a meta-analysis on positive mental health was supported by 25 studies, and overall, the intervention appeared effective (SMD: 0.405; 95% CI: 0.208-0.603). We reanalysed these data including only RCT studies and estimated a slightly larger effect size based on 21 studies (SMD: 0.439; 95% CI: 0.186-0.691), albeit with substantial heterogeneity ($I^2$: 97.5%).

### *Selecting a Local Authority*

In the UK, local government areas or districts are called 'local authorities'. Selection of Local Authorities (LA) for the recalibration explored a mix of: urban/rural, high/low advantage, and high/low ethnic diversity, among other characteristics (Appendix 3). These factors were informed by the logic model. The four LAs chosen were: Liverpool, Test Valley, Camden, and Islington.

## Approach 1: Recalibration in meta-analysis using data from the Langford review

**The use scenario and alignment with decision-making needs: I**n this approach, we set out to address a question a decision-maker may ask when trying the interpret the findings of a systematic review: '*is there any evidence that the intervention would work differently in an area like mine (for example Liverpool or Islington etc.)?*"

**How does this approach differ from current meta-analytic approaches?** Within a fixed effect model, the contribution of an individual study to the pooled effect size is determined by the inverse of the 'within study variance' (which is proportional to the sample size of the study). In a random effects model, the contribution of each study to the pooled effect size is weighted to account for variation within studies (as with a fixed effect model) as well as between study variation. In this standard meta-analytic practice therefore, large trials that may have low contextual relevance to a decision-maker may account for much of the pooled effect size, particularly in a fixed effect model.

Conventional techniques for investigating whether a study's context influences the outcome are restrictive:

1) Firstly, conventional ways of examining heterogeneity in meta-analysis tend to only allow for a restricted number of parameters to be explored simultaneously (e.g., in sub-group analysis or meta-regression).
2) Secondly, even if techniques such as sub-group analysis can be refined (see Approach 3), this still leaves the meta-analyst producing an estimate for a subgroup of studies, potentially meaning that the decision-maker overlooks the totality of evidence to only consider a subgroup of studies. This could mean that studies that differ in terms of 'surface similarity' (e.g. the country in which the intervention was conducted) are discounted as irrelevant despite being contextually similar in most other ways.

The approach outlined here involves incorporating an additional contextual relevance parameter into the calculation of the contribution of each study to the pooled effect size. This involves upwardly weighting studies that are more similar to the decision-making context of interest (here we focussed on particular LAs and downwardly weighting those less similar. This approach is named recalibration and is modelled on ideas that we developed earlier (Kneale, Thomas et al. 2019).

**How is the approach deployed?**

The approach focuses on calibrating the effect size so that, rather than representing a hypothetical population most similar to the 'mean' of the studies in the analysis (as in a typical meta-analysis), the effect size represents the population of interest. Recalibrating the effect sizes could be achieved in numerous ways; the approach taken here is as follows (further details in Appendix 4):

- ***Identifying relevant characteristics in studies and LAs and extract data***
- ***Coding the studies and harmonising the data***
- ***Sourcing and coding locality data***
- ***Creating similarity matrix based on multiple factors***
- ***Creating weight that includes the inverse of the variance***
- ***Running the models and generating the results:***

While overall the evidence suggests that physical activity and nutrition interventions are not effective in reducing fat intake, this interpretation changes when we place greater emphasis on the similarity of studies to particular settings. Namely, for both a fixed effect and random effects specification, studies that are more like Liverpool, or Camden and Islington are more effective and consequently contribute more towards the pooled effect size. In turn, the recalibrated effect size generated indicates a larger effect with a narrower confidence interval; in these areas we have greater confidence that school-based physical activity and nutrition interventions will have an impact on reducing fat intake, and while the anticipated effect size remains small it may be substantial at a population level. In contrast, in Test Valley (Hampshire), although we can observe a small change in the effect size itself, the interpretation does not change and remains that there is no evidence that the intervention is observed to reduce fat intake.

*Table 1 Estimates of effectiveness of physical activity and nutrition interventions in changing fat intake including recalibrated estimates (alternative estimates using different specifications are available on request from the authors)*

| Model specification | Model results (SMD reduction in consumption of fat in diet) | | | |
|---|---|---|---|---|
| Original Fixed effect | ES: 0.000; 95% CI: -0.001 to 0.001; I2: 95.0% | | | |
| Original Random effects | ES: -0.042; 95% CI: -0.204 to 0.120; I2: 95.0% | | | |
| | Recalibrated Local Authority Estimates | | | |
| Local Authority | Liverpool | Test Valley | Camden | Islington |

| | | | | |
|---|---|---|---|---|
| Recalibrated Fixed effect (Canberra metric) | ES: -0.070; 95% CI: -0.131 to -0.008 I2: 99.9% | ES: -0.057; 95% CI: -0.128 to 0.010; I2: 99.9% | ES: -0.080; 95% CI: -0.143 to -0.017; I2: 99.9% | ES: -0.071; 95% CI: -0.138 to -0.006; I2: 99.9% |
| Recalibrated Random effects (Canberra metric) | ES: -0.066; 95% CI: -0.130 to -0.002; I2: 99.9% | ES: -0.057; 95% CI: -0.123 to 0.010; I2: 99.9% | ES: -0.073; 95% CI: -0.138 to -0.008; I2: 99.9% | ES: -0.067; 95% CI: -0.133 to -0.001; I2: 99.9% |

Given the high levels of heterogeneity, we focus our further interpretation on the results from random effects models. In the standard random effects model, two studies (Colín-Ramírez 2009, Levy, Ruán et al. 2012) account for 21.6% of the weighting to the overall pooled effect size combined (Figure 3); both studies suggest that children ate a greater amount of fat in their diets after the intervention was conducted compared to the control group. These studies are found to be contextually dissimilar to Liverpool, and in the recalibrated model, these studies account for 16.2% of the weighting of the pooled effect size (Figure 4); in contrast, in the recalibrated model for Test Valley these studies account for 19.8% of the pooled effect size combined (Figure 5). Meanwhile, if we focus on the contribution of Haerens 2006, we can see in the model recalibrated for Liverpool, that this study contributes 16.5% towards the pooled effect size in contrast to 11.4% in the standard random effects model.

*Figure 3 Random effects model of impact of interventions on fat intake*

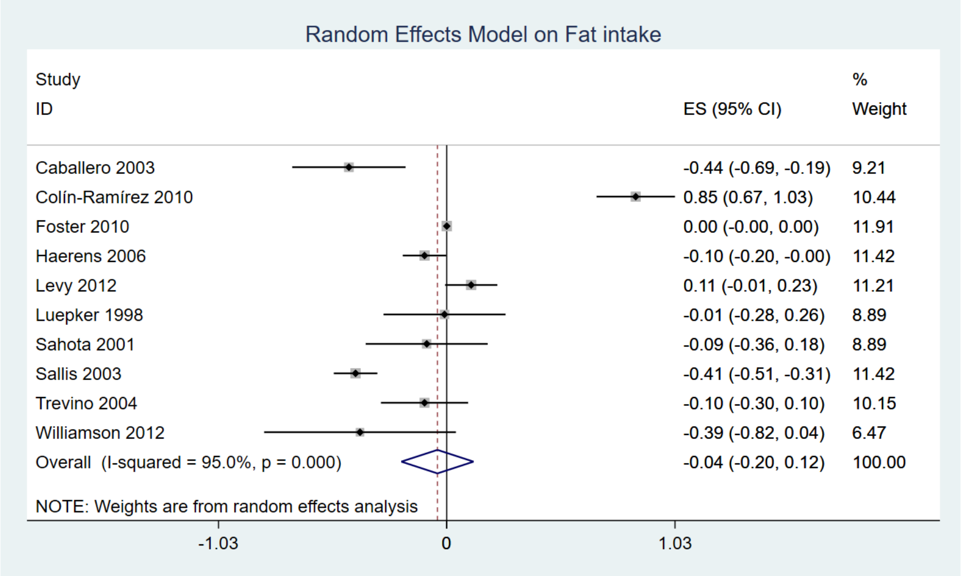


*Figure 4: Random effects model of impact of interventions on fat intake – Recalibrated for Liverpool*

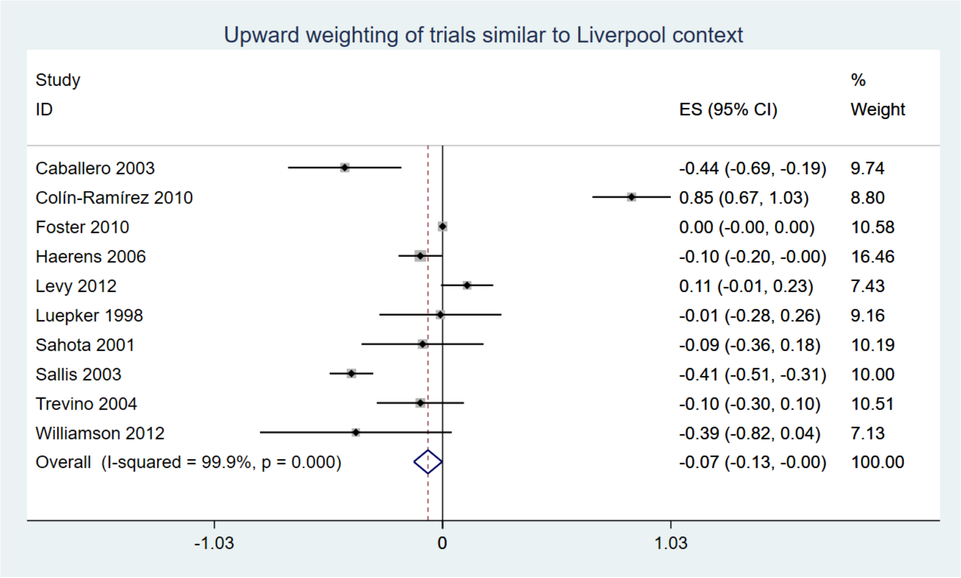
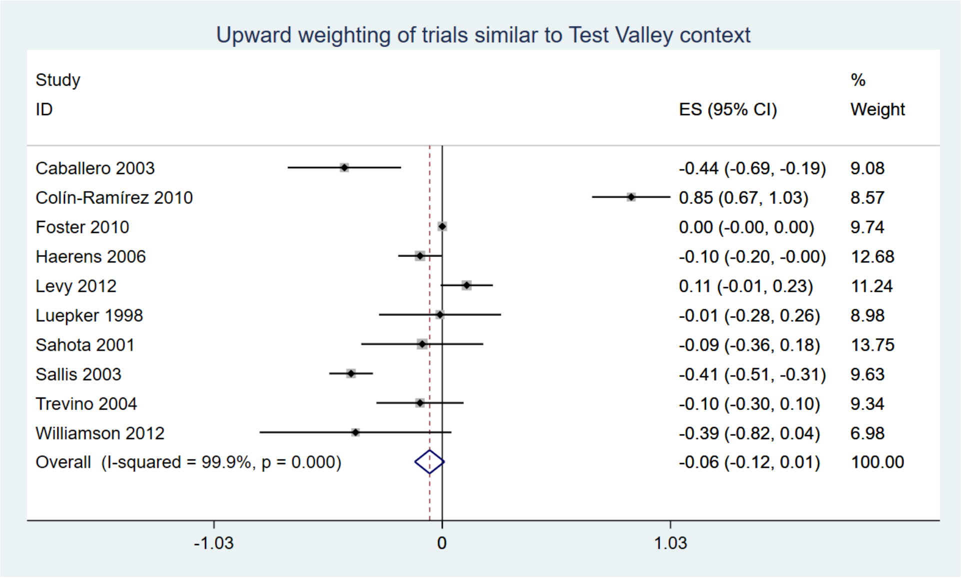


*Figure 5: Random effects model of impact of interventions on fat intake – Recalibrated for Test Valley*



Further explorations suggest that the study conducted by Haerens (2006) and Liverpool shared similarities in context in having relatively low levels of ethnic diversity, high levels of socioeconomic adversity (relatively high in the case of Haerens), similar profiles in terms of fat consumption/environment, and both are settings in high income countries; there were differences in the gender profile (Haerens involved substantially more boys than girls) and in the broader healthcare system. Of all the studies in the meta-analysis, Haerens (2006) was found to be most like Liverpool along these dimensions.

### Further interpretation of the results, and strengths and caveats

This approach aims to address the question, ‘is there any evidence that the intervention will work differently in an area like mine?’ In this example, we find evidence that studies with greater similarity to Liverpool, and Islington and Camden, may be more impactful, and consequently recalibrated estimates suggest small but potentially impactful reductions in fat consumption. In contrast, the interpretation for Test Valley is unchanged, and we would not expect any change in fat consumption. Given that all three localities where the interpretation of evidence changes have relatively high levels of socioeconomic adversity and that two of these areas have high numbers of children from minoritised communities, the data could suggest that whole school interventions involving physical activity and nutrition are likely to have greater impacts in these contexts.

