International Journal of Behavioral Nutrition and Physical Activity

REVIEW Open Access

Identification and evaluation of risk of generalizability biases in pilot versus efficacy/effectiveness trials: a systematic review and meta-analysis

Michael W. Beets[1*], R. Glenn Weaver[1], John P. A. Ioannidis[2], Marco Geraci[1], Keith Brazendale[1], Lindsay Decker[1], Anthony D. Okely[3], David Lubans[4], Esther van Sluijs[5], Russell Jago[6], Gabrielle Turner-McGrievy[1], James Thrasher[1], Xiaming Li[1] and Andrew J. Milat[7,8]

Abstract

Background: Preliminary evaluations of behavioral interventions, referred to as pilot studies, predate the conduct of many large-scale efficacy/effectiveness trial. The ability of a pilot study to inform an efficacy/effectiveness trial relies on careful considerations in the design, delivery, and interpretation of the pilot results to avoid exaggerated early discoveries that may lead to subsequent failed efficacy/effectiveness trials. "Risk of generalizability biases (RGB)" in pilot studies may reduce the probability of replicating results in a larger efficacy/effectiveness trial. We aimed to generate an operational list of potential RGBs and to evaluate their impact in pairs of published pilot studies and larger, more well-powered trial on the topic of childhood obesity.

Methods: We conducted a systematic literature review to identify published pilot studies that had a published larger-scale trial of the same or similar intervention. Searches were updated and completed through December 31st, 2018. Eligible studies were behavioral interventions involving youth (≤18 yrs) on a topic related to childhood obesity (e.g., prevention/treatment, weight reduction, physical activity, diet, sleep, screen time/sedentary behavior). Extracted information included study characteristics and all outcomes. A list of 9 RGBs were defined and coded: intervention intensity bias, implementation support bias, delivery agent bias, target audience bias, duration bias, setting bias, measurement bias, directional conclusion bias, and outcome bias. Three reviewers independently coded for the presence of RGBs. Multi-level random effects meta-analyses were performed to investigate the association of the biases to study outcomes.

Results: A total of 39 pilot and larger trial pairs were identified. The frequency of the biases varied: delivery agent bias (19/39 pairs), duration bias (15/39), implementation support bias (13/39), outcome bias (6/39), measurement bias (4/39), directional conclusion bias (3/39), target audience bias (3/39), intervention intensity bias (1/39), and setting bias (0/39). In meta-analyses, delivery agent, implementation support, duration, and measurement bias were associated with an attenuation of the effect size of − 0.325 (95CI − 0.556 to − 0.094), − 0.346 (− 0.640 to − 0.052), − 0.342 (− 0.498 to − 0.187), and − 0.360 (− 0.631 to − 0.089), respectively.

Conclusions: Pre-emptive avoidance of RGBs during the initial testing of an intervention may diminish the voltage drop between pilot and larger efficacy/effectiveness trials and enhance the odds of successful translation.

Keywords: Intervention, Childhood obesity, Youth, Physical activity, Sleep, Diet, Screen time, Scalability, Framework

* Correspondence: beets@mailbox.sc.edu
[1] Arnold School of Public Health, University of South Carolina, Columbia, SC, USA
Full list of author information is available at the end of the article

Background

Pilot testing of behavioral interventions (aka feasibility or preliminary studies) is a common part of the process of the development and translation of social science/public health interventions [1–6]. Pilot studies, within the translational pipeline from initial concept to large-scale testing of an intervention, are conducted to *"provide information of high utility to inform decisions about whether further testing [of an intervention] is warranted* [7]." In pilot studies, preliminary evidence on feasibility, acceptability, and potential efficacy of an intervention are collected [1–5]. Across major government funders, such as the National Institutes of Health (NIH), the Medical Research Council and National Institute of Health Research in the United Kingdom, the National Health and Medical Research Council of Australia, and the Canadian Institutes of Health Research, pilot studies play a prominent role in the development and funding of almost all large-scale, efficacy/effectiveness intervention trials. This is evidenced by funding mechanisms specifically for pilot studies (e.g., NIH R34) [7], the requirement of preliminary data presented in grant applications, and the inclusion of pilot studies as a key stage in the development and evaluation of complex interventions [8].

Pilot studies have received heightened attention over the past two decades. This attention has focused on what constitutes a pilot study, the type of information a pilot study can and cannot provide, whether hypothesis testing is or is not appropriate within a pilot study, the various research designs one could employ, and debates about their proper nomenclature [1–6, 9–13]. More recently, peer-reviewed scientific journals have been created with a specific focus on pilot studies, as well as an extension to the CONSORT Statement focusing on various aspects of reporting pilot/feasibility studies [9]. These articles raise important considerations in the conduct and reporting of pilot studies, and decision processes regarding whether or not to proceed with a large-scale, efficacy/effectiveness trial, yet they focus largely on topics related to threats to internal validity that may ensue.

Biases can lead to incorrect conclusions regarding the true effect of an intervention, and can be introduced anywhere along the translational pipeline of behavioral interventions – from the initial development and evaluation during a pilot study, in the large-scale randomized efficacy or effectiveness trial, to the evaluation of an intervention in a dissemination and implementation study [14, 15]. Biases relevant to internal validity, such as whether blinding or randomization were used, rates of attrition, and the selective reporting of outcomes [16] are important considerations when designing an intervention trial or evaluating published studies. However, intervention researchers need to also consider external validity in the design, conduct, and interpretation of pilot studies. The introduction of biases related to external validity can lead to prematurely scaling-up an intervention for evaluation in a larger, efficacy/effectiveness trial.

Internal validity deals with issues related to whether the receipt of the intervention was the cause for change in the outcome(s) of interest in the specific experimental context under which an intervention was tested [17]. In contrast, external validity refers to the variations in the conditions (e.g., target audience, setting) under which the intervention would exhibit the same or similar impact on the outcome(s) of interest [17]. These are important distinctions, as the vast majority of checklists for the design and conduct of a study focus on topics related to internal validity, as noted by the widely endorsed risk of bias checklists [16] and trial reporting statements [18, 19], while largely ignoring whether the casual inference, in this case the inference drawn from a pilot study, are likely to generalize to variations in study conditions that could occur in a larger-scale, more well-powered trial. Thus, if the purpose of conducting pilot studies is to "*inform decisions about whether further testing [of an intervention] is warranted* [7]", it is then reasonable to expect a great deal of emphasis would be placed on aspects of external validity, particularly when determining if a larger-scale trial is necessary.

Rationale of the proposed "risk of generalizability biases"

Biases related to external validity present in a pilot study can result in misleading information about whether further testing of the intervention, in a larger, efficacy/effectiveness trial, is warranted. We define **"risk of generalizability biases"** as the degree to which features of the intervention and sample in the pilot study are NOT scalable or generalizable to the next stage of testing in a larger, efficacy/effectiveness trial. We focus on whether aspects like who delivers an intervention, to whom it is delivered, or the intensity and duration of the intervention during the pilot study are sustained in the larger, efficacy/effectiveness trial. The use of the term "bias" in this study therefore refers to ways in which features of the pilot study lead to systematic underestimation or overestimation of the assessment regarding the viability of the tested intervention and, subsequently, influence the decision whether to progress to the next stage of evaluating the intervention in a larger, more well-powered trial is necessary.

There is a history of studies that have evaluated the same (or very similar) interventions yet produce different outcomes when conducted under efficacy or effectiveness conditions, a phenomenon referred to as "voltage drop" [20–23]. Conducting a study from an efficacy perspective may ignore important aspects of generalizability that are associated with the design and conduct of an effectiveness study [24]. Doing so can introduce external validity biases (either knowingly or unknowingly) that may change the effect the intervention has on outcomes. In Table 1, we present examples from a sample of six interventions [25–30, 32–37] related to childhood obesity that have a published efficacy and a subsequent effectiveness trial and one intervention [31] with only an efficacy evaluation published. In these studies [25–37], the authors indicate the substantially reduced or null effects observed in the effectiveness trial may be due to a feature of the efficacy study, such as delivery of the intervention by study personnel, being removed in the effectiveness trial [38]. These are but a few of the adaptations interventionists could make [39] that may lead to possible biases that distort the estimated impact of an intervention, especially during pilot testing.

Interventions that are pilot tested using highly skilled individuals, or extensive support for implementation, and/or short evaluations of the intervention may fail eventually if these features are not retained in the next phase of evaluation. Given pilot studies are often conducted with smaller sample sizes [40], it may be easier to introduce certain features, such as delivering the intervention by the researchers or providing extensive support for implementation, on a smaller scale than when testing an intervention in a larger trial that includes a larger sample size and more settings within which to provide the intervention. Pilot studies, therefore, may be more susceptible to introducing features that lead to underestimation or overestimation of an intervention's viability for testing in a larger, more well-powered trial.

The definition of risk of generalizability biases, as applied to pilot intervention studies, is grounded in concepts within the scalability, scaling-up, and dissemination/implementation of interventions for widespread uptake and population health impact [39, 41–50] and pragmatic trial design [51–53]. The scalability literature describes key considerations interventionists must consider when taking an intervention that is efficacious "to scale" for population health impact. These include the human, technical and organizational resources, costs, intervention delivery and other contextual factors required to deliver the intervention and how the intervention interacts within the setting in which it is evaluated, such as schools that have close relationships with the research team, that may not be replicable in a larger study. These elements are consistent within implementation frameworks [20–22, 54–58], which describe the need to consider the authenticity of delivery, the representativeness of the sample and settings, and the feasibility of delivering the intervention as key components in translating research findings into practice. More recently, guides for intervention development, such as PRACTIS (PRACTical planning for Implementation and Scale-up) [59], outline an iterative multi-step process and considerations for the creation of interventions to more closely align with the prototypical characteristics of the population, setting, and context where an intervention is ultimately intended to be delivered [60].

Consideration for the elements represented in the scalability and implementation framework literature are paramount for the effective translation of interventions to improve population health. Discussions surrounding their importance, however, predominately focus on the middle to end of the translational pipeline continuum, largely ignoring the relevance of these issues during the early stages of developing and evaluating interventions in pilot studies. Frameworks that focus on pilot testing, such as ORBIT (Obesity-Related Behavioral Intervention Trials) [61], describe the preliminary testing of interventions to be done with "highly selected participants" under "ideal conditions" only to move on to more representative samples if the intervention reaches clinically or statistically significant targets under optimal conditions. This perspective aligns with the efficacy-to-effectiveness paradigm that dominates much of the behavioral intervention field, where interventions are initially studied under highly controlled conditions only to move to more "real-world" testing if shown to be efficacious [21]. These pilot testing recommendations are at odds with the scalability literature and the extensive body of work by Glasgow, Green and others that argues for a focus on evaluating interventions that more closely align with the realities of the conditions under which the intervention is ultimately designed to be delivered [49]. Hence, optimal conditions [24] may introduce external validity biases that could have a substantial impact on the early, pilot results and interpretation of whether an intervention should be tested in a larger trial [20–22, 55, 62].

The identification of generalizability biases may assist researchers to avoid the introduction of such artefacts in the early stages of evaluating an intervention and, in the long run, help to avoid costly and time-consuming decisions about prematurely scaling an intervention for definitive testing. Drawing from the scalability literature and incorporating key concepts of existing reporting guidelines, such as TIDieR [63], CONSORT [9], TREND [64], SPIRIT [65], and PRECIS-2 [51, 52] we describe the development of an initial set of risk of generalizability biases and provide empirical evidence regarding their influence on study level effects in a sample of published pilot studies that are paired for comparison with a published larger-scale efficacy/effectiveness trial of the same or similar intervention on a topic related to childhood obesity. The purpose of this study was to describe the rationale for generating an initial set of "risk of generalizability biases"

Table 1 Examples of Generalizability Biases in the Childhood Obesity Literature

Bias	Likely Larger Effect	Likely Smaller/No Effect
Study	Fitzgibbon 2005 [25]	Kong 2016 [26]
Who delivered the intervention?	"…the use of specially trained early childhood educators rather than classroom teachers to deliver the intervention, thereby raising questions of generalizability."	"…using teachers in existing Head Start classrooms to deliver the intervention."
Study	Cohen 2015 [27]	Sutherland 2017 [28]
How much of the intervention was provided?	1 full day training and 1 half day training	1 90-min training
Study	Beets 2016 [29]	Beets 2018 [30]
How much support to implement the intervention was provided?	"During the first year of receiving the intervention for both the immediate and delayed program, each program received four booster sessions. During the second year of receiving the intervention (for the immediate condition only) 2 booster sessions/program were provided."	No additional onsite booster sessions or follow-up
Study	Sutherland 2016 [31]	
Who delivered the intervention?	"The provision of an in-school physical activity consultant for 1 day per week was the largest cost relating to the efficacy trial (66% of the total intervention cost). Whilst the provision of an in-school physical activity consultant was necessary under efficacy trial conditions in order to evaluate the effect of the combination of intervention strategies, the feasibility of providing a part-time consultant within schools across large geographic regions and the cost of such a model of support presents challenges in upscaling the intervention. The dissemination of an effective intervention across the community requires the use of implementation strategies which better mirror real world practice."	
Study	McKenzie 1996 [32]	Hoelscher 2004 [33] (PE outcomes)
How much support to implement the intervention was provided?	"Following initial training, CATCH PE consultants provided on-site follow-up approximately every 2 weeks. During the 2.5 years, consultants made 3089 documented school visits, averaging 55.3 per school and 51.7 min in length. Consultants performed various roles during visits, including giving feedback to teachers, modeling new lesson segments, team teaching, and providing motivation and technical support."	No onsite, on-going support provided
Study	Salmon 2008 [34]	Salmon 2011 [37]
How much of the intervention was provided?	19 lessons delivered	6 lessons delivered "…Switch-2-Activity involved an abbreviated programme; therefore, the intervention 'dose' was lower…"
How long was the intervention delivered?	10 months	7 weeks
Who delivered the intervention?	"All intervention components were delivered by one intervention specialist (a qualified Physical Education teacher) across all three schools."	"the programme was delivered by regular class teachers rather than by a specialist university research team…"

Table 1 Examples of Generalizability Biases in the Childhood Obesity Literature *(Continued)*

Bias	Likely Larger Effect	Likely Smaller/No Effect
What measures were used to collect information on outcomes?	Objective measures	Self-report
Study	West 2010 [35]	Gerards 2015 [36]
Who delivered the intervention?	"All sessions were facilitated by a clinical psychologist and accredited provider of the intervention (who co-authored the intervention materials), with assistance from graduate students in nutrition and dietetics, physical education, and psychology."	"The intervention was led by three different facilitators. These health professionals have been accredited after attending an official 3-day training course and an additional intervention day." "Finally, the West 2010 [35] study was implemented as an efficacy study, while in the current trial we tried to implement in the real life situation, which may have led to less significant study results."
Who received the intervention?	"participants were mainly white, well-educated parents with moderate levels of employment and income."	

