Received: 22 June 2021 | Revised: 2 August 2021 | Accepted: 18 August 2021

DOI: 10.1111/obr.13369

REVIEW

OBESITY Reviews WILEY

Impact of risk of generalizability biases in adult obesity interventions: A meta-epidemiological review and meta-analysis

Michael W. Beets[1] | Lauren von Klinggraeff[1] | Sarah Burkart[1] | Alexis Jones[1] | John P. A. Ioannidis[2] | R. Glenn Weaver[1] | Anthony D. Okely[3] | David Lubans[4] | Esther van Sluijs[5] | Russell Jago[6] | Gabrielle Turner-McGrievy[7] | James Thrasher[7] | Xiaoming Li[7]

[1]Department of Exercise Science, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, USA

[2]Departments of Medicine, of Epidemiology and Population Health, of Biomedical Data Science, and of Statistics, and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA

[3]Faculty of the Arts, Social Sciences and Humanities, School of Education, University of Wollongong, Wollongong, New South Wales, Australia

[4]School of Education, Priority Research Centre in Physical Activity and Nutrition, University of Newcastle, Callaghan, New South Wales, Australia

[5]Centre for Diet and Activity Research (CEDAR), MRC Epidemiology Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK

[6]Centre for Exercise Nutrition and Health Sciences, School for Policy Studies, University of Bristol, Bristol, UK

[7]Department of Health Promotion, Education, and Behavior Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, USA

Correspondence
Michael W. Beets, Department of Exercise Science, Arnold School of Public Health, University of South Carolina, 921 Assembly Street, Columbia, SC 29203, USA.
Email: beets@mailbox.sc.edu

Funding information
National Health and Medical Research Council, Grant/Award Numbers: APP1154507, APP1176858; National Heart, Lung, and Blood Institute, Grant/Award Numbers: F31HL158016, F32HL154530, R01HL149141; National Institute of General Medical Sciences, Grant/Award Number: P20GM130420; Sue and Bob O'Donnell

Summary

Biases introduced in early-stage studies can lead to inflated early discoveries. The risk of generalizability biases (RGBs) identifies key features of feasibility studies that, when present, lead to reduced impact in a larger trial. This meta-study examined the influence of RGBs in adult obesity interventions. Behavioral interventions with a published feasibility study and a larger scale trial of the same intervention (e.g., pairs) were identified. Each pair was coded for the presence of RGBs. Quantitative outcomes were extracted. Multilevel meta-regression models were used to examine the impact of RGBs on the difference in the effect size (ES, standardized mean difference) from pilot to larger scale trial. A total of 114 pairs, representing 230 studies, were identified. Overall, 75% of the pairs had at least one RGB present. The four most prevalent RGBs were duration (33%), delivery agent (30%), implementation support (23%), and target audience (22%) bias. The largest reductions in the ES were observed in pairs where an RGB was present in the pilot and removed in the larger scale trial (average reduction ES −0.41, range −1.06 to 0.01), compared with pairs without an RGB (average reduction ES −0.15, range −0.18 to −0.14). Eliminating RGBs during early-stage testing may result in improved evidence.

KEYWORDS
intervention, pilot, scaling, translation

Abbreviations: CMA, comprehensive meta-analysis; ES, effect size; MeSH, Medical Subject Headings; RGB, risk of generalizability bias.

1 | INTRODUCTION

In the United States alone, investing in clinical trials is a multibillion-dollar enterprise. For behavioral interventions, it is common to perform one or more early-stage studies before launching larger scale trials.[1,2] These early-stage studies (referred to herein as pilot/feasibility studies) lay the foundation for more definitive hypothesis testing in larger scale clinical trials by providing preliminary evidence on the effects of an intervention and evidence for the feasibility of trial (e.g., recruitment) and intervention (e.g., fidelity) related facets.[3] Pilot/feasibility studies are consistently depicted as playing key roles in translational science frameworks for developing behavioral interventions by providing evidence about whether an intervention can be done and demonstrates initial promise.[1,2,4,5] Well-designed and executed pilot/feasibility studies, thus, serve as the cornerstone for decisions regarding the execution of larger scale clinical trials, and funding for larger scale trials often requires these pilot studies (e.g., R01 grants from US National Institutes of Health [NIH]).