There are caveats to note. Firstly, the characteristics selected and extracted from studies and localities reflecting context and populations may have negligible impact on the actual effect size. In addition, the breadth of factors identified as potentially influential through the logic model was not matched by the breadth of contextual factors available in the studies. A second caveat is that each factor was given equal weighting in the dissimilarity matrix when creating study weights for the meta-analysis. Despite these, this approach is responsive to a clear need among decision-makers for contextually relevant estimates and, alongside usual meta-analytic practice, can provide additional evidence when decisions are being made about which interventions to commission. The approach outlined here builds on our previous work, but importantly we have developed the approach in terms of how factors are considered, and how the weight is constructed and scaled alongside other weighting components (see technical details in Appendix 4).

## Approach 2: Binary Logistic Regression in meta-analysis (using data from the Andermo review)

**The use scenario and alignment with decision-making needs:** This approach focusses on the scenario where the decision-maker places less value on evidence that seeks to provide a precise estimate with a measure of uncertainty, but is instead more concerned with understanding broad-brushed evidence about the effectiveness of an intervention, and the role in which contextual characteristics play in determining this decision. The decision-making question being addressed by this approach is *“Based on this heterogeneous evidence, I want to know if this intervention is generally a good idea for my area/needs?”*

**How does this approach differ from current meta-analytic approaches?** The usual approach in meta-analysis is to take a weighted average of individual study effect sizes, and as an addition, to explore which factors may help to explain any observed heterogeneity. Here we set out to examine the study level characteristics that are associated with the odds of being part of a successful or unsuccessful study, and we treat each study’s data as an example of aggregate data. This is a new approach to working with effect sizes where the focus is on understanding drivers of heterogeneity across effect sizes, rather than estimating the magnitude or occurrence of successful interventions.

The interest in this approach is motivated by (i) exploration of the influence of contextual factors without having to make evidence claims about the precision of these associations (as is the case for approaches 1 and 3); and (ii) by the possibility that allocation into a ‘successful’ and ‘unsuccessful’ group may arise from multiple factors. This approach was initially tested with studies in the Langford Review (2014), but was later deployed using data from the Andermo Review (2020) as the approach needed greater variation in the contextual data.

## How is the approach deployed?

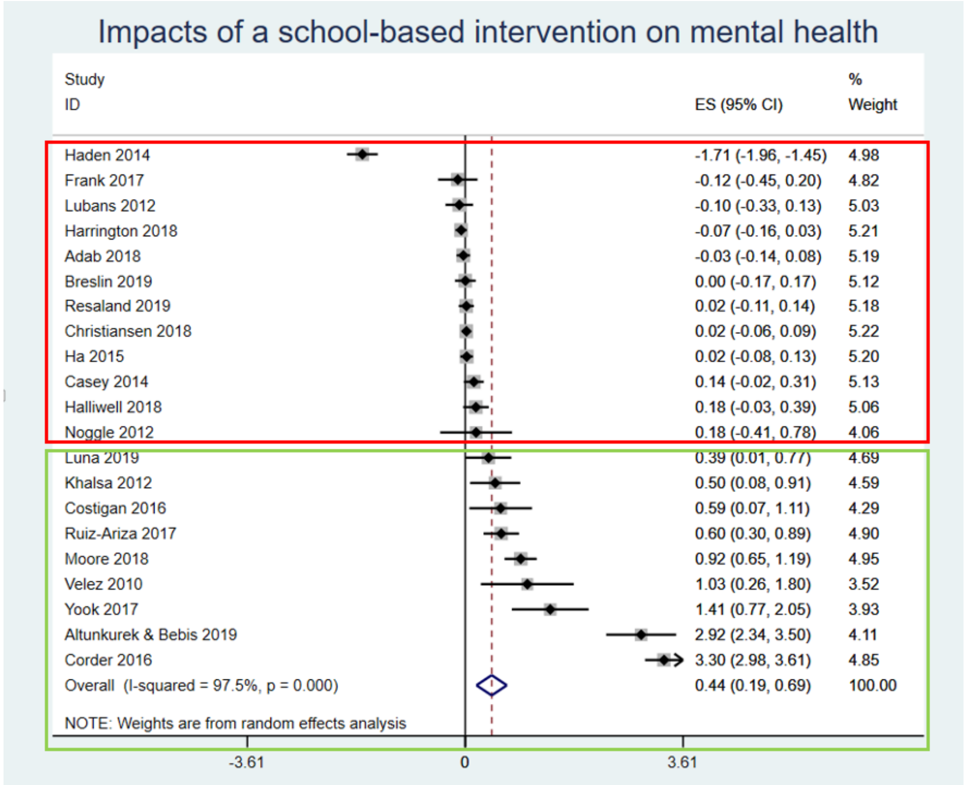
In summary, the approach entails:

- ***Using the logic model as an anchor, extract relevant characteristics from studies***
- ***Harmonise the data (create rules for categorising data)***
- ***Classify the studies as in/effective based on the outcome***

Further details are contained in Appendix 5.

In the Andermo review, we conducted a random effects meta-analysis of the 21 RCT studies. This revealed a group of 9 studies where the impact was clearly positive with a 95% confidence interval that did not cross the line of no effect (green box in Figure 6), as well as 12 studies where the impact was uncertain or harmful (red box in Figure 6). A new binary variable was created and studies allocated 0 or 1 based on the categories in Figure 6.

*Figure 6: Random effects model of impact of interventions on positive mental health*



- ***Treat each study as aggregate data and create weights that reflect the sample size***
- ***Construct regression model***

At first, we attempted a model that included all four covariates (age, gender, ethnicity, and socioeconomic status), although found that parameters were not estimated due to a lack of variation in the sample. We then constructed a model containing study-level covariates on the age and gender of the children. The output (table 2) indicates that the odds of a study reporting a significant improvement in positive mental health following a school-based physical activity intervention significantly improved when the study had older participants and, to a lesser extent, when the study had more males than females included as participants.

*Table 2 Odds of study reporting significant improvement in positive mental health by selected study-level characteristics (weighted by sample size)*

| Covariate | Odds ratio |
|---|---|
| **Older participants (aged 13+ vs younger)** | 29.23*** |
| | (2.186) |
| **More males than females** | 1.226*** |
| | (0.465) |
| Pseudo R2 | 0.3134 |
| Observations | 12,480 |

Predicted probabilities were calculated to ease interpretation (table 3). These emphasise that where the intervention was conducted among older participants with more boys involved than girls, that the predicted probability of a study being effective was much higher than if the study was conducted among younger participants with more girls than boys participating.

*Table 3 Predicted probability of study reporting significant improvement in positive mental health by selected study-level characteristics (weighted by sample size)*

| | **Younger Participants (under 13)** | **Older Participants (over 13)** |
|---|---|---|

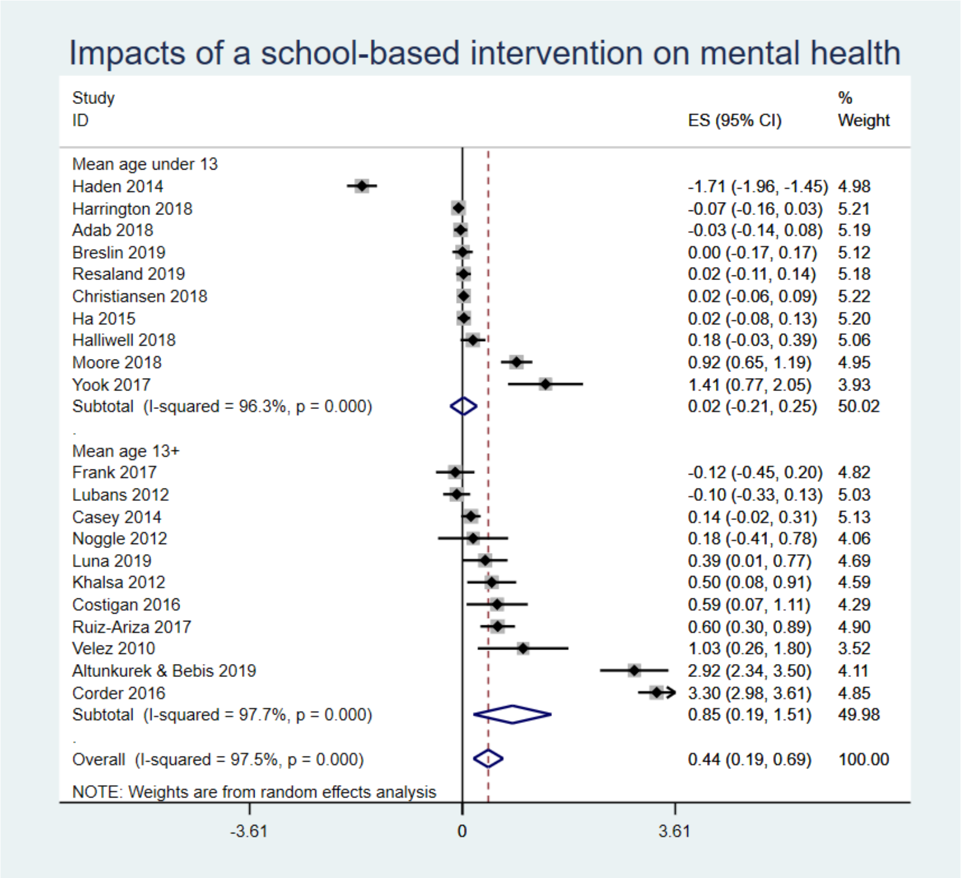
| | | |
|---|---|---|
| **More females than males** | 2.8% | 46.0% |
| **More males than females** | 3.6% | 51.1% |

### Further interpretation of the results and strengths and caveats

We can further explore in the data whether age really did influence the results of studies to this extent. Approximately half of the studies were conducted among participants with a mean age of 13 and above, with the majority of these studies being classified as effective (7/11 studies); in contrast just two studies conducted with younger children were classified as effective (2/10 studies). When we consider the number of participants within these studies, we find that while around half of participants in studies with older participants were in the effective group (47.6% of 2,253), just 3.2% of participants in trials with younger participants were in the effective group (3.2% of 10,227).

We also compared the results of this approach with a conventional approach involving subgroup analysis based on age (Figure 7, below). This showed a similar trend, offering tentative evidence that the intervention is more likely to be effective with older children, albeit with overlapping confidence intervals between subgroups of studies and high within-group heterogeneity impeding the extent to which this evidence could be considered useful for decision-making.

*Figure 7: Random effects model of impact of interventions on positive mental health, sub-grouped by average age of participants*



- ***Further interpretation of the results, and strengths and caveats:***

The results indicate that the design and targeting of current school-based physical activity interventions may be better suited to older children (and to a lesser extent, males). We believe that the message emanating from the evidence is clearer for decision-makers using the binary logistic regression approach, where the focus is less on estimating an average effect with precision, and more on adopting a configurative approach (Gough, Thomas et al. 2012) to understanding why some studies are more effective in generating an impact than others. With this approach, focussing on the results based on age, we may communicate to a decision-maker that 'there is clear evidence that school-based physical activity interventions with older children are more likely to be effective than interventions with younger children'. Such a message could only be tentative with conventional subgroup approaches.

There are caveats to this approach that need further exploration. In particular, there is an argument that this type of approach casts aside valuable information about the estimated magnitude and precision of anticipated effects. This simpler message could engender false confidence about the strength of evidence and likely impacts although we would also emphasise that this type of approach and the language of probabilities, odds, and likelihood does inherently communicate uncertainty in the estimate. Pragmatically, this type of approach is reliant both on statistical heterogeneity (for allocation of studies in groups based on effectiveness), and contextual heterogeneity (to explore associations between contextual factors and group allocation). This requires a larger meta-analytic dataset which, as our experience of selecting a review in the well-researched area of child health has shown, is surprisingly difficult. However, there are some perceived advantages. Firstly, this an advance on vote counting of effect sizes, which have been deemed to be crude, flawed and worthless (Bushman and Wang 2009), as we have incorporated information about the sample size into the weighting; there may be further ways of weighting the studies to explore in the future. In addition, unlike vote counting, we are not seeking to understand or make judgements about the intervention as a whole and its effectiveness, but only to understand characteristics that are associated with more effective interventions. Secondly, the approach may better align with decision-making needs. Thirdly, there are opportunities with larger meta-analytic datasets to extend the approach to examine groupings of studies through multinomial regression models. Finally, the substantial advantage we perceive is that this approach is simple to implement and simple to interpret.

## Approach 3: Enhanced subgroup analysis (using data from the Langford Review)

**The use scenario and alignment with decision-making needs:** This approach focusses on the scenario where a decision-maker may not have a clear idea of a specific geographic context within which the evidence is to be applied. Instead, the decision-maker in this scenario may want to develop a better understanding of the way in which contextual and population characteristics influence the effectiveness of the intervention. The idea here is to improve current practice in investigating contextual drivers of heterogeneity through adopting a ‘multivariate’ approach to subgroup analysis.

**How does this approach differ from current meta-analytic approaches?** Standard meta-analytic practice is to deploy subgroup analysis based on a single characteristic, which can lead to  methodological challenges (for example a form of confounding where repeated subgroup analyses differentiate a similar set of studies each time based on different characteristics that are treated as independent). Extensions to exploring heterogeneity such as meta regression offer opportunities to explore multiple factors, although are often not feasible or are uninformative due to the number of studies in a typical meta-analytic dataset.  This approach involves investigating natural groupings of studies based on multiple characteristics simultaneously to form the basis of subgroup analysis.