(defined below) that may lead to exaggerated early discoveries [66] and therefore increase the risk of subsequent efficacy and effectiveness trials being unsuccessful. We provide empirical support of the impact of these biases using meta-analysis on outcomes from a number of published pilot studies that led to testing an intervention in a larger efficacy/effectiveness trial on a topic related to childhood obesity and provide recommendations for avoiding these biases during the early stages of testing an intervention.

Methods

For this study, we defined behavioral interventions as interventions that target one or more actions individuals take that, when changed in the appropriate direction, lead to improvements in one or more indicators of health [67, 68]. Behavioral interventions target one or more behaviors in one of two ways – by directly targeting individuals or by targeting individuals, groups, settings or environments which may influence those individuals. Behavioral interventions are distinct from, but may be informed by, basic or mechanistic research studies that are designed to understand the underlying mechanisms that drive behavior change. Mechanistic studies are characterized by high internal validity, conducted in laboratory or clinical settings, and conducted without the intent or expectation to alter behavior outside of the experimental manipulation [69–72]. Thus, behavioral interventions are distinct from laboratory- or clinical-based training studies, pharmacological dose-response or toxicity studies, feeding and dietary supplementation studies, and the testing of new medical devices or surgical procedures.

We defined *"behavioral intervention pilot studies"* as studies designed to test the feasibility of a behavioral intervention and/or provide evidence of a preliminary effect(s) in the hypothesized direction [2, 10, 61]. These studies are conducted separately from and prior to a larger-scale, efficacy/effectiveness trial, with the results used to inform the subsequent testing of the same or refined intervention [61]. Behavioral intervention pilot studies, therefore, represent smaller, abbreviated versions or initial evaluations of behavioral interventions [10]. Such studies may also be referred to as "feasibility," "preliminary," "proof-of-concept," "vanguard," "novel," or "evidentiary" [3, 6, 61].

Study design

A systematic review was conducted for published studies that met our inclusion criteria (see below), with all reviews of database updated and finalized by December 31st, 2018. All procedures and outcomes are reported according to the PRISMA (Preferred Reporting Items for Systematic review and Meta-Analysis) [73] statement.

Data sources and search strategy

A comprehensive literature search was conducted across the following databases: PubMed/Medline; Embase/Elsevier; EBSCOhost, and Web of Science. A combination of MeSH (Medical Subject heading), EMTREE, and free-text terms, and any boolean operators and variants of terms, as appropriate to the databases, were used to identify eligible publications. Each search included one or more of the following terms for the sample's age - child, preschool, school, student, youth, and adolescent - and one of the following terms to be identified as a topic area related to childhood obesity - obesity, overweight, physical activity, diet, nutrition, sedentary, screen, diet, fitness, or sports.

To identify pairs of studies that consisted of a published pilot study with a larger, more well-powered trial of the same or similar intervention, the following procedures were used. To identify pilot studies, the following terms were used: pilot, feasibility, proof of concept, novel, exploratory, vanguard, or evidentiary. These terms were used in conjunction with the terms regarding sample age and topic area. To identify whether a pilot study had a subsequent larger, more well-powered trial published, the following was conducted. First, using a backwards approach, we reviewed published systematic reviews and meta-analyses on interventions targeting a childhood obesity-related topic that were published since 2012. The reviews were identified utilizing similar search terms as described above (excluding the pilot terms), with the inclusion of either "systematic review" or "meta-analysis" in the title/abstract. All referenced intervention studies in the reviews were retrieved and searched to identify if the study cited any preliminary pilot work that informed the intervention described and evaluated within the publication. Where no information about previous pilot work was made or statements were made about previous pilot work, yet no reference(s) were provided, contact via email with the corresponding author was made to identify the pilot publication.

All pilot studies included in the final sample for pairing with a larger, more well-powered trial required that the authors self-identified the study as a pilot by either utilizing one or more the terms commonly used to refer to pilot work somewhere within the publication (e.g., exploratory, feasibility, preliminary, vanguard), or the authors of a larger, more-well powered trial had to specifically reference the study as pilot work within the publication of the larger, more well-powered trial or protocol overview publication.

Inclusion criteria

The following inclusion criteria were used: study included youth ≤18 years, a behavioral intervention (as defined previously) on a topic related to childhood obesity, have a published pilot and efficacy/effectiveness trial of the same or similar intervention, and were published in English. An additional inclusion criterion for the efficacy/effectiveness trials was the trial had to have a comparison group for the intervention evaluated. This criterion was not used for pilot studies, as some pilot studies could use a single group pre/post-test design.

Exclusion criteria

Exclusion criteria were articles, either pilot or efficacy/effectiveness, that only provided numerical data associated with outcomes found to be statistically significant, reported only outcomes associated with compliance to an intervention, or the published pilot study only described the development of the intervention and did not present outcomes associated with preliminary testing/evaluation the intervention on one or more outcomes.

Data management procedures

For each search within each database, all identified articles were electronically downloaded as an XML or RIS file and uploaded to Covidence (Covidence.org, Melbourne, Australia) for review. Within Covidence, duplicate references were identified as part of the uploading procedure. Once uploaded, two reviewers were assigned to review the unique references and identify those that met the eligibility criteria based on title/abstract. Where disagreements occurred, a third member of the research team was asked to review the disputed reference to make a final decision. Full-text PDFs were retrieved for references that passed the title/abstract screening. These articles were reviewed and passed on to the final sample of studies for the extraction of relevant study characteristics and outcomes. For included studies, all reported outcomes (e.g., means, standard deviations, standard errors, differences, change scores, 95% confidence intervals) were extracted for each study for analyses (described below).

Defining and identification of risk of generalizability biases

Prior to reviewing the full-text articles that met the inclusion criteria, a candidate list of risk of generalizability biases was developed by the study authors, operationally defined, and their hypothesized influence on study outcomes determined based on the scalability, scaling-up, and dissemination/implementation of interventions for widespread uptake and population health impact [41–50] and pragmatic trial design [51–53] literature. After the initial set of risk of generalizability biases were developed and operationally defined, three reviewers (MB, KB, LD) independently reviewed the full-texts of the pilot and efficacy/effectiveness trial pairs for the potential presence of the biases. Each risk of generalizability bias was classified as either "present" or "absent". Where discrepancies were identified, discussion regarding the evidence for bias was conducted to resolve the disagreement. In addition, during the review of the pilot and efficacy/effectiveness pairs, additional biases were identified, discussed, defined, and added to the list of risk of generalizability biases, where necessary. A total of 9 risk of generalizability biases were identified and operationally defined. Each bias, along with the definition, the hypothesized influence, and examples, are presented in Table 2.

Meta-analytical procedures

Standardized mean difference (SMD) effect sizes were calculated for each study across all reported outcomes. The steps outlined by Morris and DeShon [85] were used to create effect size estimates from studies using different designs across different interventions (independent groups pre-test/post-test; repeated measures single group pre-test/post-test) into a common metric. For each study, individual effect sizes and corresponding 95% CIs were calculated for all outcome measures reported in the studies.

To ensure comparisons between pilot and efficacy/effectiveness pairs were based upon similar outcomes, we classified the outcomes reported across pairs (i.e., pilot and efficacy/effectiveness trial) into seven construct categories that represented all the data reported [86]. These were measures of body composition (e.g. BMI, percent body fat, skinfolds), physical activity (e.g., moderate-to-vigorous physical activity, steps), sedentary behaviors (e.g., TV viewing, inactive videogame playing), psychosocial (e.g., self-efficacy, social support), diet (e.g., kcals, fruit/vegetable intake), fitness/motor skills (e.g., running, hopping), or other. For studies reporting more than one outcome within a category, for instance reporting five dietary outcomes in the pilot and reporting two dietary outcomes in the efficacy/effectiveness trial, these outcomes were aggregated at the construct level to represent a single effect size per construct per study using a summary calculated effect size and variance computed within Comprehensive Meta-Analysis (v.3.0). The construct-level was matched with the same construct represented within the pairs. For all comparisons, outcomes were used only if they were represented in both studies within the same construct as defined above. For instance, a study could have reported data related to body composition, diet, physical activity in both the pilot and efficacy/effectiveness trial, but also reported sedentary outcomes for the pilot only and psychosocial and fitness related outcomes for the efficacy/effectiveness only. In this scenario, only the body composition, diet, and physical activity variables would be compared across the two studies within the pair. Attempts were made at one-to-one identical matches of outcomes and reported

Table 2 Operational Definitions of Risk of Generalizability Biases

Risk of Generalizability Bias	Questions to Ask	Increased Presence with Small Sample	Hypothesized Influence of the Presence of Risk of Generalizability Bias		Example	
			Pilot	Larger-Scale Efficacy/ Effectiveness	Pilot	Larger-Scale Efficacy/ Effectiveness
	What is the potential for difference(s) between…					
Intervention Intensity Bias	…the number and length of contacts in the current study and future evaluations of the intervention?	Yes	More frequent and longer contacts result in more effective intervention	Fewer and shorter contacts results in less effective intervention compared to pilot	19 lessons delivered (Salmon 2008 [34])[a]	6 lessons delivered (Salmon 2011 [37])[a]
Implementation Support Bias	…the amount of support provided to implement the intervention in the current study and future evaluations of the intervention?	Yes	Greater amounts of support to implement the intervention results in more effective intervention	Reduced support to implement the intervention results in less effective intervention compared to pilot	*"During the intervention, weekly, audio-taped debriefing meetings were held with the interventionists and project investigators to troubleshoot any problems with each session and to plan for the following sessions."* (Beech 2003 [74])	
Intervention Delivery Agent Bias	…the level of expertise of the individual(s) who deliver the intervention in the current study compared to who will deliver the intervention in future evaluations?	Yes	Higher levels of expertise delivering the intervention results in more effective intervention	Lower level of expertise to deliver the intervention results in less effective intervention compared to pilot	*"…the programme was delivered by the researcher, a PE trained specialist, with extensive experience in the primary classroom."* (Riley 2015 [75])	*"Classroom teachers were responsible for the planning and the delivery of all movement-based lessons during the intervention."* (Riley 2016 [76])
Target Audience Bias	…the demographics of those that received the intervention in the current study to those who will receive the intervention in future evaluations?	No	Delivering intervention to more conducive, convenience sample or sample that is not representative of target population results in more effective intervention	Delivering intervention to sample of whom the intervention is intended results in less effective intervention compared to pilot	*"Although our sample size was... predominately white, and well-educated…"* (Sze 2015 [77])	
Intervention Duration Bias	…the length of the intervention provided in the current study to the length of the intervention in future evaluations?	No	Shorter duration results in more effective intervention	Longer duration less effective intervention compared to pilot	4-week intervention (Wilson 2005 [78])	17-week intervention (Wilson 2011 [79])
Setting Bias	…the setting where the intervention is delivered in the current study and the intervention delivery setting in future evaluations?	No	Delivering intervention in a more conducive, convenience location that is not representative of the target setting results in more effective intervention	Delivering intervention in a location more representative of target setting results in a less effective intervention compared to pilot	Intervention delivered on university campus [b]	Intervention delivered in community setting [b]
Measurement Bias	…the measures employed in the current study and the measures used in future evaluations of the intervention for primary/ secondary outcomes?	Yes	Use of less reliable or valid measures of primary/secondary outcomes results in more effective intervention	Use of more reliable and valid measures results in less effective intervention compared to pilot	Pedometer used to measure physical activity (Lubans 2009 [80])	Accelerometer used to measure physical activity (Lubans 2012 [81])
Directional Conclusions	Are the intervention effect(s) in the hypothesized direction?	No	Less effective intervention	Reduces intervention effectiveness	*"The decline in physical activity among the participants was not anticipated…"* (Cliff 2007 [82])	

Table 2 Operational Definitions of Risk of Generalizability Biases *(Continued)*

Risk of Generalizability Bias	Questions to Ask	Increased Presence with Small Sample	Hypothesized Influence of the Presence of Risk of Generalizability Bias		Example	
			Pilot	Larger-Scale Efficacy/ Effectiveness	Pilot	Larger-Scale Efficacy/ Effectiveness
Outcome Bias	Is the primary outcome for future evaluations of the intervention measured in the current study?	No	Absences of measuring primary outcome results in more effective intervention	Absence of primary outcome collected in pilot results in less effective intervention tested in well-powered trial	Nutrients sold per day and number of items sold per day in school cafeterias (Hartstein 2008 [83])	Self-reported daily dietary intake of students (Siega-Riz 2011 [84])

[a] Although not labeled as a pilot study, the example illustrates the presence of the risk of generalizability bias in one study and altered in the subsequent trial

[b] Hypothetical example of the risk of generalizability bias as it could operate in a pilot to larger-scale efficacy/effectiveness trial

units of the outcomes within pilot and efficacy/effectiveness pairs; however, there were numerous instances where similar constructs (e.g., physical activity, weight status) were measured in the pilot and efficacy/effectiveness study but were reported in different metrics across studies (e.g., steps in the pilot vs. minutes of activity in the efficacy/effectiveness or waist circumference in cm in the pilot and waist circumference in z-scores in the efficacy/effectiveness); therefore construct matching of the standardized effect size were used.

