Beets, M. W., Weaver, R. G., Ioannidis, J. P. A., Geraci, M., Brazendale, K., Decker, L., ... Milat, A. J. (2020). Identification and Evaluation of Risk of Generalizability Biases in Pilot versus Efficacy/Effectiveness Trials: A Systematic Review and Meta-Analysis. *International Journal of Behavioral Nutrition and Physical Activity, 17*, [19 (2020)].
https://doi.org/10.1186/s12966-020-0918-y
Identification and evaluation of risk of generalizability biases in pilot versus efficacy/effectiveness trials: a systematic review and meta-analysis

Michael W. Beets1*, R. Glenn Weaver1, John P. A. Ioannidis2, Marco Geraci1, Keith Brazendale1, Lindsay Decker1, Anthony D. Okely3, David Lubans4, Esther van Sluijs5, Russell Jago6, Gabrielle Turner-McGrievy1, James Thrasher1, Xiaming Li1 and Andrew J. Milat7,8

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
1Arnold School of Public Health, University of South Carolina, Columbia, SC, USA
Full list of author information is available at the end of the article

© The Author(s). 2020 Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
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 largescale 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”
| Study | Likely Larger Effect | Likely Smaller/No Effect |
|-------|----------------------|--------------------------|
| Study Fitzgibbon 2005 [25] | “...the use of specially trained early childhood educators rather than classroom teachers to deliver the intervention, thereby raising questions of generalizability.” | Kong 2016 [26] |
| 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 |
| 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] | Haelscher 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 |
| How long was the intervention delivered? | 10 months | “...Switch-2-Activity involved an abbreviated programme; therefore, the intervention ‘dose’ was lower...” |
| 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...” |
| Bias                                      | Likely Larger Effect                                                                 | Likely Smaller/No Effect                                                                 |
|------------------------------------------|--------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|
| Study                                    | Objective measures                                                                  | Self-report                                                                             |
| What measures were used to collect       | West 2010 [35]                                                                       | Gerards 2015 [36]                                                                       |
| information on outcomes?                | “All sessions were facilitated by a clinical psychologist and accredited provider of| “The intervention was led by three different facilitators. These health professionals  |
|                                          | the intervention (who co-authored the intervention materials), with assistance from  | have been accredited after attending an official 3-day training course and an additional |
|                                          | graduate students in nutrition and dietetics, physical education, and psychology.”   | intervention day.”                                                                       |
| Who delivered the intervention?          | “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 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
| Risk of Generalizability Bias | Questions to Ask | Increased Presence with Small Sample | Hypothesized Influence of the Presence of Risk of Generalizability Bias | Example |
|-------------------------------|------------------|-------------------------------------|---------------------------------------------------------------|---------|
| 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 | 19 lessons delivered (Salmon 2008 [34])<sup>a</sup> 6 lessons delivered (Salmon 2011 [37])<sup>a</sup> |
| 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 | “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 | “… the programme was delivered by the researches, a PE trained specialist, with extensive experience in the primary classroom.” (Riley 2015 [75]) |
| 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 | “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 | 4-week intervention (Wilson 2005 [78]) |
| 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 | Intervention delivered on university campus<sup>b</sup> |
| 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 | Pedometer used to measure physical activity (Lubans 2009 [80]) |
| Directional Conclusions       | Are the intervention effect(s) in the hypothesized direction? | No | Less effective intervention | “The decline in physical activity among the participants was not anticipated…” (Cliff 2007 [82]) |
| Risk of Generalizability Bias | Questions to Ask | Increased Presence with Small Sample | Hypothesized Influence of the Presence of Risk of Generalizability Bias | Example |
|-------------------------------|------------------|-------------------------------------|-------------------------------------------------|---------|
| 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]) |

*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.