### How is the approach deployed?

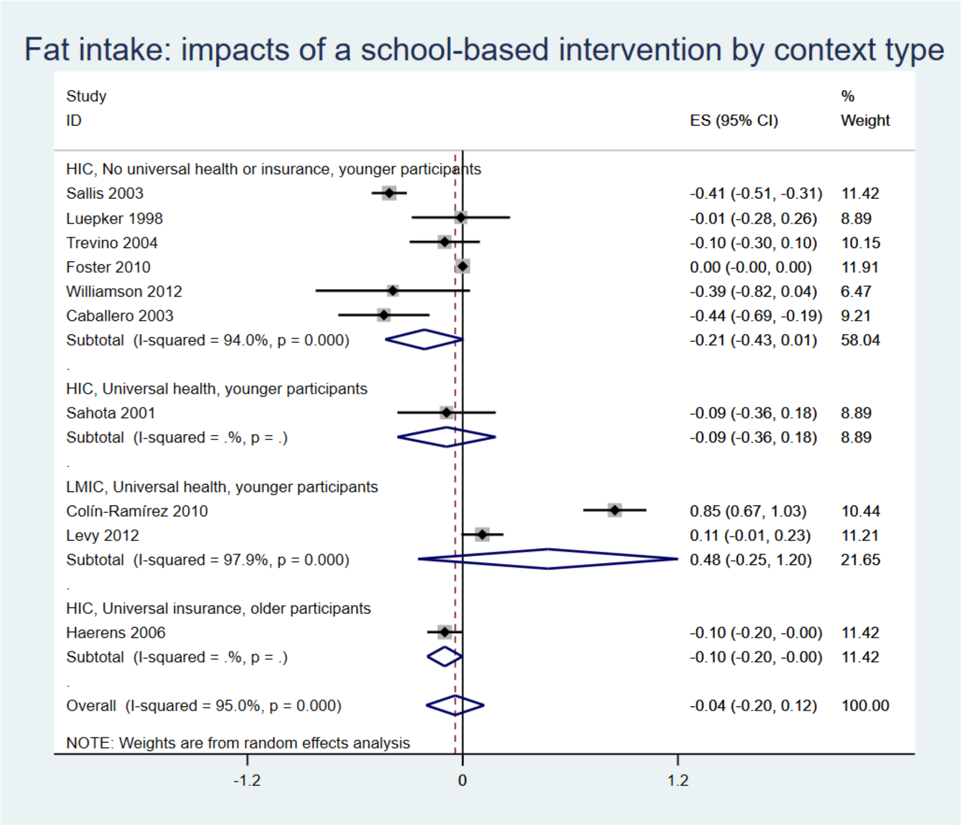
In summary, the approach entails:

- ***Using the logic model as an anchor, extract relevant characteristics from studies***

- ***Code the studies***
- ***Create similarity matrix based on multiple factors***
- ***Apply clustering algorithm to identify groupings in the data***
- ***Explore the features of the clusters***
- ***Run the analysis with new subgroups***

Further details are contained in Appendix 6. Using the data from the Langford review, four clusters of studies were identified (one larger cluster of six studies, a cluster of two studies, and two ‘outlier’ studies). The meta-analysis (Figure 8) indicates that the intervention was effective among the first subgroup of studies which took place in high income contexts among younger participants where there is no universal health service or universal health insurance; this group also had relatively low levels of socioeconomic advantage. This may suggest that whole school interventions are particularly effective among children who may otherwise have little contact with healthcare services through universal provision. However, the estimate for this subgroup of studies also overlaps with the estimate for all other subgroups, ultimately providing inconclusive evidence that studies with these contextual characteristics are more effective. We can also see that the $I^2$ remains high within this subgroup of studies, emphasizing that within the subgroup that there is substantial heterogeneity.

*Figure 8: Random effects model of impact of interventions on fat intake with enhanced subgroup analysis*



## Further interpretation of the results and strengths and caveats

In this example, we wanted to address a question that a decision-maker may have around *'What can you tell me about the impact of context on the outcome?'*. We found tentative evidence on the influence of context using this approach – namely that a group of mainly US studies have larger effect sizes which may be attributable to weaker health systems and larger pre-existing inequalities. Ultimately however, the evidence generated using this approach was inconclusive and there are additional methodological caveats to note. These include additional sources of confounding and that analyses may be underpowered to detect effects (Burke, Sussman et al. 2015); certainly the subgroups with low numbers of studies as observed in Figure 4 are of questionable analytical value. However, the advantages of this approach are that it allows multiple factors to be considered simultaneously to develop a contextual profile to help understand drivers of heterogeneity, thereby aligning with decision-making needs. While this may be a promising approach to understanding context in other, larger, reviews; in this example its application has been more limited.

## Approach 4: Qualitative Comparative Analysis (using data from the Andermo review) (1031)

### The use scenario and alignment with decision-making needs

QCA could be applied where a local decision-maker wants to know whether the evaluated interventions fit their population, area, or context. Systematic reviews have been criticised for providing insufficient evidence about the 'essential elements' of an intervention to facilitate practice (Glasziou, Chalmers et al. 2014). This approach has similar aims to binary logistic regression, whereby the aim is to examine whether contextual characteristics explain why some studies are effective and some are not. However, instead of only examining the impact of context on outcomes, QCA takes a complexity perspective and assumes that intervention, contextual/implementation factors will interact with each other in unpredictable ways. Thus, the decision-making question being addressed here is *'Which intervention factors are required to achieve success in my context?'*

### How does this approach differ from meta-analytic approaches

QCA is a relatively new approach in systematic reviews that draws on set-theory (see brief explanation in Appendix 7). It allows for multiple pathways leading to the same outcome, rather than a linear additive model as in meta-analysis, so is suitable for evaluating interventions that might have different combinations of active ingredients (Thomas, O'Mara et al. 2014). It also allows for more features (contextual, implementation, evaluation, etc.) to be explored in analysis than in a meta-analysis, because QCA does not have the same statistical assumptions that require more data points per feature (variable).

### How is the approach deployed?

We used data from the Andermo review on the analysis of the 21 RCTs measuring positive mental health (where 'positive mental health' is a broad outcome that covers a range of mental health measures, such as wellbeing and self-esteem). QCA typically involves six key stages (Thomas, O'Mara et al. 2014). For this project, we only completed stage 1, which involved building a data table to capture the outcomes, contextual features, and intervention/implementation factors of each study in a matrix. As described below, the learning from this stage was sufficient to understand the potential application of QCA to contextual analyses.

We extracted positive mental health outcomes from the review for each of the 21 cases (studies). We created sets for the 'highly effective' cases (those with a Hedges g effect size >1);

‘moderately effective’ (Hedges g between 0.4 and 0.99), ‘minimally effective’ (Hedges g between 0 and 0.39) and ‘harmful’ (Hedges g < 0).

For our contextual conditions, we extracted information on participant characteristics (age, gender, SES, and ethnicity). For intervention and implementation conditions, we employed two approaches. First, an iterative and exploratory approach to ensure we did not restrict knowingly to preconceived ideas. Through this approach we identified three potentially important conditions; **choice** (whether children were offered choice in the intervention activities), **tailoring** (whether the activities were individualised), and **fun** (whether the activities were designed to be fun). The second approach was theory-informed; using a conceptual model from a systematic review of the effects of physical activity on cognitive and mental health in children (Lubans, Richards et al. 2016) (see Appendix 7 for details of the rules for set assignment, and the extracted data for each study used to support the decisions).

## Results

Table 4 below provides basic details of each of the cases and the data table illustrating the set scores for each condition for each case. The table presents cases ordered by outcome values; in the top five rows are cases with negative impacts on children’s mental health; followed by eight cases showing minimal impacts; four cases showing moderate impacts; and in the bottom four rows are cases with the greatest impacts. This ordering allows visual identification of potential patterns of association between the outcomes and the contextual and intervention conditions.

***Approach 1: Examining contextual features****: A*mong the 21 cases, there were few interventions that were conducted in low SES contexts (n=5), delivered to participants predominantly from minority ethnic groups (n=5), or to single-sex (female) groups (n=3). QCA requires that at least one third of cases display the conditions of interest (Wagemann and Schneider 2007); therefore, it was not possible to proceed analysis with any of these individual conditions.

***Approach 2: Examining contextual fit****:* We considered assigning these cases to a ‘delivered to disadvantaged groups’ condition. However, since all but one of these cases (8 of 9 cases with at least one of the conditions) were cases with harmful or minimal outcomes (the exception being Velez et al. 2010), the cases did not display sufficient variation in outcomes to enable examination of the intervention conditions for achieving successful outcomes with disadvantaged groups.

***Approach 3: Examining whether intervention conditions may explain poor outcomes in disadvantaged contexts****:* One possible conclusion from the data table is that it is challenging to achieve successful outcomes in contexts of disadvantage. However, QCA reflects a complexity theory perspective in which it is expected that contextual and intervention features will interact to impact on outcomes. Thus, we could not dismiss the possibility that the poor outcomes observed in disadvantaged contexts could equally be explained by intervention characteristics. Thus, we proceeded with an analysis of intervention conditions. As illustrated in the intervention condition columns, the identified intervention conditions are much more prevalent among the moderate impact and high impact cases than the minimal and harmful impact cases.

*Table 4 Case details and QCA data table (n=21 cases)*

| Case details<br>First author *(Date)*<br>*Country: Brief intervention details* | Out-come | Contextual conditions | | | | Theory-informed intervention conditions | | | | | Exploratory intervention conditions | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *Low SES* | *Ethnic minority*[1] | *Female only* | *Older (>12 yr)* | *Social interaction* | *Sense of mastery* | *Body image* | *Aut-onomy* | *Self-regulation* | *Fun* | *Tailor-ing* |
| Haden *(2014) USA*: *Yoga* | 0 | 0 | 0.66 | 0 | 0 | 0 | 0.33 | 0 | 0 | 0 | 0 | 0 |
| Frank *(2016) USA: Yoga* | 0 | 1 | 1 | 0 | 1 | 0.33 | 0 | 0.33 | 0 | 1 | 0 | 0.33 |
| Lubans *(2012) Australia: PA & nutrition* | 0 | 1 | 0 | 1 | 1 | 0.33 | 0 | 0 | 0 | 0 | 0 | 0 |
| Harrington *(2018) UK: PA* | 0 | 0 | 0.66 | 1 | 1 | 0 | 0 | 1 | 0.33 | 0 | 0 | 0 |
| Adab *(2018) UK*: *PA and nutrition* | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| Breslin *(2019) Ireland*: *PA* | 0.33 | 1 | 0 | 0 | 0 | 0 | 0 | 0.66 | 0 | 0 | 0.66 | 0 |
| Resaland *(2015) Norway: PA* | 0.33 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0.66 | 0 |
| Christiansen *(2017) Denmark*: *PA* | 0.33 | 0 | 0 | 0 | 0.5 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| Ha *(2015) Hong Kong: Skipping* | 0.33 | 0 | 0 | 0 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Casey *(2014) Australia: PA* | 0.33 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0.33 | 0 | 0 |
| Halliwel *(2018) UK: Yoga* | 0.33 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0.33 | 0 | 1 | 0.33 |
| Noggle *(2012) USA: Yoga* | 0.33 | 0 | 0 | 0 | 1 | 0 | | 0 | 0.33 | 1 | 0 | 0.33 |
| Luna *(2019) Spain: PA* | 0.33 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0.33 | 0.33 | 0.66 | 0.33 |
| Khalsa *(2012) USA: Yoga* | 0.66 | 0 | 0 | 0 | 1 | 0 | 0.33 | 0 | 0 | 1 | 0.66 | 0.33 |
| Costigan *(2016) Australia: HIIT* | 0.66 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0.66 | 0 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ruiz-Ariza *(2019) Spain: HIIT* | 0.66 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0.66 | 1 | 0.66 | 0 |
| Moore *(2018) Australia: Martial arts* | 0.66 | 0 | 0 | 0 | 1 | 1 | 0.33 | 0 | 0 | 0.66 | 0 | 0.33 |
| Velez *(2010) USA: Resistance training* | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0.66 | 1 |
| Yook *(2017) South Korea: Yoga & new sport* | 1 | 0 | 0 | 0 | 0 | 1 | 0.33 | 0 | 0.33 | 1 | 0.66 | 0.33 |
| Altunkurek *(2019) Turkey: PA* | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0.33 | 1 | 0 | 0.66 |
| Corder (2016) UK: *PA* | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0.33 | 0 | 1 | 0.33 |

Refer to Appendix 7 for details of set scores; [1] Mostly, PA = General physical activity intervention (not a specific type); HIIT = High Intensity Training.

### Further interpretation of the results, strengths, and caveats

QCA was used to address a decision-maker's question '*Which intervention factors are required to achieve success in my context?'* We were unable to answer this because data on context was limited. In the other approaches, this was not an issue because data was handled differently. We did, however, find certain multiple and interacting intervention conditions that are important for effective interventions.

The absence of data related to context could be a limiting factor for future applications of QCA. This is not to suggest such data does not exist, just that it might be underreported.