All effect sizes were corrected for differences in the direction of the scales so that positive effect sizes corresponded to improvements in the intervention group, independent of the original scale's direction. This correction was performed for simplicity of interpretive purposes so that all effect sizes were presented in the same direction and summarized within and across studies. The primary testing of the impact of the biases was performed by comparing the changing in the SMD from the pilot study to the larger, efficacy/effectiveness trial for studies coded with and without a given bias present. All studies reported more than one outcome effect across the seven constructs (e.g., BMI outcomes and dietary outcomes); therefore, summary effect sizes were calculated using a random-effects multi-level robust variance estimation meta-regression model [87–89], with constructs nested within studies nested within pairs. This modeling procedure is distribution free and can handle the non-independence of the effects sizes from multiple outcomes reported within a single study.

Criteria for evidence to support risk of generalizability biases

We examined the influence of the biases on the difference in SMD between the pilot and efficacy/effectiveness trials by testing the impact of each bias, separately, on the change in the SMD from the pilot to efficacy/effectiveness trial. All data were initially entered into Comprehensive Meta-Analysis (v.3.3.07) to calculate effect sizes for each reported outcome across constructs for all studies. The computed effect sizes, variances, and information regarding the presence/absence of the risk of generalizability biases were transferred into R (version 3.5.1) where a random-effects multi-level robust variance estimation meta-regression models were computed using the package "Metafor" [90].

Next, we examined whether the empirical evidence was in the hypothesized direction (see Table 2 for the biases and hypothesized directions). The final step was to examine the relationship between the presence of a bias and the sample size in the pilot and efficacy/effectiveness pairs. We hypothesized that the risk of generalizability biases would be more prevalent within smaller sized pilots. In pilot studies, a "small" sample size was classified as any pilot study with a total of 100 participants or less [91]. In absence of an established cutoff for efficacy/effectiveness trials, we defined a "small" sample size for the larger, more well-powered trials as any trial with 312 or fewer total participants. This size was based on the median sample size in the distribution of the sample in the identified well-powered trials.

Results

A PRISMA diagram for the literature search is presented in Fig. 1. For the identification of published pilot studies, a total of 132,021 citations were identified across search engines and keywords, with 24,570 representing unique articles. After title/abstract and full-text screenings, a total of 741 articles met the final full text criteria as a pilot behavioral intervention on a topic related to childhood obesity. For the review of reviews, we identified a total of 1576 review studies. Of these, 80 reviews on a childhood obesity-related topic were identified that cited 362 unique efficacy/effectiveness interventions trials. After searching these interventions for reference to pilot work and cross-referencing the study authors with the identified pilot studies, we were able to confirm 42 pilots paired to 39 unique efficacy/effectiveness trials of the same or similar intervention [29, 74–84, 92–158]. Of these, one pilot and efficacy/effectiveness pair [94, 96] did not report similar outcomes across studies and therefore were not included in the analytical models. Three of the efficacy/effectiveness trials [84, 124, 136] had each published two separate pilot studies, reporting on different outcomes from the same pilot study [83, 100, 103, 123, 125, 159] on the same intervention evaluated in the efficacy/effectiveness publication and were included as pairs with a single efficacy/effectiveness trial and two pilots, each. Across all studies, a total of 840 individual effect sizes were initially computed, representing 379 effect sizes from the pilot studies and 461 from the efficacy/effectiveness trials. Aggregating at the construct level reduced the total individual effects to 182 across 38 pairs, with an average of 2.4 constructs represented within a pair (range 1 to 5).

The prevalence of the risk of generalizability biases across the 39 pilot and efficacy/effectiveness pairs are graphically displayed across each pair in Fig. 2. Overall, the most commonly observed biases were delivery agent bias (19/39 pairs), duration bias (15/39), implementation support bias (13/39), outcome bias (6/39), measurement bias (4/39), directional conclusion bias (3/39), and target audience bias (3/39). A single bias (setting bias) was not coded across any of the pairs, while intervention intensity bias was only identified once. In the review of 39 pairs, we found evidence of carry forward of two biases (i.e., bias present in both

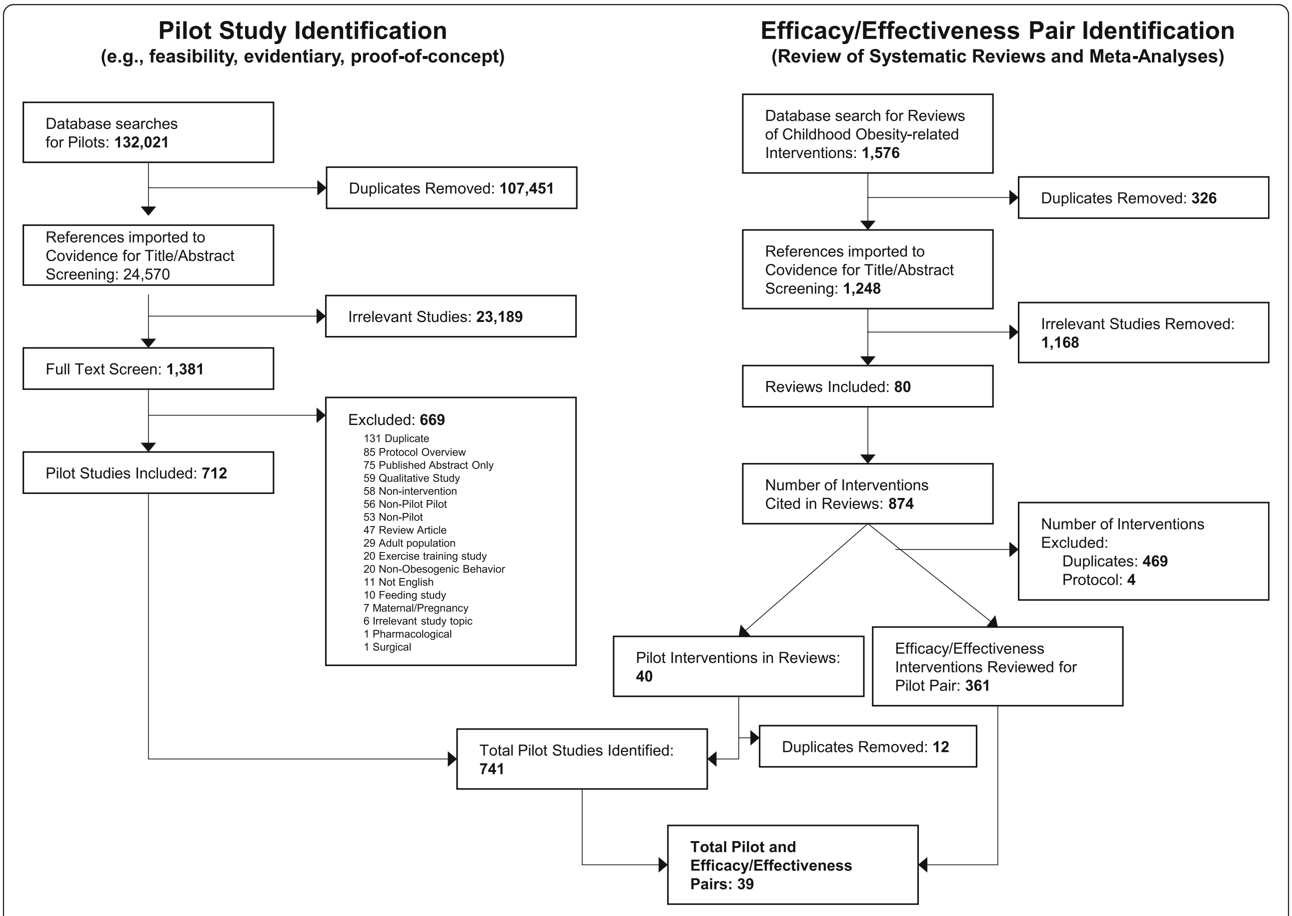

Pilot Study Identification
(e.g., feasibility, evidentiary, proof-of-concept)
Database searches for Pilots: 132,021
Duplicates Removed: 107,451
References imported to Covidence for Title/Abstract Screening: 24,570
Irrelevant Studies: 23,189
Full Text Screen: 1,381
Excluded: 669
131 Duplicate
85 Protocol Overview
75 Published Abstract Only
59 Qualitative Study
58 Non-intervention
56 Non-Pilot Pilot
53 Non-Pilot
47 Review Article
29 Adult population
20 Exercise training study
20 Non-Obesogenic Behavior
11 Not English
10 Feeding study
7 Maternal/Pregnancy
6 Irrelevant study topic
1 Pharmacological
1 Surgical
Pilot Studies Included: 712
Efficacy/Effectiveness Pair Identification
(Review of Systematic Reviews and Meta-Analyses)
Database search for Reviews of Childhood Obesity-related Interventions: 1,576
Duplicates Removed: 326
References imported to Covidence for Title/Abstract Screening: 1,248
Irrelevant Studies Removed: 1,168
Reviews Included: 80
Number of Interventions Cited in Reviews: 874
Number of Interventions Excluded:
Duplicates: 469
Protocol: 4
Pilot Interventions in Reviews: 40
Efficacy/Effectiveness Interventions Reviewed for Pilot Pair: 361
Duplicates Removed: 12
Total Pilot Studies Identified: 741
Total Pilot and Efficacy/Effectiveness Pairs: 39

Fig. 1 PRISMA diagram of literature search

pilot and efficacy/effectiveness) – delivery agent bias and implementation support bias, with 8/39 of pairs coded as carrying forward delivery agent bias, while 4/39 carrying forward implementation support bias. Outcome bias was observed in 6/39, however, given the requirement of aligning constructs for analytical comparison, no analyses were conducted on this bias. This resulted in a total of six biases, of the nine, that had sufficient data for the analytical models.

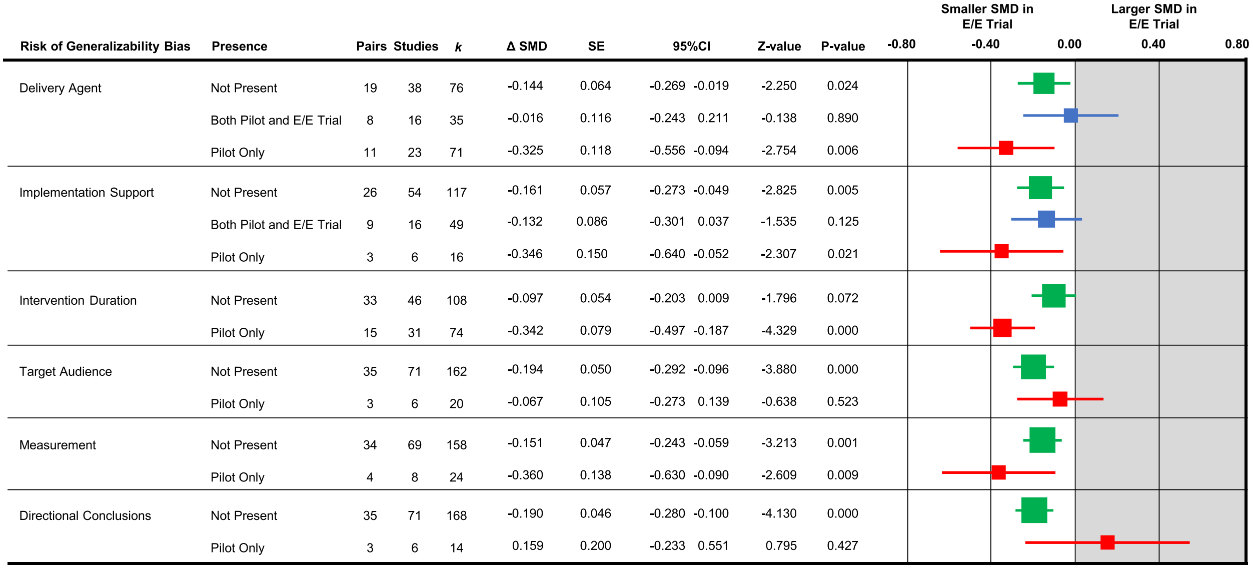
The strength of evidence supporting the potential influence of each of the six biases are presented in Fig. 3. For four of the generalizability biases – delivery agent, implementation support, intervention duration, and measurement – the difference in the SMD (i.e., the larger, more well-powered trial SMD minus the pilot SMD) was larger in the pairs of pilot studies that had the bias present and subsequently did not have the bias present in the larger, more well-powered trials, compared to pairs that did not have the biases present. Specifically, the change in the SMD was – 0.325 (95CI – 0.556 to – 0.094) for agent delivery, – 0.346 (– 0.640 to – 0.052) for implementation support, – 0.342 (– 0.498 to – 0.187) for intervention duration, and – 0.360 (– 0.631 to 0.089) for measurement. Two biases, target audience (– 0.067, – 0.274 to 0.139) and directional conclusions (0.159, – 0.233 to 0.551), were not associated with major changes in the SMD. For pairs where biases that were coded as present in both the pilot and in the larger, more well-powered trials there was no major difference in the SMD for delivery agent (SMD = – 0.016, – 0.243 to 0.212), while a small reduction in the SMD was observed for implementation support (SMD = – 0.132 (– 0.301 to 0.037).