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 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 $-0.325$ ($-0.556$ 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) | Efficacy/Effectiveness  | Intervention Support | Delivery Agent | Target Audience | Directional Conclusions | Outcome | Measurement Setting |
|------------|-------|-------------|---------------------------|-------------------------|----------------------|-----------------|-----------------|------------------------|---------|---------------------|
| 0          | Pilot | Liu 2008    | 753                       |                         |                      |                 |                 |                        |         |                     |
| 0          | E-E   | Liu 2010    | 4700                      |                         |                      |                 |                 |                        |         |                     |
| 0          | Pilot | Bundy 2009  | 12                        |                         |                      |                 |                 |                        |         |                     |
| 0          | E-E   | Engelen 2013| 221                       |                         |                      |                 |                 |                        |         |                     |
| 0          | Pilot | Stock 2007  | 383                       |                         |                      |                 |                 |                        |         |                     |
| 0          | E-E   | Santos 2014 | 647                       |                         |                      |                 |                 |                        |         |                     |
| 0          | Pilot | Kan 2012    | 2688                      |                         |                      |                 |                 |                        |         |                     |
| 0          | E-E   | Kan 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 | Sayoyo 2005 | 25                        |                         |                      |                 |                 |                        |         |                     |
| 1          | E-E   | Sayoyo 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 | Fahman 2008 | 576                       |                         |                      |                 |                 |                        |         |                     |
| 1          | E-E   | Mc Çaughn 2011| 2132                 |                         |                      |                 |                 |                        |         |                     |
| 1          | Pilot | Ni Mhurchu| 20                        |                         |                      |                 |                 |                        |         |                     |
| 1          | E-E   | Madding 2011| 322                      |                         |                      |                 |                 |                        |         |                     |
| 1          | Pilot | Abd 2014    | 488                       |                         |                      |                 |                 |                        |         |                     |
| 1          | E-E   | Abd 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 | Esther 2013 | 49                        |                         |                      |                 |                 |                        |         |                     |
| 1          | E-E   | Esther 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 | Kippen 2008 | 472                       |                         |                      |                 |                 |                        |         |                     |
| 1          | E-E   | Kippen 2010 | 506                       |                         |                      |                 |                 |                        |         |                     |
| 1          | E-E   | Kippen 2014 | 2221                      |                         |                      |                 |                 |                        |         |                     |

**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. * Sample size represents setting level (e.g., school, childcare) – child-level sample size not reported.
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

---

**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**

| Risk of Generalizability Bias | Presence | Pairs | Studies | $k$ | SMD | SE | 95%CI | Z-value | P-value | Smaller SMD in E/E Trial | Larger SMD in E/E Trial |
|-------------------------------|----------|-------|---------|-----|------|----|-------|---------|---------|--------------------------|------------------------|
| Delivery Agent                | Not Present | 19 | 38 | 76 | -0.144 | 0.054 | -0.269, -0.019 | -2.250 | 0.024 |
| Both Pilot and E/E Trial     | Not Present | 5 | 16 | 35 | -0.016 | 0.116 | -0.243, 0.211 | -1.138 | 0.286 |
| Pilot Only                    | Not Present | 11 | 23 | 71 | -0.325 | 0.118 | -0.556, -0.094 | -2.754 | 0.006 |
| Implementation Support       | Not Present | 26 | 54 | 117 | -0.181 | 0.057 | -0.273, -0.089 | -2.825 | 0.005 |
| Both Pilot and E/E Trial     | Not Present | 9 | 16 | 49 | -0.132 | 0.066 | -0.301, 0.037 | -1.830 | 0.068 |
| Pilot Only                    | Not Present | 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                    | Not Present | 15 | 31 | 74 | -0.342 | 0.079 | -0.497, -0.187 | -4.328 | 0.000 |
| Target Audience              | Not Present | 35 | 71 | 162 | -0.194 | 0.050 | -0.292, -0.096 | -3.860 | 0.000 |
| Pilot Only                    | Not Present | 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                    | Not Present | 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 | -0.289, -0.100 | -4.130 | 0.000 |
| Pilot Only                    | Not Present | 3 | 6 | 14 | 0.159 | 0.200 | -0.233, 0.551 | 0.766 | 0.447 |
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, retesting the intervention under pilot conditions (e.g.,

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.
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 under-estimation 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 is 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. 2Departments 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. 3Early Start, Faculty of Social Sciences, University of Wollongong, Wollongong, NSW, Australia. 4Priority Research Centre in Physical Activity and Nutrition, School of Education, University of Newcastle, Callaghan, New South Wales, Australia. 5Centre for Diet and Activity Research & MRC Epidemiology Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK. 6Centre for Exercise Nutrition & Health Sciences, School for Policy Studies, University of Bristol, Bristol, UK. 7New South Wales (NSW) Ministry of Health, St Leonards, NSW, Australia. 8Sydney 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, Hundleby 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 (R3T) [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. Arai 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, Bums 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;37:569–74.
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, Kenny SM. How big should the pilot study for my cluster randomised trial be? Stat Methods Med Res. 2016;25:1035–56.
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;357:191–4.
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.
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. Segawa RZ, EM Al 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.

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.ncbi.nlm.nih.gov/cdb/doi/cdb/doi10.1002/14651858.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 estimation: a tutorial in R. J Dev Life Course Criminol. 2016;8:25–112.