An advantage of QCA is that it can cope with multiple sources of heterogeneity. By being able to explore both heterogeneity in context and intervention content or implementation, it can potentially provide a more nuanced answer to what may work locally. Future analyses on interventions delivered uniquely in contexts of disadvantage may be able to identify important intervention features for those contexts.

## WP3: Evaluating and disseminating

This work package sought to address the questions of how the approaches from WP2 should be used in review production and decision-making, and how are the findings viewed by stakeholders. Insights to these questions were primarily gained through the academic seminar (December 2021) and a policy and practice focussed workshop (January 2022). Co-Production Collective produced a report on the co-production activities, which included findings relevant to these questions; relevant extracts are reported here and the full report is available from the authors.

### How should the approaches from WP2 be used in review production and decision-making?

Participants in the *academic seminar* raised important points and questions about the potential application of these approaches in future reviews and decision making. Positive comments indicated that the approaches sounded promising as a way to focus attention on local context. They also suggested that underpinning the analytical models with understanding gained from the logic model was a valuable step forward.

Concerns, however, were raised about whether the approaches could be applied in other topic areas. For example, one participant raised the question of whether a lack of variation in other systematic reviews limit the contribution of the recalibration approach in future? Would a lack of reported information about contextual features hamper application of any of the approaches?

In the academic seminar, there was further discussion around what the logic model could tell us about the strength of relationships. There appeared to be a general view that understanding the strength of the evidence on the pathways in the model would be informative both in terms of priority setting for decision making and for determining aspects of the new synthesis approaches. This echoed a sentiment of the advisory group.

Participants in *workshop 3* also discussed the use of the approaches, however, due to time restraints, this audience was only introduced to the recalibration approach and the binomial logistic regression approach. This audience received less technical detail about the approaches than in the academic seminar, and perhaps were limited in what they could comment on.

For the recalibration approach, one workshop 3 participant suggested that including context in the synthesis could be used as an excuse for not engaging with research because people could claim that it is too different from their location to be informative. For the binomial logistic regression approach, there was a general sense that giving an indication of how likely studies were to be effective, rather than reporting continuous effect sizes, could be valuable, as they could indicate interventions that would be inappropriate for the local area. To enhance this approach, it was suggested that a systematic review could additionally give ideas for feedback mechanisms and guidance on how to determine why things might go wrong when implementing an intervention.

Overall, it was challenging to incorporate both policy and methods concepts with diverse stakeholders who have differing understandings of systematic review methods and the policy making process. Workshop 3 perhaps suffered by being too ambitious in its coverage.

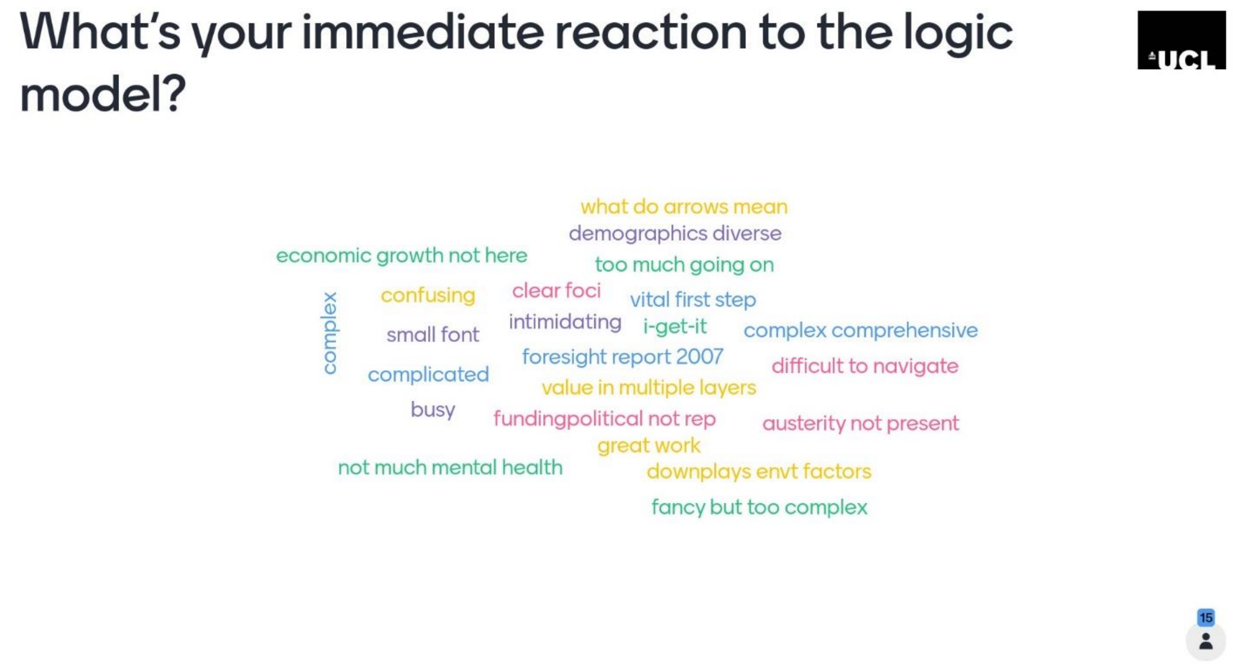
### How are the findings viewed by stakeholders?

The change of focus from child obesity to child health, and evidence that their input was recognized, were highly valued by participants. Many reported feeling that they personally benefitted from participating in the project. Some of the participants indicated that they are very keen to take part in further research, and some that they have now heard other perspectives which may lead to changes in their practice. Generally, participants in both the academic seminar and the policy and practice workshop indicated that the model more accurately reflected the complexity of childhood health than simpler models that focus on obesity status (usually operationalised as BMI) as the outcome.

As noted in Co-Production Collective's report, there was a strong feeling from people with lived experience that research and services repeat the same mistakes over and over again, embedding stigma and focus on individual rather than wider context. One participant had come into the project expecting this to be the case, was delighted to find otherwise and thrilled that their contributions were having a meaningful impact. However, they felt the final result did not go far enough in making progress. Overall, the co-production work revealed an important challenge in that systematic reviews work with existing evidence and reviewers cannot change what has been done in the past (although we can highlight gaps), so the logic model and reanalysis approaches may be constrained by historical terminology or approaches in existing research.

There was some feedback from stakeholders that the logic model was perhaps too complex to be usable by some of its intended audience. In workshop 3, participants were asked for their immediate reaction to the logic model, with the responses entered into Mentimeter to generate a word cloud. The results are visible in Figure 9. In the subsequent discussions, the logic model was seen as conceptually very progressive but needed simplification and the strength of relationships to be added for it to be useful to a broader audience. The participants also noted visual elements that could be refined to make the balance of the different factors clearer (for example, one participant noted that some of the levels had larger boxes than others, which could give a misleading impression about their relative importance).

*Figure 9: Word cloud of reactions to logic model*



What's your immediate reaction to the logic model?

## Overall discussion

### Summary

This was an ambitious 9-month project that set out to develop new approaches for handling complexity in evidence of public health interventions, with co-production as an underlying principle. Co-production inputs changed the focus and our collective understanding from a potentially stigmatising focus on obesity towards a more holistic understanding of childhood health. This focus also allowed for greater consideration of the social determinants of health and the broader macro-level factors that influence children's health. It also moves away from a narrative around personal responsibility towards the socioecological factors that are within the policy remit of decision-makers.

### Strengths of the approaches

**Systems-based logic model** – Our systems-based logic model provides a rare example of an entirely co-produced logic model. We have documented the difference that stakeholder involvement can make to the framing of the issue under consideration. The logic model demonstrates the potential to be gained by bringing together diverse groups of stakeholders, and that diverse groups can work together to conceptualise complex systems of factors that influence health. Substantively, our model offers an advance on other similar exercises, such as the Foresight Obesity Map (Butland, Jebb et al. 2007) as the data are derived from a diverse set of stakeholders who were asked to consider different influencers of obesity, and who transformed the model to reflect the concerns about childhood health as perceived by practitioners, researchers, and those with lived experience. This process could not have been possible without experienced co-producers being an integral part of the team.

**Recalibration –** We found that recalibration of meta-analytic evidence offers a deployable and replicable process of addressing uncertainty within estimates. This represents an improvement on our earlier work (Kneale, Thomas et al. 2019) in terms of how factors are considered (drawing on the logic model), and how the weights are constructed and scaled alongside other weighting components. Although the approach is not intended to replace a conventional estimate, it can provide adjunct estimates where greater importance is attached to contextually similar evidence.

**Binary logistic regression –** This is a novel approach that needs further testing and refinement. It provides a challenge to conventional meta-analytic estimates. Here, the intention is to develop a broad-brushed account of how context may influence whether a study is effective (or not), rather than seek to provide a precise estimate of effectiveness. This approach may be more suitable for complex public health interventions, where different forms of heterogeneity can undermine the meaning and interpretation of a precise effect size. There are also potential extensions of this approach through adopting a multinomial approach to categorizing the outcome and through allocating studies into different groups based on a full understanding of the studies' impacts.

**Enhanced subgroup analysis** – This approach is intended to improve on current practice, and to avoid 'confounding' that is an inherent issue in subgroup analysis, through constructing groups based on multiple characteristics. This approach did not prove fruitful in these analyses but may be informative in larger reviews.

**Qualitative Comparative Analysis for exploring context** – Explorations of study context using a QCA lens developed a way of understanding the degree of correspondence between interventions and their contexts, allowing for assessment of 'contextual fit'. While the approaches above take contextual factors in isolation of the conduct of the intervention itself, these efforts aim to understand how well the intervention – its design and implementation – fit within the context. For example, interventions without features addressing the child-specific context, such as fun and social interaction, or features addressing the needs of children in a low-income context such as those requiring little equipment, or particular ethnic context, such as cultural appropriateness may have poor contextual fit. While the issue of contextual fit could not be explored in-depth using data from either the Langford review or Andermo review, due to data availability, this may be an approach to utilize in the future with a richer dataset..

| **Question** | **Approach** |
|---|---|
| *What are the factors that are perceived by stakeholders to influence health (in my area) and what does this health system look like?* | Identify an array of different factors operating at various levels of society and visualise the pathways to the outcome – *Co-production of a systems-based logic model* |
| *Is there any evidence that the intervention would work differently in an area like mine (Liverpool, Islington, etc.)?* | Grouping studies most like decision-makers' area of interest and giving more weight to these in the meta-analysis to tailor the effect size – *Recalibration approach* |
| *Is this intervention generally a good idea for my area/needs?* | Grouping effective and ineffective studies to compare the contextual factors across the two – *Binary Logistic Regression* |
| *What can you tell me about the impact of multiple contextual factors on the outcome?* | Grouping studies based on multiple contextual factors simultaneously - *Enhanced subgroup analysis* |
| *Which intervention factors are required to achieve success in my context?* | Grouping effective and ineffective studies to see which combinations of contextual factors are associated with each – *Qualitative Comparative Analysis* |

## Assumptions and caveats of approaches developed

The approaches developed in this research could be viewed as a sequence of processes that could be deployed to help to address questions that may arise when considering evidence in context. Deployment of these approaches is contingent on meeting a number of the assumptions listed below, which also represent potential caveats to the approaches.

**Assumption 1:** Consensus can be achieved in framing the health challenge and in the identification of health factors

**Assumption 2:** Available systematic review evidence can be located and contextual evidence can be extracted

**Assumption 3:** The decision-making question about the meta-analytic evidence is clear and aligned with the approaches deployed

**Assumption 4:** Contextual factors influence the effectiveness of interventions

**Assumption 5:** The approaches are useful for decision-makers

**Assumption 6:** Suitable contextual data exists for local areas or districts

**Assumption 7:** The approaches are deployable

## Conclusion and recommendations for ways forward

We have generated a shared understanding of local systems that generate health and ill-health among children and young people and have considered differential effects of interventions that could improve health. We hope that the systems-based logic model will be an important tool for reframing discussions around childhood obesity to include childhood health and wellbeing more broadly, with greater emphasis on a broad range of influential factors.

Given the transformative impacts that co-production has brought to this work, we recommend allowing more time for it in research. Co-production offers rich insights into all stages of the research process, although it requires time to develop relationships between co-producers More time is also required if stakeholders from non-academic backgrounds, including those with lived experience and/or practitioner backgrounds, are to meaningfully engage with research methods. Nevertheless, as this project demonstrates, this investment is rewarded with valuable insights which would otherwise unavailable.

This project has identified different approaches that could be deployed to explore context in complex public health interventions. These approaches help to illuminate how a better understanding of context could change the interpretation of evidence. The detailed descriptions of these approaches (in the main report and appendices) are intended to allow systematic reviewers to replicate our approaches. In this respect, the project has met the short and medium-term impacts that we had anticipated. Through developing these, we hope to contribute further to ensuring the utility of systematic review evidence in public health decision-making and to explore further synergies between meta-analysis and secondary data sources.

# References

Ahmad, N., I. Boutron, A. Dechartres, P. Durieux and P. Ravaud (2010). "Applicability and generalisability of the results of systematic reviews to public health practice and policy: a systematic review." Trials **11**(1): 20.