The association of the presence of a bias with sample size of the pilot and efficacy/effectiveness pairs is presented in Fig. 4 for the three most prevalent biases (i.e., delivery agent, implementation support, and duration). Only 37 pairs were analyzed as two pairs [83, 84, 94, 96, 100] did not provide information on sample size at the child level, and therefore, could not be included in this analysis. Of the biases hypothesized to be influenced by smaller sample sizes, two demonstrated this pattern (i.e., implementation support and delivery agent, see Fig. 4). Of the 19 occurrences of delivery agent bias, 13 occurrences of implementation support bias, and 15 occurrences of intervention duration bias, these biases were coded in 16, 10, and 11 of the pairs with a pilot study

No. Biases	Stage	Reference	Sample Size (child level)	Intensity	Implementation Support	Delivery Agent	Target Audience	Duration	Directional Conclusions	Outcome	Measurement	Setting
0	Pilot	Liu 2008	753	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢
0	E-E	Liu 2010	4700	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢
0	Pilot	Bundy 2009	12	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢
0	E-E	Englen 2013	221	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢
0	Pilot	Stock 2007	383	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢
0	E-E	Santos 2014	647	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢
0	Pilot	Kain 2012	2688	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢
0	E-E	Kain 2014	1574	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢
1	Pilot	Cliff 2007	13	🟢	🟢	🟢	🟢	🟢	🔴	🟢	🟢	🟢
1	E-E	Cliff 2010	165	🟢	🟢	🟢	🟢	🟢	🔴	🟢	🟢	🟢
1	Pilot	Jago 2012	147	🟢	🟢	🟢	🟢	🔴	🟢	🟢	🟢	🟢
1	E-E	Jago 2016	571	🟢	🟢	🟢	🟢	🔴	🟢	🟢	🟢	🟢
1	Pilot	Sayoye 2005	25	🟢	🟢	🔴	🟢	🟢	🟢	🟢	🟢	🟢
1	E-E	Savoye 2007	209	🟢	🟢	🟠	🟢	🟢	🟢	🟢	🟢	🟢
1	Pilot	Riley 2015	54	🟢	🟢	🔴	🟢	🟢	🟢	🟢	🟢	🟢
1	E-E	Riley 2016	240	🟢	🟢	🟠	🟢	🟢	🟢	🟢	🟢	🟢
1	Pilot	Reilly 2003	60	🟢	🟢	🟢	🟢	🔴	🟢	🟢	🟢	🟢
1	E-E	Reilly 2006	545	🟢	🟢	🟢	🟢	🔴	🟢	🟢	🟢	🟢
1	Pilot	Beets 2014	895	🟢	🔴	🟢	🟢	🟢	🟢	🟢	🟢	🟢
1	E-E	Beets 2015	1991	🟢	🟠	🟢	🟢	🟢	🟢	🟢	🟢	🟢
1	Pilot	Huberty 2011	92	🟢	🔴	🟢	🟢	🟢	🟢	🟢	🟢	🟢
1	E-E	Huberty 2014	667	🟢	🔴	🟢	🟢	🟢	🟢	🟢	🟢	🟢
1	Pilot	Fahlman 2008	576	🟢	🟠	🟢	🟢	🟢	🟢	🟢	🟢	🟢
1	E-E	McCaughtry 2011	2132	🟢	🔴	🟢	🟢	🟢	🟢	🟢	🟢	🟢
1	Pilot	Ni Mhurchu 2008	20	🟢	🟢	🟢	🟢	🔴	🟢	🟢	🟢	🟢
1	E-E	Maddison 2011	322	🟢	🟢	🟢	🟢	🔴	🟢	🟢	🟢	🟢
1	Pilot	Adab 2014	488	🟢	🟢	🟢	🔴	🟢	🟢	🟢	🟢	🟢
1	E-E	Adab 2017	1392	🟢	🟢	🟢	🔴	🟢	🟢	🟢	🟢	🟢
1	Pilot	Robinson 2003	61	🟢	🟢	🟢	🟢	🔴	🟢	🟢	🟢	🟢
1	E-E	Robinson 2010	243	🟢	🟢	🟢	🟢	🔴	🟢	🟢	🟢	🟢
1	Pilot	Robbins 2012	69	🟢	🟢	🟢	🟢	🔴	🟢	🟢	🟢	🟢
1	E-E	Robbins 2016	1519	🟢	🟢	🟢	🟢	🔴	🟢	🟢	🟢	🟢
1	Pilot	Davis 2011	17	🟢	🟢	🔴	🟢	🟢	🟢	🟢	🟢	🟢
1	E-E	Davis 2013	58	🟢	🟢	🔴	🟢	🟢	🟢	🟢	🟢	🟢
1	Pilot	Eather 2013	49	🟢	🟢	🔴	🟢	🟢	🟢	🟢	🟢	🟢
1	E-E	Eather 2013	213	🟢	🟢	🔴	🟢	🟢	🟢	🟢	🟢	🟢
1	Pilot	Ebbeling 2005	103	🟢	🟢	🟢	🟢	🔴	🟢	🟢	🟢	🟢
1	E-E	Ebbeling 2012	224	🟢	🟢	🟢	🟢	🔴	🟢	🟢	🟢	🟢
1	Pilot	Grey 2004	41	🟢	🟢	🔴	🟢	🟢	🟢	🟢	🟢	🟢
1	E-E	Grey 2009	198	🟢	🟢	🟠	🟢	🟢	🟢	🟢	🟢	🟢
1	Pilot	Kipping 2008	472	🟢	🟢	🟢	🟢	🟢	🟢	🔴	🟢	🟢
1	Pilot	Kipping 2010	506	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢
1	E-E	Kipping 2014	2221	🟢	🟢	🟢	🟢	🟢	🟢	🔴	🟢	🟢

No. Biases	Stage	Reference	Sample Size (child level)	Intensity	Implementation Support	Delivery Agent	Target Audience	Duration	Directional Conclusions	Outcome	Measurement	Setting
1	Pilot	Morgan 2011	71	🟢	🟢	🔴	🟢	🟢	🟢	🟢	🟢	🟢
1	E-E	Morgan 2014	132	🟢	🟢	🟠	🟢	🟢	🟢	🟢	🟢	🟢
2	Pilot	Wilson 2005	48	🟢	🟢	🔴	🟢	🔴	🟢	🟢	🟢	🟢
2	E-E	Wilson 2011	1563	🟢	🟢	🟠	🟢	🔴	🟢	🟢	🟢	🟢
2	Pilot	Madsen 2009	174	🟢	🟢	🔴	🟢	🟢	🟢	🔴	🟢	🟢
2	E-E	Madsen 2013	156	🟢	🟢	🔴	🟢	🟢	🟢	🔴	🟢	🟢
2	Pilot	Sacher 2005	11	🟢	🟢	🔴	🟢	🔴	🟢	🟢	🟢	🟢
2	E-E	Sacher 2010	116	🟢	🟢	🔴	🟢	🔴	🟢	🟢	🟢	🟢
2	Pilot	Edwards 2006	33	🟢	🟢	🔴	🟢	🔴	🟢	🟢	🟢	🟢
2	E-E	Croker 2012	60	🟢	🟢	🔴	🟢	🔴	🟢	🟢	🟢	🟢
2	Pilot	Neumark-Sztainer 2003	201	🟢	🔴	🔴	🟢	🟢	🟢	🟢	🟢	🟢
2	E-E	Neumark-Sztainer 2010	356	🟢	🔴	🔴	🟢	🟢	🟢	🟢	🟢	🟢
2	Pilot	Patrick 2001	117	🟢	🟢	🟢	🟢	🔴	🟢	🟢	🔴	🟢
2	E-E	Patrick 2006	819	🟢	🟢	🟢	🟢	🔴	🟢	🟢	🔴	🟢
2	Pilot	Lloyd 2011	202	🟢	🟢	🔴	🔴	🟢	🟢	🟢	🟢	🟢
2	E-E	Lloyd 2017	1244	🟢	🟢	🟠	🔴	🟢	🟢	🟢	🟢	🟢
3	Pilot	Robertson 2008	27	🟢	🔴	🔴	🟢	🟢	🟢	🟢	🔴	🟢
3	E-E	Robertson 2016	128	🟢	🟠	🟠	🟢	🟢	🟢	🟢	🔴	🟢
3	Pilot	Paul 2011	30	🟢	🟠	🔴	🟢	🟢	🔴	🟢	🟢	🟢
3	E-E	Paul 2018	279	🟢	🔴	🔴	🟢	🟢	🔴	🟢	🟢	🟢
3	Pilot	Jones 2011	97	🟢	🟠	🔴	🟢	🟢	🔴	🟢	🟢	🟢
3	E-E	Jones 2015	150	🟢	🔴	🟠	🟢	🟢	🔴	🟢	🟢	🟢
3	Pilot	Cullen 2007	6 [a]	🟢	🟠	🟢	🟢	🔴	🟢	🔴	🟢	🟢
3	Pilot	Hartstein 2008	6 [a]	🟢	🟠	🟢	🟢	🔴	🟢	🔴	🟢	🟢
3	E-E	Siega-Riz 2011	3908	🟢	🔴	🟢	🟢	🔴	🟢	🔴	🟢	🟢
3	Pilot	Benjamin 2007	19 [a]	🟠	🟠	🟢	🟢	🟢	🟢	🔴	🟢	🟢
3	E-E	Alkon 2014	552	🔴	🔴	🟢	🟢	🟢	🟢	🔴	🟢	🟢
3	Pilot	Smith 2013	17	🟢	🟠	🔴	🟢	🟢	🟢	🔴	🟢	🟢
3	E-E	Hoza 2014	94	🟢	🔴	🔴	🟢	🟢	🟢	🔴	🟢	🟢
3	Pilot	Beech 2003	60	🟢	🔴	🔴	🟢	🔴	🟢	🟢	🟢	🟢
3	E-E	Klesges 2010	303	🟢	🟠	🟠	🟢	🔴	🟢	🟢	🟢	🟢
3	Pilot	Dudley 2010	38	🟢	🟠	🔴	🟢	🔴	🟢	🟢	🟢	🟢
2	Pilot	Andruschko 2018	20	🟢	🟠	🔴	🟢	🔴	🟢	🟢	🟢	🟢
3	E-E	Okely 2017	1199	🟢	🔴	🟠	🟢	🔴	🟢	🟢	🟢	🟢
4	Pilot	Lubans 2009	106	🟢	🟢	🟢	🔴	🔴	🟢	🔴	🔴	🟢
4	E-E	Lubans 2012	357	🟢	🟢	🟢	🔴	🔴	🟢	🔴	🔴	🟢
4	Pilot	Ni Mhurchu 2009	29	🟢	🟠	🔴	🟢	🔴	🟢	🟢	🔴	🟢
4	E-E	Maddison 2014	251	🟢	🔴	🟠	🟢	🔴	🟢	🟢	🔴	🟢

Fig. 2 Presence of risk of generalizability biases in pilot and larger-scale efficacy/effectiveness pairs. Note: Red circle (🔴) indicates bias present, green circle (🟢) bias not present, orange circle (🟠) bias identified in pilot or well-powered but not the other. E-E = Efficacy/Effectiveness. [a] Sample size represents setting level (e.g., school, childcare) – child-level sample size not reported



Risk of Generalizability Bias	Presence	Pairs	Studies	k	Δ SMD	SE	95%CI		Z-value	P-value
Delivery Agent	Not Present	19	38	76	-0.144	0.064	-0.269	-0.019	-2.250	0.024
	Both Pilot and E/E Trial	8	16	35	-0.016	0.116	-0.243	0.211	-0.138	0.890
	Pilot Only	11	23	71	-0.325	0.118	-0.556	-0.094	-2.754	0.006
Implementation Support	Not Present	26	54	117	-0.161	0.057	-0.273	-0.049	-2.825	0.005
	Both Pilot and E/E Trial	9	16	49	-0.132	0.086	-0.301	0.037	-1.535	0.125
	Pilot Only	3	6	16	-0.346	0.150	-0.640	-0.052	-2.307	0.021
Intervention Duration	Not Present	33	46	108	-0.097	0.054	-0.203	0.009	-1.796	0.072
	Pilot Only	15	31	74	-0.342	0.079	-0.497	-0.187	-4.329	0.000
Target Audience	Not Present	35	71	162	-0.194	0.050	-0.292	-0.096	-3.880	0.000
	Pilot Only	3	6	20	-0.067	0.105	-0.273	0.139	-0.638	0.523
Measurement	Not Present	34	69	158	-0.151	0.047	-0.243	-0.059	-3.213	0.001
	Pilot Only	4	8	24	-0.360	0.138	-0.630	-0.090	-2.609	0.009
Directional Conclusions	Not Present	35	71	168	-0.190	0.046	-0.280	-0.100	-4.130	0.000
	Pilot Only	3	6	14	0.159	0.200	-0.233	0.551	0.795	0.427