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, Ekulund 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, Hemmings 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, Neelion SEB, Hill S, Pan Y, Nguyen V, Rose R, Savage E, Forresti 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;14:2135.

95. Beets MW, Weaver RG, Moore JB, Turner-McGrew G, Pete 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, Trantier PJ, Naughton GA, Wyer 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;43:590–100.

99. Croker H, Viner RM, Nichols D, Houron 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 and minority adolescents. Health Psychol. 2011;30:463.

103. Dudley DA, Okely AD, Pearson P, Peat J. Engaging adolescent girls from overweight and obese families: A Feasibility Study. JIMR Mhealth UHealth 2015;54:697. doi:10.20912/2015/mhealth-4605. PMC:PMC4704914.

104. Eather N, 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;43:590–100.

105. Eather N, Morgan PJ, Lubans DR. Feasibility and preliminary efficacy of the 4FitFun intervention for improving physical fitness in a sample of primary school children: a pilot study. Phys Educ Sport Pedagog. 2013;18:389–411.

106. Ebbinghaus CB, Feldman HA, Chomitz VR, Antontelli TA, Gortmaker SL, Ogartanian SK, Ludwig DS. A randomized trial of sugar-sweetened beverages and adolescent body weight. N Engl J Med. 2012;367:1407–16.
Klesges RC, Obarzanek E, Kumanyika S, Murray DM, Klesges LM, Relyea GE, Kipping RR, Jago R, Lawlor DA. Diet outcomes of a pilot school-based school-based obesity prevention programme (HeLP) to prevent obesity in UK primary-school children: a cluster randomised controlled trial. Lancet Child Adolesc Health. 2018;2:35–45.

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.

Maddison R, Marsh S, Foley L, Epstein LH, Olds T, Dews O, Heke I, Carter K, Jiang Y, Ni Mhruchu C. Screen-time weight-loss intervention targeting children at home (SWITCH): a randomised controlled trial. Int J Behav Nutr Phys Act. 2014;11:111.

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.

Madsen KA, Thompson HR, Wlasuk L, Quelizia E, Schmidt C, Newman TB. After-school program to reduce obesity in minority children: a pilot study. J Child Health Care. 2009;13:16–22.

McAuliffe N, Fahlin 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.

Ni Mhruchu 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.

Neumark-Sztainer D, Story M, Hannan P, Res JX. New moves: a school-based obesity prevention program for adolescent girls. Prev Med. 2003;37:471–51.

Okely AD, Lubans DR, Morgan PJ, Cotton W, Peralta L, Miller J, Batterham M, Janssen I. Promoting physical activity among adolescent girls: the girls in sport group randomized trial. Int J Behav Nutr Phys Act. 2017;14:81.

Patrick K, Caffes 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.

Patrick K, Sallis JF, Prochaska JJ, Lydon-Dell DF, Caffes 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.

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.

Paul IM, Savage JS, Anzman-Frasca S, Marini ME, Beiler JS, Hess LB, Loken E, Birch L. 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.

Pereira LA, Kelly L, Montgomery KD, Hamilton A, Fisher A, McCall JH, Conte RL, Paton JY, Grant S. Physical activity to prevent obesity in young children: cluster randomised controlled trial. BMJ. 2006;333:1041.

Reilly JJ, McDowell ZC. Physical activity interventions in the prevention and treatment of paediatric obesity: systematic review and critical appraisal. Proc Nutr Soc. 2003;62:619–9.

Robbins LB, Ling J, Sharma DB, Dalimonte-Merculino 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.

Robbins LB, Pfeiffer KA, Maier KS, Lo Y-J, Weslek 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.

Robertson W, Fleming J, Kamal A, Hamborg T, Khan KA, Griffiths F, Stewart-Brown S, Stallard N, Petrou S, Simskis 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.

Robertson W, Friede T, Blissert J, Rudolf MC, Walls MA, Stewart-Brown S. Pilot of Families for Health’, a community-based family intervention for obesity. Arch Dis Child. 2009;93:921–6.

Robinson TN, Killen JD, Kraemer HC, Wilson DM, Matheson DM, Haskell WL, Pratt 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. Etnh Dis. 2003;13:51–6.
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, Rabbin R, Chanone J-P, Milin 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, Dztura J, Shaw M, Serrecchia JB, Barbetta G, Rose P, Lavetis 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, Chanone 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, Dztura 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. Beets MW, Glenn Weaver R, Turner-McGrievy G, Saunders RP, Webster CA, Brazendale K, Chandler J. Evaluation of a statewide dissemination and implementation of physical activity intervention in after-school 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.