Andermo, S., M. Hallgren, S. Jonsson, S. Petersen, M. Friberg, A. Romqvist, B. Stubbs and L. S. Elinder (2020). "School-related physical activity interventions and mental health among children: a systematic review and meta-analysis." Sports medicine-open **6**(1): 1-27.

Berlin, J. A. and R. M. Golub (2014). "Meta-analysis as evidence: building a better pyramid." JAMA **312**(6): 603-606.

Brownson, R. C., J. E. Fielding and C. M. Maylahn (2009). "Evidence-based public health: a fundamental concept for public health practice." Annual review of public health **30**: 175-201.

Burchett, H. E. D., L. Blanchard, D. Kneale and J. Thomas (2018). "Assessing the applicability of public health intervention evaluations from one setting to another: a methodological study of the usability and usefulness of assessment tools and frameworks." Health research policy and systems **16**(1): 1-12.

Burke, J. F., J. B. Sussman, D. M. Kent and R. A. Hayward (2015). "Three simple rules to ensure reasonably credible subgroup analyses." Bmj **351**.

Bushman, B. J. and M. C. Wang (2009). Vote-counting procedures in meta-analysis. The handbook of research synthesis. H. Cooper, L. V. Hedges and J. C. Valentine. New York Russell Sage Foundation. **236:** 193-213.

Butland, B., S. Jebb, P. Kopelman, K. McPherson, S. Thomas, J. Mardell and V. Parry (2007). "Foresight. Tackling obesities: future choices. Project report." Foresight. Tackling obesities: future choices. Project report.

Colín-Ramírez, E. (2009). "Actividad física y dieta para la prevención de factores de riesgo cardiovascular (RESCATE)." Rev Esp Nutr Comunitaria **15**(2): 71-80.

DfE. (2021). "Free school meals: autumn 2020." Retrieved Jan 17th, 2022, from https://www.gov.uk/government/publications/free-school-meals-autumn-2020.

Glass, G. V. (2000). "Meta-analysis at 25." Retrieved 26th November, 2014, from http://www.gvglass.info/papers/meta25.html.

Glasziou, P. P., I. Chalmers, S. Green and S. Michie (2014). "Intervention synthesis: a missing link between a systematic review and practical treatment (s)." PLoS Medicine **11**(8): e1001690.

Gough, D., J. Thomas and S. Oliver (2012). "Clarifying differences between review designs and methods." Systematic reviews **1**(1): 1.

Hedges, L. V. (2013). "Recommendations for practice: justifying claims of generalizability." Educational Psychology Review **25**(3): 331-337.

Ioannidis, J. P., S. Greenland, M. A. Hlatky, M. J. Khoury, M. R. Macleod, D. Moher, K. F. Schulz and R. Tibshirani (2014). "Increasing value and reducing waste in research design, conduct, and analysis." The Lancet **383**(9912): 166-175.

Kneale, D., A. Rojas-García, R. Raine and J. Thomas (2017). "The use of evidence in English local public health decision-making." Implementation Science **12**(1): 53.

Kneale, D., J. Thomas, A. O'Mara-Eves and R. D. Wiggins (2019). "How can additional secondary data analysis of observational data enhance the generalisability of meta-analytic evidence for local public health decision-making? ." Research synthesis methods **10**(1): 44-56.

Langford, R., C. P. Bonell, H. E. Jones, T. Pouliou, S. M. Murphy, E. Waters, K. A. Komro, L. F. Gibbs, D. Magnus and R. Campbell (2014). "The WHO Health Promoting School framework for improving the health and well-being of students and their academic achievement." Cochrane Database of Systematic Reviews **4**.

Levy, T. S., C. M. Ruán, C. A. Castellanos, A. S. Coronel, A. J. Aguilar and I. M. G. Humarán (2012). "Effectiveness of a diet and physical activity promotion strategy on the prevention of obesity in Mexican school children." BMC public health **12**(1): 1-13.

Lubans, D., J. Richards, C. Hillman, G. Faulkner, M. Beauchamp, M. Nilsson, P. Kelly, J. Smith, L. Raine and S. Biddle (2016). "Physical activity for cognitive and mental health in youth: a systematic review of mechanisms." Pediatrics **138**(3).

NHS Digital (2020). National Child Measurement Programme, England 2019/20 School Year. N. D. Population Health Team. London, NHS Digital.

Norström, A. V., C. Cvitanovic, M. F. Löf, S. West, C. Wyborn, P. Balvanera, A. T. Bednarek, E. M. Bennett, R. Biggs and A. de Bremond (2020). "Principles for knowledge co-production in sustainability research." Nature Sustainability **3**(3): 182-190.

O'Muircheartaigh, C. and L. V. Hedges (2014). "Generalizing from unrepresentative experiments: a stratified propensity score approach." Journal of the Royal Statistical Society: Series C **63**(2): 195-210.
Oliver, S., L. Langer, P. Nduku, H. Umayam, K. Conroy, C. Maugham, T. Bradley, M. Bangpan, D. Kneale and C. Roche (2021). Engaging Stakeholders with Evidence and Uncertainty: Developing a Toolkit. London, Centre of Excellence for Development Impact and Learning (CEDIL).
Oliver, S., C. Roche, R. Stewart, M. Bangpan, K. Dickson, K. Pells, N. Cartwright, D. Gough and J. Hargreaves (2018). Stakeholder Engagement for Development Impact Evaluation and Evidence Synthesis CEDIL Inception Paper. London, Centre of Excellence for Development Impact and Learning (CEDIL), London International Development Centre
ONS. (2020). "Population denominators by broad ethnic group and for White British, local authorities in England and Wales: 2011 to 2019." Retrieved Jan 17th, 2022, from https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/adhocs/008781populationdenominatorsbybroadethnicgroupandforwhitebritishlocalauthoritiesinenglandandwales2011to2017.
ONS (2021). Mid-Year Population Estimates, UK, June 2020. London, Office for National Statistics.
Paul-Ebhohimhen, V. and A. Avenell (2008). "Systematic review of the use of financial incentives in treatments for obesity and overweight." Obesity Reviews **9**(4): 355-367.
Petticrew, M., P. Tugwell, E. Kristjansson, S. Oliver, E. Ueffing and V. Welch (2011). "Damned if you do, damned if you don't: subgroup analysis and equity." Journal of epidemiology and community health **66**: 95-98.
PHE (2019). Whole systems approach to obesity: A guide to support local approaches to promoting a healthy weight. London, Public Health England.
PHE (2021). Public Health Outcomes Framework.
StataCorp (2021). STATA 17 Base Reference Manual. College Station, TX, Stata Press.
StataCorp (2021). Stata Statistical Software: Release 17. College Station, TX, StataCorp LLC.
The Health Foundation. (2021). "Map of child poverty." Retrieved Jan 17th, 2022, from https://www.health.org.uk/evidence-hub/money-and-resources/poverty/map-of-child-poverty.
van Jaarsveld, C. H. and M. C. Gulliford (2015). "Childhood obesity trends from primary care electronic health records in England between 1994 and 2013: population-based cohort study." Archives of disease in childhood: archdischild-2014-307151.
Wagemann, C. and C. Q. Schneider (2007). Stanards of Good Practice in Qualitative Comparative Analysis and Fuzzy Sets. http://compasss.org/working-papers-series/, Comparative Methods for Systematic Cross-Case Analysis.
Wang, S., J. R. Moss and J. E. Hiller (2006). "Applicability and transferability of interventions in evidence-based public health." Health promotion international **21**(1): 76-83.

# Appendix 1 – Advisory Group Details

In total, 9 people agreed to join the advisory group including academics (3), public health practitioners (2), teachers (1) and people with lived experience (2).

The advisory group were given email updates and opportunities to comment throughout the project and met (virtually) three times. All meetings were hosted on Zoom.

The first advisory group session (June 2021) involved informing the group of the purpose of the study, providing an opportunity for members to get to know each other, clarifying expectations for the workshops, and determining how they should be organised. The advisory group was also asked to help identify suitable systematic reviews to use as case studies.

The second advisory group meeting (July 2021) occurred after the first two workshops to discuss the draft logic model and the implications for the next work packages.

The third advisory group meeting (November 2021) occurred during work package 2, to keep the group updated on progress and consider whether the new approaches were methodologically and theoretically sound. We also updated on refinements to the logic model.

A fourth meeting had been planned, but in the end, it was deemed that final inputs would be best gained through email contact. This was in part due to the timeframe of the work, which was already tight but further compounded the emergence of a new coronavirus variant placing strain on people's time. It was also driven by the requirements of the final stage: the anticipated purpose of a fourth meeting had been to help the research team to interpret workshop/survey findings, consider how the approaches outlined could be scaled up, and to develop guidelines for their use. These were briefly discussed in the third meeting. However, the number and complexity of new approaches that were explored in the project meant that little methodological detail could be covered in the third meeting, and it was decided that giving members time to review materials at their convenience would enable them to better engage with work.

## Appendix 2 – Selection criteria and details for systematic reviews considered for re-analysis

Drawing on the learning from the development of the systems-based logic model, systematic reviews were sought for re-analysis through ad hoc searching that focused broadly on children's health. Using terms to represent our inclusion criteria and our interest in interventions conducted in schools to improve child health, we first searched the Cochrane Database of Systematic Reviews. We then progressed to search other databases (ProQuest and PubMed) to see if we could identify reviews that matched better our inclusion criteria; ad hoc searching was also conducted through Google Scholar search to see if any other reviews matched our criteria more closely. We appraised candidate reviews based on the following criteria:

- **Review design**: Only systematic reviews were eligible. We preferred a Cochrane review as these are more likely to provide access to the underlying data. The publication date was also a consideration and only reviews published within the previous decade were considered as candidate reviews
- **Intervention**: Target reviews must have considered the effectiveness of school-based interventions that aimed to improve children's wellbeing thorough improvements in physical activity, mental health, and/or healthy eating.
- **Contextual characteristics extracted**: We prioritised reviews reporting data on the socio-demographic characteristics of the participants and settings within included trials. These characteristics could have reflected factors such as parental

occupation/educational attainment, free-school meal status, housing status, access to green space, school setting (e.g. urban or rural); children's ethnicity, gender, and age; and health at baseline including healthy weight.

- **Outcomes of interest:** Included reviews need to measure BMI as well as at least one outcome of the following:  physical activity, food intake, mental health, behaviour, or academic progress.
- **Quantitative synthesis:** Reviews were only considered if they included a meta-analysis of 10 of more trials.
- **Heterogeneity:** Meta-analyses within included reviews needed to also exhibit a considerable degree of heterogeneity in their findings. Implicitly, underlying the approaches developed in the research is the assumption that the intervention works more effectively in some settings than others, and that the contextual features of these settings can provide an indication on which settings may be more and less conducive to a successful intervention.

## Table 1: Systematic reviews of school-based interventions considered for re-analysis

## *(Green – full match; Red – no match; Yellow – Partial match)*

| | **RCT only and number of trials** | **Takes Socio-economic status into account** | **Takes other socio-demo into account** | **Intervention** | **BMI Meta-Analysis with 10 or more data sets (ds) (either study or subsets of study data)** | **High heterogeneity** | **Investigated Heterogeneity** | **Meta-Analysis with 10 or more data sets (ds) either study or subsets of study data** | **High heterogeneity** | **Investigated Heterogeneity** | **Cochrane** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Outcome BMI and other | | | | | | | | | | | |
| Langford 2014 | 🟩 **67** | 🟩 Yes where trials reported it | 🟩 Age, gender | WHO Health Promoting School framework | 🟩 Largest 13 ds | 🟩 Heterogeneity substantial | 🟩 Sensitivity analysis | 🟩 Various – physical activity, fat intake, fruit and veg intake<br>10-13 Ds | 🟩 Heterogeneity substantial | 🟩 Sensitivity analysis | 🟩 |
| Pineda 2020 | 🟨 **100** studies Mixed design | 🟩 Yes see metaregression | 🟩 Gender | School food environment | 🟩 12 ds | 🟨 ?reported | 🟩 Meta regression | 🟩 3-Food intake related 16 ds | 🟨 ?reported | 🟩 Meta regression | 🟥 |
| Micha 2018 | 🟩 **91** studies Mixed design | 🟨 Largely not reported in trials | 🟩 Gender, age | School food policies | 🟨 6 ds | 🟨 | States but not reported | 🟩 Food intake related – 13 ds | 🟨 ?reported | 🟩 Meta regression | 🟥 |
| Norris 2020 | 🟩 **42** | 🟥 | 🟨 'Few trials reported gender specific outcome' | Physical activity | 🟨 2 ds | | | 🟩 Various education, cognitive, physical Largest 16 | 🟩 | 🟩 Moderator analysis | 🟥 |
| Outcome physical activity | | | | | | | | | | | |
| Love 2018 | 🟩 17 | 🟩 Yes | 🟩 Gender | Physical activity | 🟥 | 🟥 | | 🟩 17 ds on physical activity | 🟨 not reported level authors assumed heterogeneity | 🟩 Sensitivity analysis | 🟥 |
| Evans 2012 | 🟨 27 studies RCT and other including **'unclear'** | 🟥 | 🟩 Age | Programs to improve daily fruit and veg intake | 🟥 | 🟥 | | 🟩 11 ds | | | 🟥 |
| Pfedderer 2021 | 🟨 Mixed design | 🟥 | 🟩 Rural/urban | Physical activity | 🟥 | 🟥 | | 🟩 40 ds for physical activity | 🟩 High | 🟩 Subgroup analysis – RCT or not, and other items | 🟥 |
| BMI | | | | | | | | | | | |
| Podnar 2020 | 🟨 **200** studies Mixed design | 🟩 Explores subgroup analysis 26 studies on **economically** | | Activity, fitness, sedentary behaviour | 🟩 Y various with **30** plus outcomes | 🟩 Y - high | 🟩 Sensitivity/subgroup analysis – gender, design, quality, age | 🟥 | 🟥 | 🟥 | 🟥 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 🟨 | **deprived pops BUT** MA forest plot not reported | | | 🟩 | 🟩 | 🟩 | 🟥 | 🟥 | 🟥 | 🟥 |
| Brown 2019* | 153 RCTS | By country income | | Most physical activity and diet | Y –various school or wider community analysis with subgrouping by setting outcomes at most 20 | Y – in some high in others 0<br>Heterogeneity beyond subgroup | 🟩 | 🟥 | 🟥 | 🟥 | 🟩 |
| Vasque 2014* | 52 studies | 🟥 | Age, gender | Physical activity | 50 ds | high | Moderator analysis | 🟥 | 🟥 | 🟥 | 🟥 |
| Jones 2020 | 29 mixed design | 🟥 | 🟥 | Physical activity | 🟥 | 🟥 | 🟥 | Physical activity 10/11 ds | High | 🟩 | 🟥 |
| Mental health | | | | | | | | | | | |
| Andermo 2020 | 30 mixed design | SES | Age, gender | Physical activity | 🟥 | 🟥 | 🟥 | Mental health approx. 20 ds | High | Moderator analysis | 🟥 |
| Academic achievement | | | | | | | | | | | |
| Watson 2017 | 16 mixed design | 🟥 | 🟨 | Physical activity | 🟥 | 🟥 | 🟥 | Academic achievement around 10 | High | Insufficient data to do | 🟥 |