Fig. 3 Forest plot of the change in the standardized mean difference (SMD) of the presence, absence, or carry forward of six risk of generalizability biases from a pilot to larger-scale efficacy/effectiveness (E/E) trial

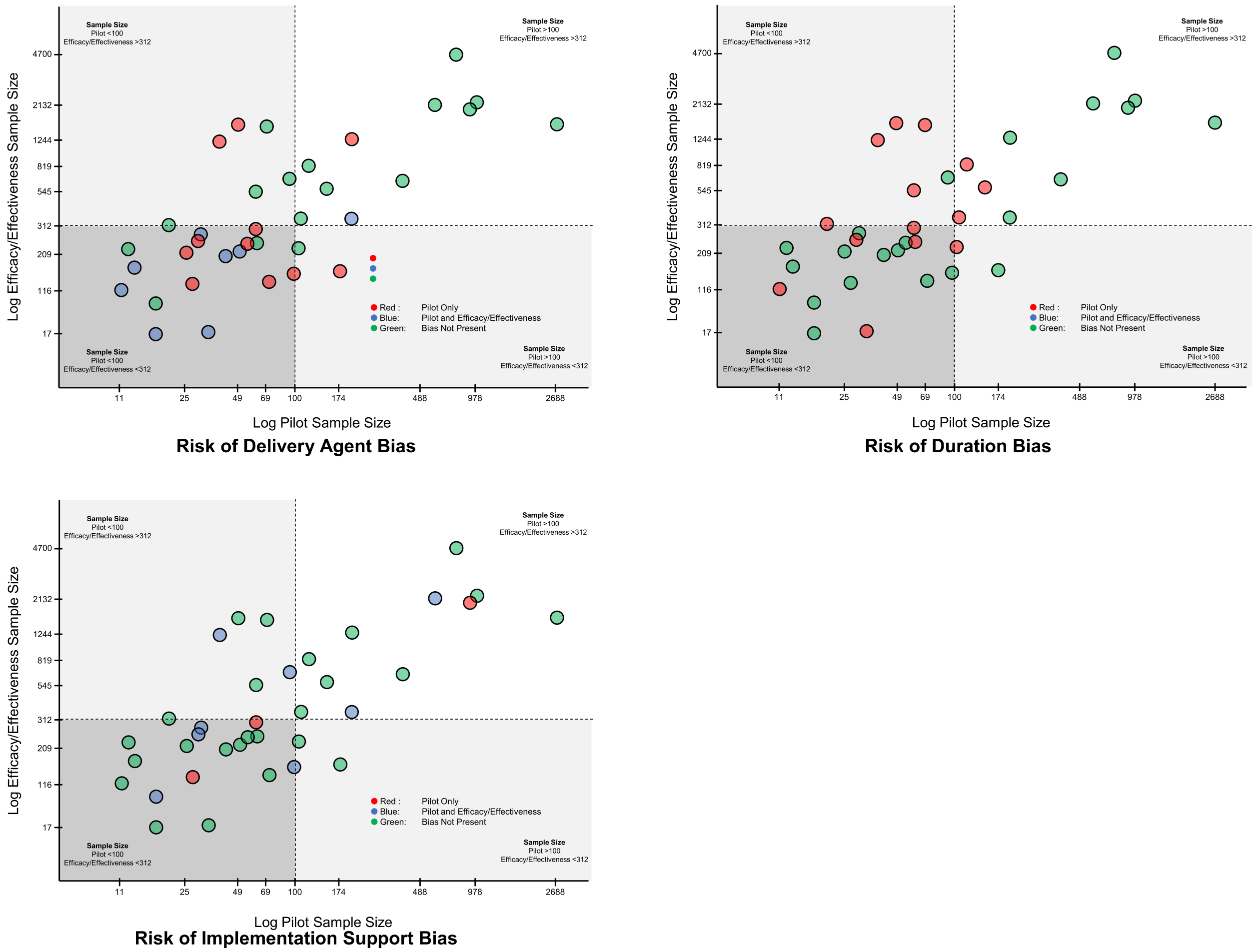
classified as having a small sample size ($N = 100$ or less), respectively, [91].

Discussion

The purpose of the current study was to define a preliminary set of risk of generalizability biases, specific to the early stages of testing of an intervention, provide a conceptual basis for their presence and to present evidence of their influence within a sample of pilot and the larger, more well-powered efficacy/effectiveness trial pairs on a topic related to childhood obesity. The identification of these biases should assist interventionists in avoiding the unintentional effects of biases related to external validity during the early stages of designing, conducting, and interpreting the outcomes from an intervention, as well as for reviewers of grants and manuscripts to determine whether the presence of one or more of the proposed biases may lead to exaggerated early discoveries [66] and subsequent failed efficacy/effectiveness trials.

In this study we identified 9 biases in pilot tested interventions that investigators, to a large extent, have control over whether or not they are introduced. These biases do not have to be introduced unless there is a strong and compelling rationale for their inclusion. One possible argument for including one or more of the risk of generalizability biases in a pilot (e.g., having a doctoral student deliver an intervention, testing the intervention over a short/abbreviated time period) are the resources available to conduct the study. Across the 39 pilot and efficacy/effectiveness pairs a total of 31 indicated the receipt of funding: 11 pilots were associated with NIH funding sources, 3 with sources from the National Institute for Health Research, 2 from the CDC, 11 from a foundation, and 4 from university or department/college level grants. "Well-funded" pilots, those with funding from the NIH, CDC or NIHR, contained biases at a similar rate as those considered to have lower amounts of funding (university/departmental award or foundation). Of the "well-funded" pilot studies, over 50% included risk of delivery agent bias, or risk of duration bias, while 42% included risk of implementation support bias.

While we could not confirm the total grant funding award for many of the pilot studies, of those where publicly available information was available, they received sizable awards to conduct the pilot study (e.g., NIH awards of R21 grants for 2 years and US$275,000 total direct costs). Interestingly, the resources to conduct a pilot, as evidenced by the receipt of federal grants, therefore, does not appear to be associated with the introduction or absence of a risk of generalizability bias. Thus, there must be alternative reasons that lead interventionists to include risk of generalizability biases in their pilot studies. At this time, however, it is unclear what rationale may be used for justifying the inclusion of risk of generalizability bias, particularly for those risk of generalizability biases that demonstrated the strongest relationship with differences in effect size estimations. Possible reasons may include the pressure to demonstrate initial feasibility and acceptability and potential efficacy which would then increase the chance of receiving funding for a larger study, the need for "statistically significant' effects for publication, existing paradigms that endorse highly controlled studies prior to more real-world contexts or a combination of one or more of these



Risk of Delivery Agent Bias

Risk of Duration Bias

Risk of Implementation Support Bias

Fig. 4 Association of the three most prevalent risk of generalizability biases with pilot and efficacy/effectiveness sample size. Note: The x- and y-axis represent the log of the total sample size per study. The tick marks represent the actual total sample size across the range of sample sizes in the studies.

reasons [24, 160, 161]. This may be a function of the pressures of securing grant funding for promotion or keeping a research laboratory operating [162].

With the creation of any new intervention there is a risk of it not being feasible, acceptable or potentially efficacious. Testing a new intervention on a small scale is a logical decision given the high-risk associated with the intervention not resulting in the anticipated effects [163]. Smaller scale studies are less resource intensive, compared to efficacy/effectiveness studies and thus, are a natural choice for pilot studies. It is also important to recognize that early "evidence of promise" from studies that may have design weaknesses is often used to secure further research funding and as such pilot studies often have in-built design limitations. Because a study is small in scale, it does not imply that the risks of generalizability biases described herein should be introduced. Our findings indicate, however, that a "small sample" size appears to serve as a proxy for the introduction of some of the biases that demonstrated the most influence on study level effects. This susceptibility to the biases, such as delivery agent bias and implementation support bias can, from a practical standpoint, operate more easily with smaller sample sizes. Interestingly, not all small sample pilot studies had evidence of delivery agent bias, implementation support bias, or duration bias, indicating small sample size studies can be conducted without the biases.

It is reasonable to assume that certain aspects of an intervention would (and at times should) be modified based upon the results of the pilot testing. Piloting an intervention affords this opportunity – the identification of potentially ineffective elements and their removal or the identification of missing components within an intervention that are theoretically and/or logically linked to the final interventions' success in a larger-scale trial. If changes are necessary and, perhaps substantial, re-testing the intervention under pilot conditions (e.g., smaller sized study) is necessary. In fact, the ORBIT model calls for multiple pilot tests of an intervention to ensure it is ready for efficacy/effectiveness testing [61]. Within the sample of pilot and efficacy/effectiveness trial pairs, we identified many pilot studies whose findings suggested the next testing of the intervention should have been another pilot, instead of the larger-scale, efficacy/effectiveness trial identified. Part of the decision to move forward, despite evidence suggesting further refinement and testing of the refinements is necessary, could be attributed to incentives such as the need to secure future grant funding. In the efficacy/effectiveness literature, optimistically interpreting findings, despite evidence of the contrary, is referred to as "spin" [164, 165]. How such a concept applies to pilot studies is unclear and needs further exploration to whether "spin" is operating as a bias during the early stages of testing an intervention. Across our literature searches, we found no evidence of multiple pilot studies being conducted prior to the efficacy/effectiveness trial. Of the pilot to efficacy/effectiveness pairs that had two pilot studies published, these were pilot studies reporting different outcomes from the same pilot testing, rather than a sequential process of pilots. This suggests that published pilot studies, at least within the field of childhood obesity, are conducted only once, with interventionists utilizing the results (either positive or null) to justify the larger-scale evaluation of the intervention.

Our findings highlight that intervention researchers need to carefully consider whether information obtained from pilot tests of an intervention delivered by highly trained research team members, with extensive support for intervention delivery, over short timeframes with different measures than are to be used in the larger-trial can be sustained and is consistent with what is intended to-be-delivered in the efficacy/effectiveness trial. Including one or more of these biases in a pilot study could result in inflated estimates of effectiveness during the pilot and lead interventionists to believe the intervention is more effective than the actual effect achieved when delivered in a efficacy/effectiveness trial without these biases [14, 26, 166]. These are critical decisions because, if the purpose of a pilot study is to determine whether a large-scale trial is warranted, yet the outcomes observed from the pilot study are contingent upon the features included in the pilot that are not intended to be or cannot be carried forward in an efficacy/effectiveness trial, the likelihood of observing limited or null results in the efficacy/effectiveness trial is high. This scenario renders the entire purpose of conducting a pilot evaluation of an intervention a meaningless exercise that can waste substantial time and resources, both during the pilot and the larger-scale evaluation of an ineffective intervention.

Based on these findings, the following is recommended:

1. Carefully consider the impact of the risk of generalizability biases in the design, delivery, and interpretation of pilot, even in small sample size pilots and their potential impact on the decision to progress to a larger-scale trial
2. All pilots should be published, and efficacy/effectiveness studies should reference pilot work
3. When reporting pilot studies, information should be presented on the presence of the risk of generalizability biases and their impact on the outcomes reported discussed
4. When reviewers (e.g., grant, manuscript) review pilot intervention studies, evidence of the presence and impact of the risk of generalizability biases should be considered
5. If a pilot was "unsuccessful", it should not be scaled-up but rather modified accordingly and re-piloted

Despite the initial evidence presented to support the utility of the risk of generalizability biases, there are several limitations that need to be considered. First, the sample in this study was limited to only 39 pilot and efficacy/effectiveness pairs, despite identifying over 700 published pilot and over 360 efficacy/effectiveness intervention studies. The publication of pilots, in addition to the clear reference to pilot work in efficacy/effectiveness studies needs to be made to ensure linkages between pilot and efficacy/effectiveness studies can be made. Second, a possibility exists that the over- or underestimation of effects reported herein are also due to unmeasured biases, beyond the risk of generalizability biases investigated here, and thus, readers need to take this into consideration when evaluating the impact of the risk of generalizability biases. Third, the absence of a risk of generalizability bias does not infer that there was no bias. Rather, it simply refers to the inability to identify evidence in a published study of the presence of a given risk of generalizability bias. Hence, one or more of the risk of generalizability biases could have been present, yet not reported in a published study and therefore be undetectable. Fourth, it is possible that in the search we missed some pilot and larger-scale study pairs due to a lack of clear labeling of pilot studies. Finally, the evidence presented was only gathered from a single topic area – childhood obesity. It is unclear if the risk of generalizability biases exists and operate similarly within other intervention topics or if new risk of generalizability biases would be discovered that were not identified herein. Future studies need to explore this to develop an exhaustive list of recommendations/considerations for interventionists developing, testing, and interpreting outcomes from pilot intervention studies.