A challenge many researchers face in the design and execution of a pilot/feasibility study is the ability to translate initially promising findings of an intervention into an intervention that demonstrates efficaciousness when evaluated in a larger scale trial. The transition from early-stage, often small, pilot/feasibility studies to progressively larger trials is commonly associated with a respective drop in the impact of an intervention that can render an intervention tested in the larger trial completely inert.[6–8] Referred to as a scaling penalty[9,10] in the dissemination/implementation literature, a similar phenomenon is observed in the sequence of studies from pilot/feasibility to progressively larger size trials.[11]

This initial promise followed by reduced effectiveness may be from the introduction of biases during the early-stage of testing that lead to exaggerated early effects. Biases, in the scientific literature on clinical trials, are typically thought of in relation to internal validity, which focus on issues resulting from randomization and blinding procedures, incomplete data and selective reporting of outcomes.[12,13] Although important, internal validity issues do not address other contextual factors that may lead to overestimation or underestimation of effects, especially in behavioral interventions. In the behavioral science field, a newly developed set of biases have been conceptualized that address contextual factors in behavioral interventions that, when present, could lead to inflated effects.[11] Referred to as risk of generalizability biases (RGBs), these focus on contextual factors associated with external validity or the degree to which features of the intervention and sample in the early-stage pilot/feasibility study are not scalable or generalizable to the next stage of testing in a larger, more well-powered trial. RGBs focus broadly on the conduct of behavioral interventions by including items related to where an intervention was delivered, by whom and to whom an intervention was delivered, and other support necessary to deliver the intervention. The RGBs focus on changes in these from the early-stage pilot/feasibility studies to the larger scale trial and how such changes can potentially lead to diminished effects in the large-scale trial. The RGBs, therefore, represent contextual features that any number of behavioral interventions, regardless of theoretical and methodological approach or behavior targeted, encounter in the design and execution of an intervention.

An initial study[11] provided preliminary support for the impact of RGBs in childhood obesity trials. This study showed larger scale trials that were informed by pilot/feasibility trials with an RGB had substantially greater reductions in outcomes in comparison with larger scale trials informed by pilot/feasibility studies without an RGB. This, however, was demonstrated in a relatively small number of interventions (total of 39) that had a published pilot/feasibility study and a published larger scale trial on a topic related to childhood obesity. Given pilot/feasibility studies provide evidence to inform decisions about investing in larger scale trials, they should be conducted without the introduction of biases. We believe the RGBs represent a unique set of biases behavioral interventionist face and have the potential to inform important aspects in the design and execution of pilot/feasibility studies. The purpose of this study was to build upon previous evidence of the influence of RGBs and evaluate their impact in a sample of published pilot/feasibility studies and larger scale trials of the same behavioral intervention on a topic related to adult obesity.

2 | METHODS

The methods are similar to the methods used in a previous meta-epidemiological investigation of RBGs in trials of childhood obesity.[11] Specifically, comprehensive, meta-epidemiological review procedures were used to identify behavioral interventions focused on adult obesity (age range ≥18 years) that have a published preliminary, early-stage testing of the intervention and a published larger scale trial of the same intervention. Consistent with our prior work, behavioral interventions were defined as social science/public health intervention involving a coordinated set of activities targeted at one or more levels including interpersonal, intrapersonal, policy, community, macro-environments, micro-environments, and institutions[1,14–16] and obesity-related topics could include diet, exercise, physical activity, screen time, sleep, sedentary behavior, or combination of these behaviors. "Behavioral intervention pilot studies" were defined as studies which test the feasibility of a behavioral intervention and/or provide evidence of a preliminary effect(s) in a hypothesized direction.[1,17,18] These studies are conducted separately from, and prior to, larger scale trials, with the results used to inform the subsequent testing of the same or refined intervention.[1,18] Behavioral intervention pilot studies can also be referred to as "feasibility," "preliminary," "proof-of-concept," "vanguard," "novel," or "evidentiary".[1,19,20]

2.1 | Data sources and search strategy

To identify pilot/feasibility studies and larger scale trials of behavioral interventions on a topic related to adult obesity the following

procedures were used. First, a combination of controlled vocabulary terms (e.g., MeSH and Emtree), free-text terms, and Boolean operators were used to identify eligible reviews and meta-analysis across OVID Medline/PubMed; Embase/Elsevier; EBSCOhost; and Web of Science. Searches for meta-analyses and reviews allowed for the identification of a large number of behavioral intervention studies in a time effective manner and is a primary mechanism of study identification in meta-epidemiological studies.[21] Each search contained one or more of the following terms for participant age—adult (i.e., 18 years and older)—and one of the follow terms related to obesity—obesity, weight, physical activity, diet, screen, sleep, sedentary, exercise, and study design—systematic review or meta-analysis of behavioral interventions. A detailed record of the search strategy is provided in the Data S1.