Traffic light system used to aid decision. * Not all school-based interventions

## Appendix 3 – Local Authority (district) characteristics used as the basis of selection

Table 1: Socio-demographic factors per local authority

| | Camden | Islington | Liverpool | Test Valley | England |
|---|---|---|---|---|---|
| Deprivation[1] | 20.1 | 27.5 | 42.4 | 11.9 | x |
| Ethnic group – % white[2] | 66.3% | 68.2% | 88.9% | 95% | 85.4% |
| Households with a least one early-yrs or school age child [3] | 18.4% | 16.2% | 20.0% | 24.9% | 23% |
| % of the economically active population aged 16+ who are unemployed [4] | 4.1% | 4.5% | 4.4% | 2.5% | 4.3% (Harlow 4.8, Barking and Dagenham 6.5) |
| % of people aged 16-64 who are employed[4] | 53.8 | 65.4 | 62.7 | 71.5 | 65.1 |
| Density of fast food outlets: rate per 100,000 population[5] | 174.2 | 106.9 | 129.2 | 48.9 | x |
| Provision of green space per resident (sqm)<br>Using Green Space Index per ward[6] | 2.716963<br>(St pancreas Somers Town | 4.697460<br>(St Mary's) | 33.510300<br>(Everton) | 32.731757<br>(Anna) | 29.84 |
| People per sq.km[7] | 4,453 | 16,321 | 4,453 | 201 | 432 |
| % of pupils eligible for free school meals[8] | 30.7 | 24.2 | 29.6 | 14.6* | 19.7 |

[1]Source gov.uk Index of Multiple Deprivation 2019 – 7 weighted domains income, employment, education, health, crime, barriers to housing and services, living environment, [2] 2019-2020 ons.gov.uk, [3] https://www.gov.uk/government/collections/statistics-childcare-and-early-years [4] 2019-2020 ons.gov.uk, [5] Food Standards Agency, Food Hygiene Rating Scheme 2017. Definition of fast foods used: energy intense and available quickly. For example but not limited to burger bars, kebab and chicken shops, chip shops and pizza outlets. [6] Green Space Index 2021 – based on ordinance survey - https://www.fieldsintrust.org/green-space-index/technical-notes 1. a score of 1 indicates a minimum standard of provision; 2. The total provision of parks and green spaces;3. The provision per person; 4. The number of people who are not within a ten-minute walk of a park or green space. [7] 2019-2020 ons.gov.uk.[8] Source gov.uk Free school Meals. Autumn term 2020/21

*Hampshire total

# Appendix 4 – Additional technical details for Approach 1 (Recalibration)

**How is the approach deployed?**
Details of each of these steps is outlined in the Appendix and these steps include:

- ***Identify relevant characteristics in studies and LAs:***

We started through using the logic model as an anchor for understanding the types of contextual characteristics that may be important in determining the effectiveness of an intervention. The studies included within the Langford, Bonell et al. (2014) review were then examined and data on the following characteristics were extracted, either from the review or directly from the primary studies; these reflected a mixture of sociodemographic and socio-political characteristics:

3) *The gender split of participants* - this reflected the logic model co-creators' observations that the relationship between developmental stage and weight was moderated by gender, and that traditional approaches to weight management stigmatised diverse groups including girls
4) *The ethnicity of the children* - this reflected the logic model co-creators' observations that traditional approaches to weight management stigmatised children from minoritised ethnic groups; in addition, a number of the 'cultural' factors described may have reflected differences by ethnic group (for example differences in food culture)
5) *Socioeconomic status of the school/participants* - economic factors were perceived as shaping children's health through numerous ways in the logic model
6) *Whether the study tool place within a low- and middle-income country (LMIC) or high-income country (HIC)* - this may serve as a proxy for some of the macroeconomic factors identified in the logic model
7) *The health care system of the country in which the healthcare system tool place* (whether there was universal healthcare available free at the point of use) - this reflects some of the broader health policy features in the model (for example the importance of joined up policy-making in shaping children's health)
8) *The baseline level of obesity of children* – as represented in the logic model, being labelled as obese entailed being stigmatised by wider society; given the relationship
9) *The baseline fat intake of children* – this was the primary outcome under observation (no equivalent was available for LAs, although an indicator of the availability of unhealthy foods was used)

- ***Code the studies:***

As there was heterogeneity in the way in which the characteristics above were measured across the studies (see data table below), we adopted a strategy analogous to the data coding strategies used in Qualitative Comparative Analysis (QCA) and created rules for re-categorising the data (this is also similar to creating ordinal variables in standard secondary data analysis). The description of the rules for coding data is provided within the data table.

- ***Source and code Local Authority data:***

Similar indicators as those described above were obtained for the selected LAs on the proportion of children from ethnic minority groups (ONS 2020), the gender split among children (ONS 2021), the proportion of children eligible for free school meals (DfE 2021), the quintile of child poverty in which the LA ranked (The Health Foundation 2021), the proportion of children in year 6 who were overweight or obese (PHE 2021); and the number of fast-food outlets per 100,000 population (ref).

- ***Create similarity matrix based on multiple factors:***

After harmonisation of the measures from the studies and LAs, a similarity matrix was constructed based on the data table, with matrix values reflecting the similarity of studies from one another and from each of the LAs. This matrix was constructed using STATA (StataCorp 2021), initially using a Dice binary coefficient, which places high weight on correspondence between values, and implementing the 'proportions' option (to ensure that non-zero values were not all treated as equal (StataCorp 2021). Due to issues with the matrix values, a later iteration treated the data as continuous and created the matrix based on the Canberra method of estimating distances (StataCorp 2021); this method was selected as it is known to be sensitive to small changes near zero which reflected the nature of the data. From the matrices, variables were then constructed reflecting the similarity of each study to a given LA.

- ***Create weight that includes the inverse of the variance:***

Next, we constructed two sets of weights for the meta-analysis – one analogous to a weighting factor in a fixed effect model and one similar to a random effects weight – but both including a parameter reflecting the similarity of the study to a LA. One of the decisions to be made at this point was around how much the 'contextual salience' parameter should contribute to the weighting of the pooled effect size. To ensure that the parameter did not overly influence the contribution of each study, we scaled this parameter according to the maximum and minimum standard error values in the data. A weight for fixed effect models was then created through taking the study variance and adding the scaled similarity parameter, and taking the inverse of this; the weight for the random effects model parameter additionally incorporated tau2.

# Appendix 4 (continued) – Data table and coding decisions made for recalibration and enhanced subgroup analysis approachs

| Trial | Gender (% male) | Gender coded | Mean age | Minoritised ethnic group (%) | Ethnicity coded | Low SEP or FSM coded | Low SEP (parent charas | School percent on free school meal | Country economic status | Healthcare set | Region | Fat intake baseline coded | Obesity baseline |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Caballero 2003** | 51.70% | 0.33 | 7.6 | 100% | 1.00 | 0.66 | (qual evidence) | Qualitative statement | 1 | 0 | Southern USA | 0.49 | 1: 47.0% of children overweight or obese in intervention group |
| **Colín-Ramírez 2010** | 56% | 0.66 | 9.4 | Missing | 0.33 | 0.66 | (all schools in low SES) | Missing | 0 | 0.66 | Mexico | 0.49 (coded as missing because of incompatibility with other outcomes) | 0.49 |
| **Foster 2010** | 47.40% | 0.33 | 11.3 | 82.90% | 1.00 | 0.66 | 27.10% | At least 50% | 1 | 0 | USA | 0.49 | 1: 50.3% of children overweight or obese in intervention group |
| **Haerens 2006** | 63.40% | 1.00 | 13.1 | Missing | 0.00 | 0.66 | 67.40% | Missing | 1 | 0.66 | West Flanders, Belgium | 1= (boys aged 11-14 recommended to have 86g; girls 72g; average 79g; the baseline value in this study is 47% higher) | 0.66: BMI z-score 0.14 suggesting that intervention group differed little from Flemish standard (56th percentile on average – author calculation) |
| **Levy 2012** | 48.40% | 0.33 | 11.0 | Missing | 0.33 | 0.33 | 34.90% | Unclear | 0 | 0.66 | Mexico | 0= (boys aged 7-10 recommended to have 76g; girls 68g; average 72g; the baseline value in this study is half this) | 0.66: Approx 30-32% of children overweight or obese |
| **Luepker 1998** | 51.80% | 0.33 | 8.8 | 31% | 0.33 | 0.66 | Missing | Qualitative statement | 1 | 0 | USA | 0.33 (25-35% of calories should be from fats; the baseline value was 39% so slightly higher) | 0.33: Average BMI was 17.49 and average age was 8.76 (this approximately between 15-50th percentile for girls according to WHO reference charts and between 50-85th percentile for boys) |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Sahota 2001** | 51% | 0.33 | 8.4 | Missing for individuals but potentially higher given evidence | 0.33 | 0.00 | (qual evidence) | Qualitative statement | 1 | 1 | Leeds | 0.49 (coded as missing because of incompatibility with other outcomes) | 0.66: BMI z-score 0.12 suggesting that intervention group differed little from 1990 standard (54th percentile on average – author calculation) |
| **Sallis 2003** | 51% | 0.33 | 12.5 | 44.50% | 0.33 | 0.33 | Missing | 39.50% | 1 | 0 | San Diego, CA, USA | 0.66 (8 fatty items a day) | 0.66: Average BMI was 20.1 and average age was 12.5 (this approximately between 50-85th percentile for girls according to WHO reference charts and between 50-85th percentile for boys) |
| **Trevino 2004** | 50% | 0.33 | 9.79 | 80% | 1.00 | 1.00 | 94% | Qualitative statement | 1 | 0 | San Antonio TX | 1 (25-35% of calories from fats but these should be poly and monounsaturated fats not saturated; children on the trial had 38% of energy from saturated fats) | 1: Average BMI was 20.6 and average age was 9.8 (this approximately between 85-97th percentile for girls according to WHO reference charts and between 85-97th percentile for boys) |
| **Williamson 2012** | 42.80 % | 0.00 | 10.5 | 63% | 0.66 | 1.00 | 77% | 77% | 1 | 0 | Louisiana LA | 0.49 | 0.66: Mean baseline BMI percentile score was 70.4 |
| **Liverpool** | 51.3% | 0.33 | - | 18.7% | 0 | 1 | top quintile of child poverty (1) | 29.5% | 1 | 1 | 0.66 | *0.66* | 0.66: 25.7% of children overweight or obese |
| **Test valley** | 50.6% | 0.33 | - | 7.1% | 0 | 0 | lowest quintile (5) | 14.6%* | 1 | 1 | 0.33 | *0* | 0.33: 17.8% of children overweight or obese |
| **Camden** | 51.5% | 0.33 | - | 48.2% | 0.66 | 0.33 | quintile 3 | 34.4% | 1 | 1 | 0.66 | *1* | 0.33: 21.9% of children overweight or obese |
| **Islington** | 51.3% | 0.33 | - | 47.5% | 0.66 | 0.66 | quintile 2 | 36.4% | 1 | 1 | 0.66 | *0.66* | 0.66: 25.0% of children overweight or obese |
| | | | | | | | | | | | | | |

**Coding rules:**

**Gender -** imbalance towards males; 1=20%+ more males than females; 0.66= 10-19.9% more males than females; 0.33= difference between males and females <10%; 10-19.9% more females than makes

**Ethnicity -** representation of minoritised ethnic groups: 1 = over 75% from minoritised groups; 0.66 = 45-75% of participants from minoritised group; 0.33 = 20-45% from minoritised ethnic group or trial takes place in areas described as diverse; 0 = less than 20% from minoritised group or information missing in HIC trial [Note for LMIC studies missing value 0.33]

**Low Socioeconomic Position or Free School Meals**: 1 = 75% or over claiming FSM or described as disadvantaged; 0.66 qualitative evidence provided suggesting high levels of disadvantage or levels 50-75% for disadvantage or FSM; 0.33 = Missing (no element of disadvantage described); 0 = described as disproportionately advantaged relative local population

**Country economic status**: 1= High income country; 0=Low and middle income country

**Healthcare set ideas:** 1=Universal gov funded; 0.66 = universal insurance (public/private); 0 = non universal system

**Fat intake**: 1=Clear evidence children exceeding guidelines; 0.66 suggestive evidence that children exceeding guidelines; 0.49 missing data or incompatibility in indicators; 0= clear evidence that children are consuming less fat than expected (note the data are unavailable for LAs, and data on obesity were substituted)

Data for LAs is based on level of fast food outlets per 100,000 population (see table x), areas with 0-49.9 were coded as 0; 50-99.9 were coded 0.33; areas with 100-149.9 were coded as 0.66; and those with over 150 coded as 1; these measures are intended to reflect the obesogenic environment.