In conclusion, pilot studies represent an essential and necessary step in the development and eventual widespread distribution of public health behavioral interventions. The evidence presented herein indicates there are risk of generalizability biases that are introduced during the pilot stage. These biases may influence whether an intervention will be successful during a larger, more well-powered efficacy/effectiveness trial. These risk of generalizability biases should be considered during the early planning and design phase of a pilot and the interpretation of the results both for interventionists and reviewers of grants and scientific manuscripts. Thus, testing an intervention at the early stages under conditions that it would not be tested again may not provide sufficient evidence to evaluate whether a larger-scale trial is warranted. Future studies need to continue to refine and expand the list of risk of generalizability biases and evaluate their presence with study level effects across different social science and public health behavioral intervention topic areas.

Acknowledgements
None.

Authors' contributions
MB secured the funding for the study and conceptualized the research questions. All authors contributed equally to interpreting the data and drafting and revising the manuscript for scientific clarity. All authors read and approved the final manuscript.

Authors' information
NA

Funding
Research reported in this publication was supported by the National Heart, Lung, And Blood Institute of the National Institutes of Health under Award Number R01HL149141. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Availability of data and materials
Access to the data will be made available upon completion of the entire project.

Ethics approval and consent to participate
This research was approved by the Institutional Review Board of the University of South Carolina.

Consent for publication
NA

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Arnold School of Public Health, University of South Carolina, Columbia, SC, USA. [2]Departments of Medicine, of Health Research and Policy, of Biomedical Data Science, and of Statistics, and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA. [3]Early Start, Faculty of Social Sciences, University of Wollongong, Wollongong, NSW, Australia. [4]Priority Research Centre in Physical Activity and Nutrition, School of Education, University of Newcastle, Callaghan, New South Wales, Australia. [5]Centre for Diet and Activity Research & MRC Epidemiology Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK. [6]Centre for Exercise Nutrition & Health Sciences, School for Policy Studies, University of Bristol, Bristol, UK. [7]New South Wales (NSW) Ministry of Health, St Leonards, NSW, Australia. [8]Sydney Medical School, The University of Sydney, Sydney, Australia.

Received: 5 August 2019 Accepted: 23 January 2020
Published online: 11 February 2020

References

1. Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. J Eval Clin Pract. 2004;10:307–12.
2. Leon AC, Davis LL, Kraemer HC. The role and interpretation of pilot studies in clinical research. J Psychiatr Res. 2011;45:626–9.
3. Stevens J, Taber DR, Murray DM, Ward DS. Advances and controversies in the design of obesity prevention trials. Obesity. 2007;15:2163–70.
4. Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, Robson R, Thabane M, Giangregorio L, Goldsmith CH. A tutorial on pilot studies: the what, why and how. BMC Med Res Methodol. 2010;10:1.
5. van Teijlingen E, Hundley V. The importance of pilot studies. Nurs Stand. 2002;16:33–6.
6. Eldridge SM, Lancaster GA, Campbell MJ, Thabane L, Hopewell S, Coleman CL, Bond CM. Defining Feasibility and Pilot Studies in Preparation for Randomised Controlled Trials: Development of a Conceptual Framework. PLoS One. 2016;11:e0150205.
7. Pilot Effectiveness Trials for Treatment, Preventive and Services Interventions (R34) [http://grants.nih.gov/grants/guide/rfa-files/RFA-MH-16-410.html]. Accessed Feb 2018.
8. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. Br Med J. 2008;337:a1655.
9. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L. Lancaster GA, group pc: CONSORT 2010 statement: extension to randomised pilot and feasibility trials. Pilot Feasibility Stud. 2016;2:64.
10. Arain M, Campbell MJ, Cooper CL, Lancaster GA. What is a pilot or feasibility study? A review of current practice and editorial policy. BMC Med Res Methodol. 2010;10:67.
11. Arnold DM, Burns KE, Adhikari NK, Kho ME, Meade MO, Cook DJ. McMaster critical care interest G: the design and interpretation of pilot trials in clinical research in critical care. Crit Care Med. 2009;37:S69–74.
12. Duffett M, Choong K, Hartling L, Menon K, Thabane L, Cook DJ. Pilot randomized trials in pediatric critical care: a systematic review. Pediatr Crit Care Med. 2015;16:e239–44.
13. Eldridge SM, Costelloe CE, Kahan BC, Lancaster GA, Kerry SM. How big should the pilot study for my cluster randomised trial be? Stat Methods Med Res. 2016;25:1039–56.
14. Hoddinott P. A new era for intervention development studies. Pilot Feasibility Stud. 2015;1:36.
15. de Bruin M, McCambridge J, Prins JM. Reducing the risk of bias in health behaviour change trials: improving trial design, reporting or bias assessment criteria? A review and case study. Psychol Health. 2015;30:8–34.
16. The Cochrane Handbook for Systematic Reviews of Interventions: Handbook is 5.1 [updated March 2011] [http://handbook.cochrane.org]. Accessed Jan 2018.
17. Shadish W, Cook T, Campbell D. Experimental and quasi-experimental designs for generalized casual inferences. Belmont: Wadsworth; 2002.
18. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG, Consort. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Int J Surg. 2012;10:28–55.
19. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet. 2001;357:1191–4.
20. Glasgow RE, Emmons KM. How can we increase translation of research into practice? Types of evidence needed. Annu Rev Public Health. 2007; 28:413–33.
21. Glasgow RE, Lichtenstein E, Marcus AC. Why don't we see more translation of health promotion research to practice? Rethinking the efficacy-to-effectiveness transition. Am J Public Health. 2003;93:1261–7.

22. Klesges LM, Estabrooks PA, Dzewaltowski DA, Bull SS, Glasgow RE. Beginning with the application in mind: designing and planning health behavior change interventions to enhance dissemination. Ann Behav Med. 2005;29(Suppl):66–75.
23. Chambers DA, Glasgow RE, Stange KC. The dynamic sustainability framework: addressing the paradox of sustainment amid ongoing change. Implement Sci. 2013;8:117.
24. Flay BR. Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. Prev Med. 1986;15:451–74.
25. Fitzgibbon ML, Stolley MR, Schiffer L, Van Horn L, KauferChristoffel K, Dyer A. Two-year follow-up results for hip-hop to health Jr.: a randomized controlled trial for overweight prevention in preschool minority children. J Pediatr. 2005;146:618–25.
26. Kong A, Buscemi J, Stolley MR, Schiffer LA, Kim Y, Braunschweig CL, Gomez-Perez SL, Blumstein LB, Van Horn L, Dyer AR, Fitzgibbon ML. Hip-Hop to Health Jr. Randomized effectiveness trial: 1-year follow-up results. Am J Prev Med. 2016;50:136–44.
27. Cohen KE, Morgan PJ, Plotnikoff RC, Callister R, Lubans DR. Physical activity and skills intervention: SCORES cluster randomized controlled trial. Med Sci Sports Exerc. 2015;47:765–74.
28. Sutherland RL, Nathan NK, Lubans DR, Cohen K, Davies LJ, Desmet C, Cohen J, McCarthy NJ, Butler P, Wiggers J, Wolfenden L. An RCT to facilitate implementation of school practices known to increase physical activity. Am J Prev Med. 2017;53:818–28.
29. Beets MW, Weaver RG, Turner-McGrievy G, Huberty J, Ward DS, Pate RR, Freedman D, Hutto B, Moore JB, Bottai M, et al. Physical activity outcomes in afterschool programs: a group randomized controlled trial. Prev Med. 2016;90:207–15.
30. Beets MW, Glenn Weaver R, Brazendale K, Turner-McGrievy G, Saunders RP, Moore JB, Webster C, Khan M, Beighle A. Statewide dissemination and implementation of physical activity standards in afterschool programs: two-year results. BMC Public Health. 2018;18:819.
31. Sutherland R, Reeves P, Campbell E, Lubans DR, Morgan PJ, Nathan N, Wolfenden L, Okely AD, Gillham K, Davies L, Wiggers J. Cost effectiveness of a multi-component school-based physical activity intervention targeting adolescents: the 'Physical activity 4 Everyone' cluster randomized trial. Int J Behav Nutr Phys Act. 2016;13:94.
32. McKenzie TL, Nader PR, Strikmiller PK, Yang M, Stone EJ, Perry CL, Taylor WC, Epping JN, Feldman HA, Luepker RV, Kelder SH. School physical education: effect of the child and adolescent trial for cardiovascular health. Prev Med. 1996;25:423–31.
33. Hoelscher DM, Feldman HA, Johnson CC, Lytle LA, Osganian SK, Parcel GS, Kelder SH, Stone EJ, Nader PR. School-based health education programs can be maintained over time: results from the CATCH institutionalization study. Prev Med. 2004;38:594–606.
34. Salmon J, Ball K, Hume C, Booth M, Crawford D. Outcomes of a group-randomized trial to prevent excess weight gain, reduce screen behaviours and promote physical activity in 10-year-old children: switch-play. Int J Obes. 2008;32:601–12.
35. West F, Sanders MR, Cleghorn GJ, Davies PS. Randomised clinical trial of a family-based lifestyle intervention for childhood obesity involving parents as the exclusive agents of change. Behav Res Ther. 2010;48:1170–9.
36. Gerards SM, Dagnelie PC, Gubbels JS, van Buuren S, Hamers FJ, Jansen MW, van der Goot OH, de Vries NK, Sanders MR, Kremers SP. The effectiveness of lifestyle triple P in the Netherlands: a randomized controlled trial. PLoS One. 2015;10:e0122240.
37. Salmon J, Jorna M, Hume C, Arundell L, Chahine N, Tienstra M, Crawford D. A translational research intervention to reduce screen behaviours and promote physical activity among children: Switch-2-activity. Health Promot Int. 2011;26:311–21.
38. Yoong SL, Wolfenden L, Clinton-McHarg T, Waters E, Pettman TL, Steele E, Wiggers J. Exploring the pragmatic and explanatory study design on outcomes of systematic reviews of public health interventions: a case study on obesity prevention trials. J Public Health (Oxf). 2014;36:170–6.
39. McCrabb S, Lane C, Hall A, Milat A, Bauman A, Sutherland R, Yoong S, Wolfenden L. Scaling-up evidence-based obesity interventions: a systematic review assessing intervention adaptations and effectiveness and quantifying the scale-up penalty. Obes Rev. 2019;20(7):964–82. https://onlinelibrary.wiley.com/doi/full/10.1111/obr.12845.
40. Billingham SA, Whitehead AL, Julious SA. An audit of sample sizes for pilot and feasibility trials being undertaken in the United Kingdom registered in the United Kingdom clinical research network database. BMC Med Res Methodol. 2013;13:104.
41. Indig D, Lee K, Grunseit A, Milat A, Bauman A. Pathways for scaling up public health interventions. BMC Public Health. 2017;18:68.
42. Milat AJ, Bauman A, Redman S. Narrative review of models and success factors for scaling up public health interventions. Implement Sci. 2015;10:113.
43. Milat AJ, King L, Bauman A, Redman S. Scaling up health promotion interventions: an emerging concept in implementation science. Health Promot J Austr. 2011;22:238.
44. Milat AJ, King L, Bauman AE, Redman S. The concept of scalability: increasing the scale and potential adoption of health promotion interventions into policy and practice. Health Promot Int. 2013;28:285–98.
45. Milat AJ, Newson R, King L, Rissel C, Wolfenden L, Bauman A, Redman S, Giffin M. A guide to scaling up population health interventions. Public Health Res Pract. 2016;26:e2611604.
46. O'Hara BJ, Bauman AE, Eakin EG, King L, Haas M, Allman-Farinelli M, Owen N, Cardona-Morell M, Farrell L, Milat AJ, Phongsavan P. Evaluation framework for translational research: case study of Australia's get healthy information and coaching service(R). Health Promot Pract. 2013;14:380–9.
47. O'Hara BJ, Phongsavan P, King L, Develin E, Milat AJ, Eggins D, King E, Smith J, Bauman AE. 'Translational formative evaluation': critical in up-scaling public health programmes. Health Promot Int. 2014;29:38–46.
48. Redman S, Turner T, Davies H, Williamson A, Haynes A, Brennan S, Milat A, O'Connor D, Blyth F, Jorm L, Green S. The SPIRIT action framework: a structured approach to selecting and testing strategies to increase the use of research in policy. Soc Sci Med. 2015;136-137:147–55.
49. World Health Organization. Begining with the End in Mind: Planning pilot projects and other programmatic research for sucessful scaling up. France: WHO; 2011. https://apps.who.int/iris/bitstream/handle/10665/44708/9789241502320_eng.pdf;jsessionid=F51B37DE2EF6215F95067CD7C13D4234?sequence=1.
50. Chambers DA, Norton WE. The Adaptome: advancing the science of intervention adaptation. Am J Prev Med. 2016;51:S124–31.
51. Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. BMJ. 2015;350:h2147.
52. Zwarenstein M, Treweek S, Loudon K. PRECIS-2 helps researchers design more applicable RCTs while CONSORT extension for pragmatic trials helps knowledge users decide whether to apply them. J Clin Epidemiol. 2017;84:27–9.
53. Treweek S, Zwarenstein M. Making trials matter: pragmatic and explanatory trials and the problem of applicability. Trials. 2009;10:37.
54. Glasgow RE, Bull SS, Gillette C, Klesges LM, Dzewaltowski DA. Behavior change intervention research in healthcare settings: a review of recent reports with emphasis on external validity. Am J Prev Med. 2002;23:62–9.
55. Green LW, Glasgow RE. Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. Eval Health Prof. 2006;29:126–53.
56. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, Griffey R, Hensley M. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Admin Pol Ment Health. 2011;38:65–76.
57. Durlak JA, DuPre EP. Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. Am J Community Psychol. 2008;41:327–50.
58. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci. 2009;4:50.
59. Koorts H, Eakin E, Estabrooks P, Timperio A, Salmon J, Bauman A. Implementation and scale up of population physical activity interventions for clinical and community settings: the PRACTIS guide. Int J Behav Nutr Phys Act. 2018;15:51.
60. Davidson KW, Goldstein M, Kaplan RM, Kaufmann PG, Knatterud GL, Orleans CT, Spring B, Trudeau KJ, Whitlock EP. Evidence-based behavioral medicine: what is it and how do we achieve it? Ann Behav Med. 2003;26:161–71.
61. Czajkowski SM, Powell LH, Adler N, Naar-King S, Reynolds KD, Hunter CM, Laraia B, Olster DH, Perna FM, Peterson JC, et al. From ideas to efficacy: the ORBIT model for developing behavioral treatments for chronic diseases. Health Psychol. 2015;34:971–82.
62. Glasgow RE, Klesges LM, Dzewaltowski DA, Bull SS, Estabrooks P. The future of health behavior change research: what is needed to improve translation of research into health promotion practice? Ann Behav Med. 2004;27:3–12.

63. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, Altman DG, Barbour V, Macdonald H, Johnston M, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ. 2014;348:g1687.
64. Des Jarlais DC, Lyles C, Crepaz N, Group T. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. Am J Public Health. 2004;94:361–6.
65. Chan AW, Tetzlaff JM, Altman DG, Laupacis A, Gotzsche PC, Krleza-Jeric K, Hrobjartsson A, Mann H, Dickersin K, Berlin JA, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. Ann Intern Med. 2013;158:200–7.
66. Ioannidis JP. Scientific inbreeding and same-team replication: type D personality as an example. J Psychosom Res. 2012;73:408–10.
67. Cutler DM. Behavioral health interventions: what works and why? In: Anderson NB, Bulatao RA, Cohen B, editors. Critical Perspectives on Racial and Ethnic Differences in Health in Late Life. Washington, DC: The National Academies Press; 2004. p. 643–76.
68. Collins LM, Nahum-Shani I, Almirall D. Optimization of behavioral dynamic treatment regimens based on the sequential, multiple assignment, randomized trial (SMART). Clin Trials. 2014;11:426–34.
69. Rubio DM, Schoenbaum EE, Lee LS, Schteingart DE, Marantz PR, Anderson KE, Platt LD, Baez A, Esposito K. Defining translational research: implications for training. Acad Med. 2010;85:470–5.
70. Efficacy and Mechanism Evaluation programme: Mechansitic Studies, Expanation and Examples [https://www.nihr.ac.uk/documents/mechanistic-studies-explanation-and-examples/12146]. Accessed Mar 2018.
71. Casadevall A, Fang FC. Descriptive science. Infect Immun. 2008;76:3835–6.
72. Behavioral and Social Sciences Research Definitions [https://obssr.od.nih.gov/about-us/bssr-definition/]. Accessed Apr 2018.
73. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009;6:e1000097.
74. Beech BM, Klesges RC, Kumanyika SK, Murray DM, Klesges L, McClanahan B, Slawson D, Nunnally C, Rochon J, McLain-Allen B. Child-and parent-targeted interventions: the Memphis GEMS pilot study. Ethn Dis. 2003;13:S1–40.
75. Riley N, Lubans DR, Morgan PJ, Young M. Outcomes and process evaluation of a programme integrating physical activity into the primary school mathematics curriculum: the EASY minds pilot randomised controlled trial. J Sci Med Sport. 2015;18:656–61.
76. Riley N, Lubans DR, Holmes K, Morgan PJ. Findings from the EASY minds cluster randomized controlled trial: evaluation of a physical activity integration program for mathematics in primary schools. J Phys Act Health. 2016;13:198–206.
77. Sze YY, Daniel TO, Kilanowski CK, Collins RL, Epstein LH. Web-Based and Mobile Delivery of an Episodic Future Thinking Intervention for Overweight and Obese Families: A Feasibility Study. JMIR Mhealth Uhealth 2015;3(4):e97. https://doi.org/10.2196/mhealth.4603. PMC: PMC4704914.
78. Wilson DK, Evans AE, Williams J, Mixon G, Sirard JR, Pate R. A preliminary test of a student-centered intervention on increasing physical activity in underserved adolescents. Ann Behav Med. 2005;30:119.
79. Wilson DK, Van Horn ML, Kitzman-Ulrich H, Saunders R, Pate R, Lawman HG, Hutto B, Griffin S, Zarrett N, Addy CL. Results of the "active by choice today"(ACT) randomized trial for increasing physical activity in low-income and minority adolescents. Health Psychol. 2011;30:463.
80. Lubans DR, Morgan PJ, Callister R, Collins CE. Effects of integrating pedometers, parental materials, and E-mail support within an extracurricular school sport intervention. J Adolesc Health. 2009;44:176–83.
81. Lubans DR, Morgan PJ, Okely AD, Dewar D, Collins CE, Batterham M, Callister R, Plotnikoff RC. Preventing obesity among adolescent girls: one-year outcomes of the nutrition and enjoyable activity for teen girls (NEAT girls) cluster randomized controlled trial. Arch Pediatr Adolesc Med. 2012;166:821–7.
82. Cliff DP, Wilson A, Okely AD, Mickle KJ, Steele JR. Feasibility of SHARK: a physical activity skill-development program for overweight and obese children. J Sci Med Sport. 2007;10:263–7.
83. Hartstein J, Cullen KW, Reynolds KD, Harrell J, Resnicow K, Kennel P. Studies to treat or prevent pediatric type 2 diabetes prevention study group: impact of portion-size control for school a la carte items: changes in kilocalories and macronutrients purchased by middle school students. J Am Diet Assoc. 2008;108:140–4.
84. Siega-Riz AM, El Ghormli L, Mobley C, Gillis B, Stadler D, Hartstein J, Volpe SL, Virus A, Bridgman J. The effects of the HEALTHY study intervention on middle school student dietary intakes. Int J Behav Nutr Phys Act. 2011;8:7.
85. Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychol Methods. 2002;7:105–25.
86. Waters E, de Silva-Sanigorski A, Hall BJ, Brown T, Campbell KJ, Gao Y, Armstrong R, Prosser L, Summerbell CD. Interventions for preventing obesity in children. Cochrane Database Syst Rev. 2011;Issue 12. Art. No.: CD001871. https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD001871.pub3/full.
87. Tipton E. Small sample adjustments for robust variance estimation with meta-regression. Psychol Methods. 2015;20:375–93.
88. Tanner-Smith EE, Tipton E, Polanin JR. Handling complex meta-analytic data structures using robust variance estimates: a tutorial in R. J Dev Life Course Criminol. 2016;2:85–112.
89. Konstantopoulos S. Fixed effects and variance components estimation in three-level meta-analysis. Res Synth Methods. 2011;2:61–76.
90. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Softw. 2010;36:1–48.
91. Stice E, Shaw H, Marti CN. A meta-analytic review of obesity prevention programs for children and adolescents: the skinny on interventions that work. Psychol Bull. 2006;132:667–91.
92. Adab P, Pallan MJ, Cade J, Ekelund U, Barrett T, Daley A, Deeks J, Duda J, Gill P, Parry J. Preventing childhood obesity, phase II feasibility study focusing on south Asians: BEACHeS. BMJ Open. 2014;4:e004579.
93. Adab P, Pallan MJ, Lancashire ER, Hemming K, Frew E, Barrett T, Bhopal R, Cade JE, Canaway A, Clarke JL. Effectiveness of a childhood obesity prevention programme delivered through schools, targeting 6 and 7 year olds: cluster randomised controlled trial (WAVES study). BMJ. 2018;360:k211.
94. Alkon A, Crowley AA, Neelon SEB, Hill S, Pan Y, Nguyen V, Rose R, Savage E, Forestieri N, Shipman L. Nutrition and physical activity randomized control trial in child care centers improves knowledge, policies, and children's body mass index. BMC Public Health. 2014;14:215.
95. Beets MW, Weaver RG, Moore JB, Turner-McGrievy G, Pate RR, Webster C, Beighle A. From policy to practice: strategies to meet physical activity standards in YMCA afterschool programs. Am J Prev Med. 2014;46:281–8.
96. Benjamin SE, Ammerman A, Sommers J, Dodds J, Neelon B, Ward DS. Nutrition and physical activity self-assessment for child care (NAP SACC): results from a pilot intervention. J Nutr Educ Behav. 2007;39:142–9.
97. Bundy AC, Luckett T, Tranter PJ, Naughton GA, Wyver SR, Ragen J, Spies G. The risk is that there is 'no risk': a simple, innovative intervention to increase children's activity levels. Int J Early Years Educ. 2009;17:33–45.
98. Cliff DP, Okely AD, Morgan PJ, Steele JR, Jones RA, Colyvas K, Baur LA. Movement skills and physical activity in obese children: randomized controlled trial. Med Sci Sports Exerc. 2011;43:90–100.
99. Croker H, Viner RM, Nicholls D, Haroun D, Chadwick P, Edwards C, Wells JC, Wardle J. Family-based behavioural treatment of childhood obesity in a UK National Health Service setting: randomized controlled trial. Int J Obes. 2012;36:16.
100. Cullen KW, Hartstein J, Reynolds KD, Vu M, Resnicow K, Greene N, White MA. Studies to treat or prevent pediatric type 2 diabetes prevention study group: improving the school food environment: results from a pilot study in middle schools. J Am Diet Assoc. 2007;107:484–9.
101. Davis AM, James RL, Boles RE, Goetz JR, Belmont J, Malone B. The use of TeleMedicine in the treatment of paediatric obesity: feasibility and acceptability. Matern Child Nutr. 2011;7:71–9.
102. Davis AM, Sampilo M, Gallagher KS, Landrum Y, Malone B. Treating rural pediatric obesity through telemedicine: outcomes from a small randomized controlled trial. J Pediatr Psychol. 2013;38:932–43.
103. Dudley DA, Okely AD, Pearson P, Peat J. Engaging adolescent girls from linguistically diverse and low income backgrounds in school sport: a pilot randomised controlled trial. J Sci Med Sport. 2010;13:217–24.
104. Eather N, Morgan PJ, Lubans DR. Improving the fitness and physical activity levels of primary school children: results of the Fit-4-fun group randomized controlled trial. Prev Med. 2013;56:12–9.
105. Eather N, Morgan PJ, Lubans DR. Feasibility and preliminary efficacy of the Fit4Fun intervention for improving physical fitness in a sample of primary school children: a pilot study. Phys Educ Sport Pedagog. 2013;18:389–411.
106. Ebbeling CB, Feldman HA, Chomitz VR, Antonelli TA, Gortmaker SL, Osganian SK, Ludwig DS. A randomized trial of sugar-sweetened beverages and adolescent body weight. N Engl J Med. 2012;367:1407–16.