**<u>Obesity</u>**: 1 = Over 40% categorised as obese or overweight and/or total mental BMI percentile estimated to be over 85th percentile; 0.66 = Over 20-40% categorised as obese or overweight and/or total mental BMI percentile estimated to be between 50th-85th percentile; 0.49=missing; 0.33=10-20% categorised as obese or overweight and/or total mental BMI percentile estimated to be between 15-50th percentile
*Estimate for Hampshire

# Appendix 5 – Additional technical details for Approach 2 (Binary Logistic Regression)

- ***Using the logic model as an anchor, extract relevant characteristics***

A similar process was undertaken where the logic model was consulted as the basis for extracting data from studies. As was the case for Approach 1, there were limitations in the data available within studies. In this example, we were exploring the impact of physical activity interventions on positive mental health, and we were able to examine the gender split of participants, whether the children were from a minoritised ethnic group, the socioeconomic status of the school/participants, and the health care system of the country in which the healthcare system tool place. The justification for selecting these factors is described in approach 1. In addition, we also explored the age of the children.

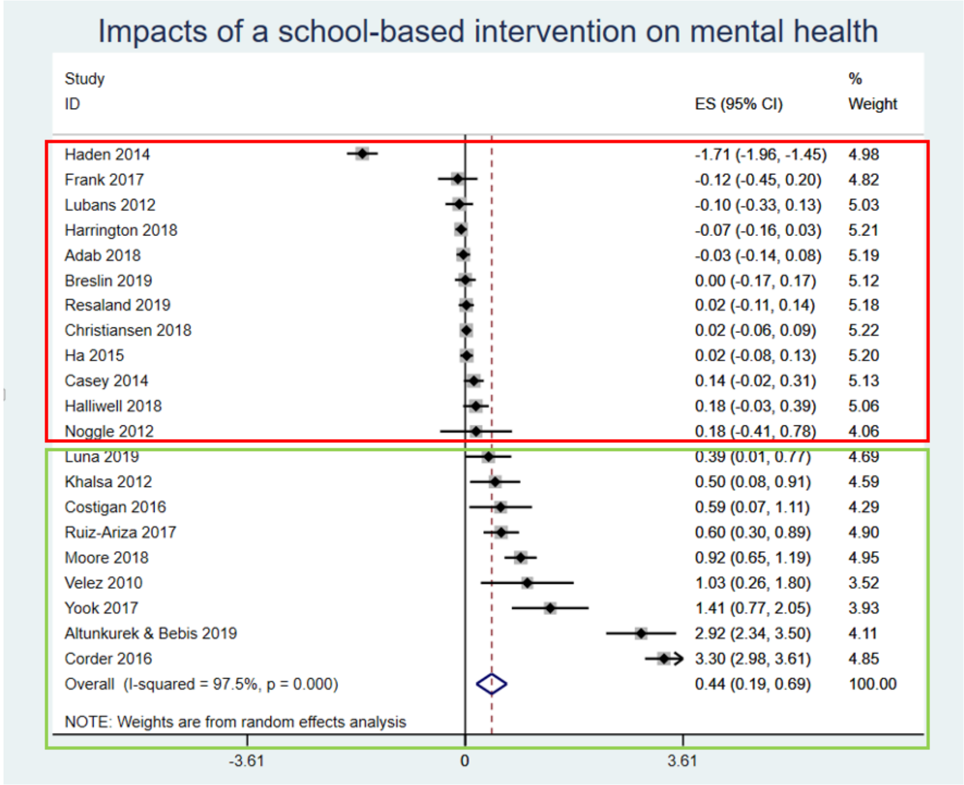
- ***Harmonise the data (create rules for categorising data)***

Coding rules were applied to harmonise the data in a similar process as described in Approach 1. Initially, the values were coded 0 to 1 with potential values in between, although these were later transformed to binary variables (values of only 0 or 1, see data table in Appendix 4).

- ***Classify the studies as in/effective based on the outcome***

In the Andermo review, we conducted a random effects meta-analysis of the 21 RCT studies. This highlighted that there was a group of 9 studies where the impact was clearly positive with a 95% confidence interval that did not cross the line of no effect (green box in Figure 2), as well as 12 studies where the impact was uncertain or harmful (red box in Figure 2). A new binary variable was created and studies allocated 0 or 1 based on the categories in Figure 2.

Figure 2 Random effects model of impact of interventions on positive mental health



- ***Treat each study as aggregate data and create weights that reflects the sample size***

From this point on, each study was treated as aggregate observation to be weighted by its sample size (as might be the case for other types of aggregate data).

(see main body of report for further details)

# Appendix 6 – Additional technical details for Approach 3 (Enhanced Sub-group Analysis)

## How is the approach deployed?

1. ***Using the logic model as an anchor, extract relevant characteristics***

This process is identical to the process described in Approach 1 above. In addition to the characteristics described for Approach 1, we also considered the age of participants in this analysis.

2. ***Code the studies:***

As there is heterogeneity in the way in which the characteristics were measured across the studies, coding rules are once again applied to harmonise the data (see Approach 1).

3. ***Create similarity matrix based on multiple factors:***

As was the case above, a dissimilarity matrix forms the basis of this analysis, and this time it is used to identify groups of studies. As was the case in approach 1 above, a dissimilarity matrix was constructed using Canberra distance metric.

4. ***Apply clustering algorithm to identify groupings in the data***

In this stage, hierarchical cluster analysis is conducted on the dissimilarity matrix, with a number of potential options available on the number of clusters in the data. We used the Duda-Hart values to assess the number of clusters that should be retained. In this case, the values suggested that a four-cluster solution would be optimal.

5. ***Explore the features of the clusters***

Using the data from the Langford review, four clusters of studies were identified. The first cluster consisted of six studies that were all conducted within high income settings without universal healthcare, with participants with a mean age under 13 years, and without a strong gender imbalance towards male participants; this group of studies also had evidence of relatively high levels of ethnic diversity and low socioeconomic advantage among participants. A second group of two studies were conducted in Low and Middle Income countries (both in Mexico) with similarities in the profile of participants across domains. A third group consisted of a single study (Sahota (ref))  which had been conducted in the UK where the profile of participants was described as slightly more advantaged than expected in the local population and which had been conducted among younger participants; a final group also consisted of a single study (Harerens (ref)) which had been conducted in a high income setting with low levels of ethnic diversity among participants, who were also older on average than participants in other studies and included more males.

6. ***Run the analysis with new subgroups***

(see main body of report for details from this point)

## Appendix 7 – Additional technical details for Approach 4 (Contextual QCA)

### QCA set theory – a brief description

QCA uses set theory; sets are essentially categories, which may be binary, e.g. either the case is either a member or a non-member of a set, or they may be ‘fuzzy’ whereby the case (here, trials) is either a full member of that set, a partial member, or a non-member. Numerical descriptors between 1 and 0 are assigned to indicate set membership. For this analysis we employed fuzzy-set membership and used a value of 1 to indicate full set membership, 0 to indicate non-set membership, 0.66 to indicate that a case is more in than out of the set and 0.33 to indicate that a case is more out than in a set.

### Details of Luban’s theory

The overarching model hypothesised three mechanisms through which physical activity may improve children’s cognitive and mental health; neurobiological mechanisms (e.g. endorphins), psychosocial mechanisms (e.g. social connectedness) and behavioural mechanisms (e.g. improved sleep). Since the studies did not measure neurobiological we focused on the four conditions indicated in hypothesis for psychosocial mechanisms; **relatedness** (opportunity for social interaction), **perceived competence** (mastery in the physical domain) **body image** (improvements in appearance self-perceptions) and **autonomy** (independence) and the one measured behavioural mechanism **(self-regulation)**. We identified that ‘autonomy’ had significant overlap with our condition for ‘choice’ and so we collapsed these two conditions.

Table 1: Condition definitions and scoring rules

| | Conditions definition of mechanisms that were theory-based and data driven, and rules for scoring |
|---|---|
| Social interaction | Opportunity for social interaction/one-to-one support from peers/professionals as a main aspect of the intervention = 1, where there was an opportunity for social interaction, but this opportunity was limited (e.g. only in certain aspects of the intervention) = 0.66, where authors say social interaction was important but not evident in intervention = 0.33 |
| Sense of mastery | Opportunity for mastery in a physical domain. For example, efforts to support perception of mastering a new sport so everyone starts as a novice or rewards for participation rather than skill. Where opportunity for mastery is a main aspect of the intervention = 1, where opportunity is a limited part of the intervention = 0.66, where authors say mastery is important but not evident in intervention = 0.33 |
| Body image | Promoting positive self-perception of body image is a main aspect of the intervention = 1, where there is a focus on body image, but it is a limited part of the intervention = 0.66, where authors say the focus is body awareness but not specific to body image or where authors say it is important, but it is not evident in the intervention = 0.33 |
| Autonomy | Participants (individual or student group) were given a choice of a main activity of the intervention or where there is some control over what/when/how the intervention is conducted = 1, choice/control but limited to certain activities = 0.66, authors say choice/control is important but not evident in intervention = 0.33 |
| Self-regulation/coping | Opportunities to facilitate self-regulation or coping or resilience is/are a main aspect of the intervention = 1, where there is a focus on self- regulation/coping/resilience but in a limited |

| | |
|---|---|
| | way = 0.66, authors say self-regulation/coping/resilience is important but not evident in intervention = 0.33 |
| Fun | Designed to be fun and was a main aspect of the intervention = 1, where there was a focus on fun but for a limited part of the intervention = 0.66, where the authors say fun is important, but it is not evident in the intervention = 0.33 |
| Tailoring | Where the intervention is personalised to the participant = 1, where some of the intervention is tailored to the participant = 0.66, where the authors say tailoring is important, but it is not evident in the intervention = 0.33 |

### Table 2: Extracted data/notes per case to support scoring decisions

| Study | Outcome set | Social interaction | Mastery in physical domain | Body image | Autonomy | Self-regulation | Fun | Tailoring |
|---|---|---|---|---|---|---|---|---|
| Haden 2014 | 0 | 0 = nothing relating to social interaction | *0.33 = yoga is "focused on individual progression rather than competition" - but seems very prescriptive in physical requirements - " " 30-minute asana practice (standing for 15 minutes, seated for 5 minutes, backbends/inversions for 5–10 minutes) with each pose held for a count of five, or occasionally taught as a “vinyasa flow” linking all the poses together for one breath"* | *0 = authors do not consider (as mechanism/outcome)* | 0 = home practice which involved choice in activity and when was not prescribed but was encouraged but unclear if it was taken up | 0 = Authors do not consider (as mechanism/outcome) | 0 = not reported | 0 = not reported |
| Frank 2016 | 0 | 0.33 - limited potential for group discussion - Prior to beginning each lesson, behavioural expectations are reviewed, and the agenda for the day’s lesson is reviewed. Then, instructors attempt to activate student background | 0 = not mentioned | 0.33 “Lessons are divided into four units focusing on stress management, body and emotional awareness, self-regulation, and building healthy relationships”. = (score not higher as it is not specifically about improvements or image), | 0 = not reported | 1 = coping was a key outcome “We also anticipated that the treatment group would report significantly higher levels of active primary and secondary coping strategies and lower levels of somatic symptoms as compared to controls” | 0 = not reported | 0.33 = Some but limited “Prior to beginning each lesson, behavioural expectations are reviewed, and the agenda for the day’s lesson is reviewed. Then, instructors attempt to activate student background related to the topic in question and may engage in brief conversation with the group to stimulate interest.” |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | related to the topic in question and may engage in brief conversation with the group to stimulate interest. | | | | | | |
| Lubans 2012 | 0 | 0.33 = mentions importance of environmental (teacher, family, and peer support' - but nothing in intervention that seems to support interactions | 0 = not reported | 0 = does not consider body image | 0 = Intended to have choice - see protocol “The sessions are organized into 4-week units and the sequencing of activities is selected by the students. For example, girls may choose Zumba for 4 weeks followed by 4 weeks of Pilates”'<br>BUT no indication that choice was actually offered | 0 = authors do not consider (as mechanism/outcome | 0 = Very negative – “To reinforce the targeted behaviors, the girls were sent text messages weekly during the second and third terms and biweekly during the fourth term of the program’s delivery (eg, “Sitting down for long periods of time is bad for you, but what makes it worse is that people often eat junk while sitting down in front of the TV. Try to avoid eating dinner while watching TV”)” | 0 = not reported |
| Harrington 2018 | 0 | 0 = not social - peer 'leaders' about empowerment rather than | 0 = process evaluation shows that both peer leaders and girls felt judged | 1 = measures impact on body attractiveness | 0.33 - Yes but process evaluation shows much of time | 0 = authors do not consider (as mechanism/outcome | 0 = evidence seems to be equivocal “Some girls and peer leaders felt | 0 = not reported |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | social support 'uses peer leadership and marketing to empower adolescent girls to influence school decisions, develop themselves as role models, and promote PA to peer | | | spent deciding on activities rather than delivering activities | | that their enjoyment of PE, sports and physical activity had increased as a 0consequence of teachers allowing them to have an opinion/ voice on what is offered and how it is offered.<br>“The PE teachers give us a choice of what we want to do in PE rather than telling us what we’ve got to do. It<br>makes us enjoy PE more because it’s doing something that we want to do rather than being… Forced to do it” (Girls Subgroup, School 4”<br>In the exit survey 46% of girls reported liking physical activity a bit or a lot more and 45% reported liking sport and PE a bit or a lot more, which reinforced these feelings. | |
| Adab 2018 | 0 | 0 = nothing relating to social interaction | 0 = not clear | 1 = Measured as an outcome satisfaction with body image | 0 = not reported | 0 = authors do not consider (as mechanism/outcome | 0 = not reported | 0 = not reported |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Breslin 2019 | 0.33 | 0 = nothing relating to social interaction | 0 = not mentioned | 0.66 = intervention Topic 3/12 sessions relate to body image ' Muscles (upper body' 'muscles (lower body)', 'feel good'. | 0 = not reported | 0 = authors do not consider (as mechanism/outcome | Games are reported as fun but not described = 0.33 | 0 = tailoring and flexibility in delivery for teacher but not school child |
| Resaland 2015 | 0.33 | 0 = theory does not cover interaction - In line with these theories the AS | 1 = in line with these theories the ASK intervention emphasizes creating autonomy supporting and mastery oriented teacher-student interaction in order to enhance students' physical activity behaviour by positively influencing their perception of competence, self-efficacy, and intrinsic motivation for physical activity. | 0 = does not consider body image | 0 = not reported | 0 = authors do not consider (as mechanism/outcome | 0.66 " ASK intervention were planned so that activities were varied and enjoyable for the children" "Special attention was given to creating an encouraging and motivating atmosphere during lessons, in order to support positive feelings and attitudes towards physical activity." (from rationale paper) | 0 = teacher directed |
| Christiansen 2017 | 0.33 | 1 = these involved a) students working in teams, b) ensuring a high degree of student co-creation through choices, reflection and feedback, and c) focusing on individual skills development rather than on competition. | 1 = "the physical activity intervention programme was grounded in SDT and designed to target the three innate psychological needs: competence, autonomy and relatedness in order to improve intrinsic motivation for physical activity for all students ... Central features were included across all courses. These involved a) students working in teams, b) ensuring a high degree of student co-creation through | 1 = body image measured | 0. = student participants involved in planning the intervention | 0 = authors do not consider (as mechanism/outcome | 0 = not reported | 0 = students encouraged to be involved in planning' of intervention only but unclear if participants were, and tailoring in implementation was done at school level only |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | choices, reflection and feedback, and c) focusing on individual skills development rather than on competition” | | | | | |
| Ha 2015 | 0.33 | 0 = nothing relating to social interaction | 0 = not reported - and possible downside of student sports leaders, skipping coaches and ambassadors | 0 = does not consider body image | 0 = not reported | 0 = authors do not consider (as mechanism/outcome | 0 = authors say “future interventions may aim to incorporate more diverse types of rope skipping activities in order to make the activity more interesting to students, and hence might facilitate their engagement in rope skipping, or more generally in PA” | 0 = not reported |
| Casey 2014 | 0.33 | 0 = nothing relating to social interaction | 1 = focus placed upon the tactical dimensions of the game, rather than skill performance | 0 = body shape mentioned as a potential barrier amongst other things but researchers focus was broader | 0 = not reported | 0.33 = perceived behavioural control as a mediator which may or may not be a mechanism | 0 = not reported | 0 = not reported |
| Halliwel 2018 | 0.33 | 0 = nothing relating to social interaction | 0 | 1= authors focus on benefits of yoga for body image | 0.33 = instructor told children to find a memory of when they felt “really | 0 = authors do not consider (as mechanism/outcome) | 1 = majority of children said it was fun (greater than 70%) | 0.33 = in one aspect yes that is in they need to find their inner superheros/warriors and go on an adventure |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | good" about themselves and focus on how the memory ca | | | |
| Noggle 2012 | 0.33 | 0 = nothing relating to social interaction | 0 | 0 = does not consider body image | 0.33 = children were given opportunity to accept or not adjustments to activities | 1 = coping, resilience discussed as mechanism and measured | 0 = Focus is on mindfulness rather than fun - seems quite didactic / health focused - Each session had a theme or talking point that was discussed throughout the session by the instructor. Themes included a basic yoga approach and methodology (postures, breathing, relaxation, meditation, and awareness), nonviolence, mind-body interactions and awareness, body systems, stress management, emotional intelligence, self-talk and critical voice, contentment, discipline, decision making, values and principles, commitment, and acceptance.<br>hat yoga did not do anything one way or the other for students. One negative | 0.33 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | comment suggested that yoga caused overstretching in the foot. | |
| Luna 2019 | 0.33 | 1 = group work - Alternative sports are characterised by being motivating, cooperative, socialising and adapted to participants' characteristics. The selection and organisation of teams (five teams per classroom) was developed | 1 = In this pilot programme, the application of an alternative sport that was novel and unknown to students meant that everyone started with the same theoretical and practical sports knowledge, and there were few initial differences in their levels of technical–tactical sports skill. | 0 = does not consider body image | 0.33 = in week 3 activities include 'Selection and assignment of anthems, badges, mascots and t-shirts representing a team.' | 0.33 = talks about strategies for self-management in regards to conflict but does not measure it | 0.66 = seems learning rather than fun focused "The programme was based on the pedagogical sport education model within a quality physical education framework, and approached from the perspective of social and emotional learning. Elements start with education - end with 'festive' - 'The physical-sport programme was completed following the sport education model structure [17]: (1) season: lengthy didactic units; (2) membership: development of a team spirit and cooperation; (3) regular competition: showing technical–tactical abilities; (4) data register: giving evidence of and analysing the process that has been followed; and (5) festivity: a festive atmosphere." | 0.33 = the overall emphasis was on self-inquiry and not purely didactic teaching - The majority of yoga postures were simple and adaptable for all physical fitness levels. Physically demanding techniques were eventually introduced as optional variations of the standard poses toward the end of the program, based on students' progress |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Khalsa 2012 | 0.66 | 0 = nothing relating to social interaction | 0.33 mentions it is an important factor in background section | 0 = does not consider body image | 0 = not reported | 1= coping and reliance are measured and are highlighted as aspects where intervention may have impact | 0.66 = “this secular program includes simple yoga postures, breathing exercises, visualization, and games with an emphasis on fun and relaxation and minimizing risk without unduly complex or physically athletic or demanding techniques”. | 0.33 = an important component of the program |
| Costigan 2016 | 0.66 | 1 = rewards for participation rather than skill | 1 = rewards for participation rather than skill | 1= ‘body attractiveness’ is a mechanism and outcome | 0 = not reported | 0 = authors do not consider (as mechanism/outcome) | 0.66- = " First, sessions were designed to be enjoyable by including a fun warmup and cooldown activity or game, and participants worked with a partner of their choice (one participant undertook the ‘‘work’’ phase of the sessions, while their partner completed the ‘‘rest’’ phase)." | 0 = not reported |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Ruiz-Ariza 2019 | 0.66 | 0 | 0 = not discussed (although  similar intervention to costigan - rewards not mentioned) | 0 = not considered | 0.66 =  “to promote autonomy, participants were also given the opportunity a) to choose music (student playlists used weeks 2–8), b) to select specific exercises to be completed during a workout (weeks 4–6), and c) to choose a workout (between two workouts previously completed; weeks 7 and 8) once exercises were mastered” | 1= self-control of emotions is one of mechanisms discussed and measured | 0.66 = cooperative element described as “For some researchers, the social character of cooperative PA, playful entertainment, and group decision-m aking in cooperative exercises, are some determinant factors of this kind of PA” | 0 = not reported |
| Moore 2018 | 0.66 | 1 =  “A 1:4 or 1:3 ratio of researchers to participants was always maintained for thorough supervision and guidance throughout all resistance training sessions” | 0.33= notes mastery as concept within resilience- resilience is focus of psychoeducation, intervention is non-aggressive | 0 = not considered | 0 = not reported | 0.66 = self-regulation, coping and reliance are discussed as aspects intervention may promote although does not measure it | 0 = not reported | 0.33 = “Psycho-education—based on facilitator guided group discussion. Topics included respect, goal-setting, self-concept and self-esteem, courage, resilience, bullying and peer pressure, self-care and caring for others, values, and, optimism and hope” |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Velez 2010 | 1 | 1= not one-to-one but intensive support from professionals | 1 = “Self-concept is a domain of overall global self-esteem and is defined as a group of perceptions that an individual has of themselves regarding particular characteristics (i.e., physical, cognitive, etc.) (31). Self-concept has been touted to be the driving force behind motivation for behaviour changes, including changes in physical activity (20,31). Physical self-concept in particular holds considerable importance during adolescence because the physical changes during that phase promote a heightened awareness of self-consciousness (15).” | 1 = focus is on body, feeling attractive is an outcome | 0 = not reported | 0 = authors do not consider (mechanism/outcome) | 0.66 = “In keeping consistent with Csikszentmihalyi’s theory (12) that boredom or being overwhelmed are the 2 main factors for noncompliance with exercise programs, the structure of the resistance training program was designed to ensure that the participants would be adequately challenged to maintain motivation. The low level of dropouts suggests that the level of difficulty and the structure of the program were generally liked and well received by the adolescents’” | 1 = “RPE (10) was used to assess the participants’ perception of the intensity of the workout and was administered after each exercise was completed. The RPE scale consists of ratings from 6–20, where 6 is very, very light work and 20 is maximal exertion. Each score was taken to ensure that the subjects were working at a moderate to moderate-high intensity level based on their personal feelings of exertion.” |
| Yook 2017 | 1 | 1= session 6 goals to recognise joy of collaboration | 0.33 = new sport but not explained if / why 'new' is important | 0 = not considered | 0.33 = encouraged to pay attention to their own body | 1 = reliance as a mechanism and outcome | 0.66 = “the physical activity used in this study is a new sport that consists of a game-focused activity and it naturally gives positive emotions. " physical activity programmes for teenagers should involve not only physical fitness, but also psychological | 0.33 = “the physical activity programme combining a new sport and mindfulness yoga was carried out for 8 weeks; both new sport and yoga were separately practiced once per week. Mindful yoga directed participants to pay attention to their bodily senses, feelings, and |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | fitness and development of positive emotions and personality" "The physical activity intervention combining new sport and mindfulness yoga was theoretically based on the healthy psychological growth model (Lopez & Snyder, 2003). This model assumes that all human beings possess psychological strengths and the ability to achieve optimum psychological health, and moves on to claim that happiness depends 50% on genetics, 10% on circumstances, and 40% on intentional activities (Lyubomirsky, 2007)." | thoughts to recognise themselves. The programme also drew emotions from the body-focused programme to recognise the bodily senses and emotions during that time” |
| Altunkurek and Bebis 2019 | 1 | 1= both individual and coaching sessions | 0 = not mentioned | 1 = measures self-perception | 0.33 = group preference of some activities but not individual | 1 = coping is an outcome measured | 0 = not reported | 0.66 = “During these sessions [on health-related matters], each student determined the agenda of the coaching session. They fo- cused on matters directly related to improving health and on matters indirectly related to health, |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | such as family and friend relationships, school achievement, and stress management.” |
| Corder 2016 | 1 | 1 = sociability was measured and mentorship was a key component | 1 = rewards for participation rather than skill | 0 = not considered | 0.33 = choice per mentor/peer leader not individual | 0 = Authors do not consider (mechanism/outcome) | 1 = over 70% Young People said so - | 0.33 = flexibility of when additional activities beyond core occur - as in try activities at home but not certain this was taken up |