107. Ebbeling CB, Feldman HA, Osganian SK, Chomitz VR, Ellenbogen SJ, Ludwig DS. Effects of decreasing sugar-sweetened beverage consumption on body weight in adolescents: a randomized, controlled pilot study. Pediatrics. 2006; 117:673–80.
108. Edwards C, Nicholls D, Croker H, Van Zyl S, Viner R, Wardle J. Family-based behavioural treatment of obesity: acceptability and effectiveness in the UK. Eur J Clin Nutr. 2006;60:587.
109. Engelen L, Bundy AC, Naughton G, Simpson JM, Bauman A, Ragen J, Baur L, Wyver S, Tranter P, Niehues A. Increasing physical activity in young primary school children—it's child's play: a cluster randomised controlled trial. Prev Med. 2013;56:319–25.
110. Fahlman MM, Dake JA, McCaughtry N, Martin J. A pilot study to examine the effects of a nutrition intervention on nutrition knowledge, behaviors, and efficacy expectations in middle school children. J Sch Health. 2008;78:216–22.
111. Grey M, Berry D, Davidson M, Galasso P, Gustafson E, Melkus G. Preliminary testing of a program to prevent type 2 diabetes among high-risk youth. J Sch Health. 2004;74:10–5.
112. Grey M, Jaser SS, Holl MG, Jefferson V, Dziura J, Northrup V. A multifaceted school-based intervention to reduce risk for type 2 diabetes in at-risk youth. Prev Med. 2009;49:122–8.
113. Hoza B, Smith AL, Shoulberg EK, Linnea KS, Dorsch TE, Blazo JA, Alerding CM, McCabe GP. A randomized trial examining the effects of aerobic physical activity on attention-deficit/hyperactivity disorder symptoms in young children. J Abnorm Child Psychol. 2015;43:655–67.
114. Huberty JL, Beets MW, Beighle A, Saint-Maurice PF, Welk G. Effects of ready for recess, an environmental intervention, on physical activity in third-through sixth-grade children. J Phys Act Health. 2014;11:384–95.
115. Huberty JL, Siahpush M, Beighle A, Fuhrmeister E, Silva P, Welk G. Ready for recess: a pilot study to increase physical activity in elementary school children. J Sch Health. 2011;81:251–7.
116. Jago R, Edwards M, Sebire S, Bird E, Tomkinson K, Kesten J, Banfield K, May T, Cooper A, Blair P. Bristol girls dance project: a cluster randomised controlled trial of an after-school dance programme to increase physical activity among 11-to 12-year-old girls. Public Health Res. 2016;4(6):1–175.
117. Jago R, Edwards MJ, Sebire SJ, Tomkinson K, Bird EL, Banfield K, May T, Kesten JM, Cooper AR, Powell JE. Effect and cost of an after-school dance programme on the physical activity of 11–12 year old girls: the Bristol girls dance project, a school-based cluster randomised controlled trial. Int J Behav Nutr Phys Act. 2015;12:128.
118. Jago R, Sebire SJ, Cooper AR, Haase AM, Powell J, Davis L, McNeill J, Montgomery AA. Bristol girls dance project feasibility trial: outcome and process evaluation results. Int J Behav Nutr Phys Act. 2012;9:83.
119. Jones RA, Okely AD, Hinkley T, Batterham M, Burke C. Promoting gross motor skills and physical activity in childcare: a translational randomized controlled trial. J Sci Med Sport. 2016;19:744–9.
120. Jones RA, Riethmuller A, Hesketh K, Trezise J, Batterham M, Okely AD. Promoting fundamental movement skill development and physical activity in early childhood settings: a cluster randomized controlled trial. Pediatr Exerc Sci. 2011;23:600–15.
121. Kain J, Concha F, Moreno L, Leyton B. School-based obesity prevention intervention in Chilean children: effective in controlling, but not reducing obesity. J Obes. 2014;2014:618293.
122. Kain J, Uauy R, Vio F, Cerda R, Leyton B. School-based obesity prevention in Chilean primary school children: methodology and evaluation of a controlled study. Int J Obes. 2004;28:483.
123. Kipping R, Payne C, Lawlor DA. Randomised controlled trial adapting American school obesity prevention to England. Arch Dis Child. 2008;93: 469–73.
124. Kipping RR, Howe LD, Jago R, Campbell R, Wells S, Chittleborough CR, Mytton J, Noble SM, Peters TJ, Lawlor DA. Effect of intervention aimed at increasing physical activity, reducing sedentary behaviour, and increasing fruit and vegetable consumption in children: active for life year 5 (AFLY5) school based cluster randomised controlled trial. BMJ. 2014;348:g3256.
125. Kipping RR, Jago R, Lawlor DA. Diet outcomes of a pilot school-based randomised controlled obesity prevention study with 9–10 year olds in England. Prev Med. 2010;51:56–62.
126. Klesges RC, Obarzanek E, Kumanyika S, Murray DM, Klesges LM, Relyea GE, Stockton MB, Lanctot JQ, Beech BM, McClanahan BS. The Memphis Girls' health enrichment multi-site studies (GEMS): an evaluation of the efficacy of a 2-year obesity prevention program in African American girls. Arch Pediatr Adolesc Med. 2010;164:1007–14.
127. Liu A, Hu X, Ma G, Cui Z, Pan Y, Chang S, Zhao W, Chen C. Evaluation of a classroom-based physical activity promoting programme. Obes Rev. 2008;9:130–4.
128. Lloyd J, Creanor S, Logan S, Green C, Dean SG, Hillsdon M, Abraham C, Tomlinson R, Pearson V, Taylor RS. Effectiveness of the healthy lifestyles Programme (HeLP) to prevent obesity in UK primary-school children: a cluster randomised controlled trial. Lancet Child Adolesc Health. 2018;2:35–45.
129. Lloyd JJ, Wyatt KM, Creanor S. Behavioural and weight status outcomes from an exploratory trial of the healthy lifestyles Programme (HeLP): a novel school-based obesity prevention programme. BMJ Open. 2012;2:e000390.
130. Maddison R, Marsh S, Foley L, Epstein LH, Olds T, Dewes O, Heke I, Carter K, Jiang Y, Ni Mhurchu C. Screen-time weight-loss intervention targeting children at home (SWITCH): a randomized controlled trial. Int J Behav Nutr Phys Act. 2014;11:111.
131. Madsen K, Thompson H, Adkins A, Crawford Y. School-community partnerships: a cluster-randomized trial of an after-school soccer program. JAMA Pediatr. 2013;167:321–6.
132. Madsen KA, Thompson HR, Wlasiuk L, Queliza E, Schmidt C, Newman TB. After-school program to reduce obesity in minority children: a pilot study. J Child Health Care. 2009;13:333–46.
133. McCaughtry N, Fahlman M, Martin JJ, Shen B. Influences of constructivist-oriented nutrition education on urban middle school Students' nutrition knowledge, self-efficacy, and behaviors. Am J Health Educ. 2011;42:276–85.
134. Ni Mhurchu C, Roberts V, Maddison R, Dorey E, Jiang Y, Jull A, Tin ST. Effect of electronic time monitors on children's television watching: pilot trial of a home-based intervention. Prev Med. 2009;49:413–7.
135. Neumark-Sztainer D, Story M, Hannan PJ, Rex J. New moves: a school-based obesity prevention program for adolescent girls. Prev Med. 2003;37:41–51.
136. Okely AD, Lubans DR, Morgan PJ, Cotton W, Peralta L, Miller J, Batterham M, Janssen X. Promoting physical activity among adolescent girls: the girls in sport group randomized trial. Int J Behav Nutr Phys Act. 2017;14:81.
137. Patrick K, Calfas KJ, Norman GJ, Zabinski MF, Sallis JF, Rupp J, Covin J, Cella J. Randomized controlled trial of a primary care and home-based intervention for physical activity and nutrition behaviors: PACE+ for adolescents. Arch Pediatr Adolesc Med. 2006;160:128–36.
138. Patrick K, Sallis JF, Prochaska JJ, Lydston DD, Calfas KJ, Zabinski MF, Wilfley DE, Saelens BE, Brown DR. A multicomponent program for nutrition and physical activity change in primary care: PACE+ for adolescents. Arch Pediatr Adolesc Med. 2001;155:940–6.
139. Paul IM, Savage JS, Anzman SL, Beiler JS, Marini ME, Stokes JL, Birch LL. Preventing obesity during infancy: a pilot study. Obesity (Silver Spring). 2011;19:353–61.
140. Paul IM, Savage JS, Anzman-Frasca S, Marini ME, Beiler JS, Hess LB, Loken E, Birch LL. Effect of a responsive parenting educational intervention on childhood weight outcomes at 3 years of age: the INSIGHT randomized clinical trial. JAMA. 2018;320:461–8.
141. Reilly JJ, Kelly L, Montgomery C, Williamson A, Fisher A, McColl JH, Conte RL, Paton JY, Grant S. Physical activity to prevent obesity in young children: cluster randomised controlled trial. BMJ. 2006;333:1041.
142. Reilly JJ, McDowell ZC. Physical activity interventions in the prevention and treatment of paediatric obesity: systematic review and critical appraisal. Proc Nutr Soc. 2003;62:611–9.
143. Robbins LB, Ling J, Sharma DB, Dalimonte-Merckling DM, Voskuil VR, Resnicow K, Kaciroti N, Pfeiffer KA. Intervention effects of “girls on the move” on increasing physical activity: a group randomized trial. Ann Behav Med. 2018;53:493–500.
144. Robbins LB, Pfeiffer KA, Maier KS, Lo Y-J, Wesolek SM. Pilot intervention to increase physical activity among sedentary urban middle school girls: a two-group pretest–posttest quasi-experimental design. J Sch Nurs. 2012;28:302–15.
145. Robertson W, Fleming J, Kamal A, Hamborg T, Khan KA, Griffiths F, Stewart-Brown S, Stallard N, Petrou S, Simkiss D. Randomised controlled trial evaluating the effectiveness and cost-effectiveness of'Families for Health', a family-based childhood obesity treatment intervention delivered in a community setting for ages 6 to 11 years. Health Technol Assess. 2017;21:1.
146. Robertson W, Friede T, Blissett J, Rudolf MC, Wallis MA, Stewart-Brown S. Pilot of'Families for Health': community-based family intervention for obesity. Arch Dis Child. 2008;93:921–6.
147. Robinson TN, Killen JD, Kraemer HC, Wilson DM, Matheson DM, Haskell WL, Pruitt LA, Powell TM, Owens A, Thompson N. Dance and reducing television viewing to prevent weight gain in African-American girls: the Stanford GEMS pilot study. Ethn Dis. 2003;13:S1–65.

148. Robinson TN, Matheson DM, Kraemer HC, Wilson DM, Obarzanek E, Thompson NS, Alhassan S, Spencer TR, Haydel KF, Fujimoto M. A randomized controlled trial of culturally tailored dance and reducing screen time to prevent weight gain in low-income African American girls: Stanford GEMS. Arch Pediatr Adolesc Med. 2010;164:995–1004.
149. Sacher P, Chadwick P, Wells J, Williams J, Cole T, Lawson M. Assessing the acceptability and feasibility of the MEND Programme in a small group of obese 7–11-year-old children. J Hum Nutr Diet. 2005;18:3–5.
150. Santos RG, Durksen A, Rabbani R, Chanoine J-P, Miln AL, Mayer T, McGavock JM. Effectiveness of peer-based healthy living lesson plans on anthropometric measures and physical activity in elementary school students: a cluster randomized trial. JAMA Pediatr. 2014;168:330–7.
151. Savoye M, Berry D, Dziura J, Shaw M, Serrecchia JB, Barbetta G, Rose P, Lavietes S, Caprio S. Anthropometric and psychosocial changes in obese adolescents enrolled in a weight management program. J Am Diet Assoc. 2005;105:364–70.
152. Smith AL, Hoza B, Linnea K, McQuade JD, Tomb M, Vaughn AJ, Shoulberg EK, Hook H. Pilot physical activity intervention reduces severity of ADHD symptoms in young children. J Atten Disord. 2013;17:70–82.
153. Stock S, Miranda C, Evans S, Plessis S, Ridley J, Yeh S, Chanoine J-P. Healthy buddies: a novel, peer-led health promotion program for the prevention of obesity and eating disorders in children in elementary school. Pediatrics. 2007;120:e1059–68.
154. Li Y-P, Hu X-Q, Schouten EG, Liu A-L, Du S-M, Li L-Z, Cui Z-H, Wang D, Kok FJ, Hu FB. Report on childhood obesity in China (8): effects and sustainability of physical activity intervention on body composition of Chinese youth. Biomed Environ Sci. 2010;23:180–7.
155. Morgan PJ, Lubans DR, Callister R, Okely AD, Burrows TL, Fletcher R, Collins CE. The 'Healthy dads, healthy Kids' randomized controlled trial: efficacy of a healthy lifestyle program for overweight fathers and their children. Int J Obes. 2011;35:436–47.
156. Morgan PJ, Collins CE, Plotnikoff RC, Callister R, Burrows T, Fletcher R, Okely AD, Young MD, Miller A, Lloyd AB, et al. The 'Healthy dads, healthy Kids' community randomized controlled trial: a community-based healthy lifestyle program for fathers and their children. Prev Med. 2014;61:90–9.
157. Savoye M, Shaw M, Dziura J, Tamborlane WV, Rose P, Guandalini C, Goldberg-Gell R, Burgert TS, Cali AM, Weiss R, Caprio S. Effects of a weight management program on body composition and metabolic parameters in overweight children: a randomized controlled trial. JAMA. 2007;297:2697–704.
158. Ni Mhurchu C, Maddison R, Jiang Y, Jull A, Prapavessis H, Rodgers A. Couch potatoes to jumping beans: a pilot study of the effect of active video games on physical activity in children. Int J Behav Nutr Phys Act. 2008;5:8.
159. Andruschko J, Okely AD, Pearson P. A school-based physical activity and motor devleopment program for low-fit adolescent females: The Sport4Fun pilot randomized controlled trial. J Motor Learn Dev. 2018;6:345–56.
160. Nosek BA, Spies JR, Motyl M. Scientific Utopia: II. Restructuring incentives and practices to promote truth over Publishability. Perspect Psychol Sci. 2012;7:615–31.
161. Flay BR, Biglan A, Boruch RF, Castro FG, Gottfredson D, Kellam S, Moscicki EK, Schinke S, Valentine JC, Ji P. Standards of evidence: criteria for efficacy, effectiveness and dissemination. Prev Sci. 2005;6:151–75.
162. Wawer J. How to stop salami science - promotion of healthy trends in publishing behaviour. Account Res. 2018. https://doi.org/10.1080/08989621.2018.1556099.
163. Bacchetti P, Deeks SG, McCune JM. Breaking free of sample size dogma to perform innovative translational research. Sci Transl Med. 2011;3:87ps24.
164. Khan MS, Lateef N, Siddiqi TJ, Rehman KA, Alnaimat S, Khan SU, Riaz H, Murad MH, Mandrola J, Doukky R, Krasuski RA. Level and prevalence of spin in published cardiovascular randomized clinical trial reports with statistically nonsignificant primary outcomes: a systematic review. JAMA Netw Open. 2019;2:e192622.
165. Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. JAMA. 2010;303:2058–64.
166. Beets MW, Glenn Weaver R, Turner-McGrievy G, Saunders RP, Webster CA, Moore JB, Brazendale K, Chandler J. Evaluation of a statewide dissemination and implementation of physical activity intervention in afterschool programs: a nonrandomized trial. Transl Behav Med. 2017;7:690–701.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.