All identified systematic reviews and/or meta-analysis were uploaded in an EndNote Library (v. X9.2). Each resulting title/abstract of the systematic reviews and meta-analysis were screened by at least two independent reviewers (LV, SB, AJ) in Covidence (Covidence.org, Melbourne, Australia) prior to full-text review. All articles included with the systematic reviews and/or meta-analyses were retrieved and uploaded into an NVivo (v.12, Doncaster, Australia) file for text-mining. Articles from the reviews were text-mined using text search query to identify them as either (1) self-identified preliminary testing of an intervention or (2) larger scale trial referring to prior preliminary work. This was done using terms such as "pilot, feasible, preliminary, protocol previously, rationale, elsewhere described, prior work" to flag sections of text. Once flagged, each section of text was reviewed to determine whether it met inclusion criteria and then tagged as either a larger scale or smaller-scale study. After being identified and tagged as a larger scale trial, a "follow back" approach was used to identify references to preliminary testing of an intervention within the body of the article (Figure 1, Steps 4.1–4.2). Where larger scale trials indicated previous published pilot/feasibility testing of the intervention, the referenced article was retrieved and reviewed to determine if it met the definition of pilot/feasibility study. For studies self-identified as pilot/feasibility, studies were "follow forward" using the Web of Science Reference Search interface to identify any subsequent published study referencing the identified pilot/feasibility study as preliminary work (Figure 1, Steps 5.1–5.2; L. V. and S. B.). Successfully paired pilot/feasibility studies and larger scale trials identified in both the forward and backward approaches were catalogued (Figure 1, Steps 4.3 and 5.3), and narrative and analytic information was extracted (A. J., S. B., and L. V.) prior to meta-analytic modeling.

2.2 | Inclusion criteria

To be included, studies were required to study adults 18 years of age or older participating in a behavioral intervention on a topic related to obesity. Pairs of studies were included if they had published pilot/feasibility study and a larger scale trial of the same or similar intervention and were published in English.

2.3 | Exclusion criteria

Pairs were excluded if either the pilot study reported only outcomes associated with compliance to an intervention (e.g., feasibility metrics, attendance, and adherence, $N = 13$) and no measures on primary or secondary outcomes, or the published pilot study or larger scale trial provided point estimates for the outcomes but did not provide a measure of variance (e.g., SD, SE, 95% CI, $N = 15$).

2.4 | Data management procedures

2.4.1 | Coding RGBs

The RGBs were coded in each pilot and larger scale trial pair according to previously established criteria[11] and are defined in Table 1. Within each pair of pilot/feasibility study and larger scale trial, two reviewers independently reviewed the entire article to identify the presence or absence of one or more RGBs. For the purpose of this review, eight of the nine originally defined RGBs were extracted for analyses, with the

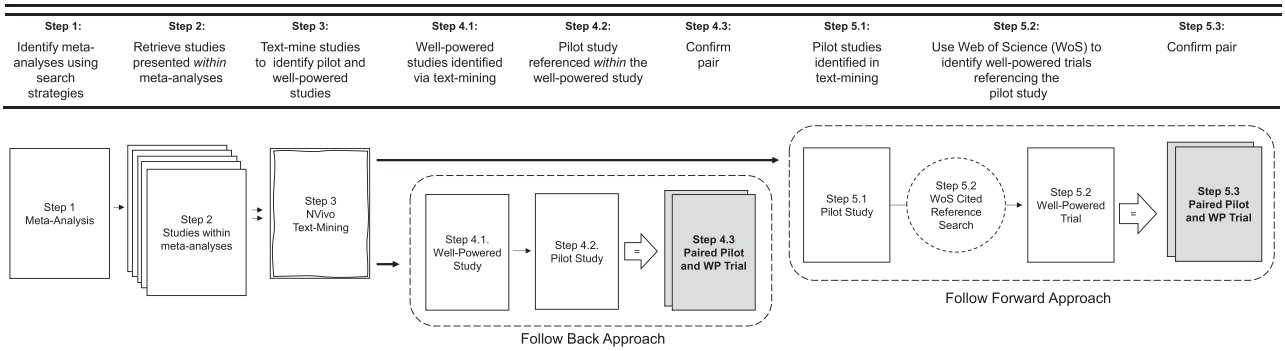

FIGURE 1 Diagram of search procedures of locating pilot studies and larger scale trial pairs

TABLE 1 Operational definition of risk of generalizability biases

Risk of generalizability bias	What is the potential for difference(s) between …	Example of bias
Intervention intensity bias	… the number and length of contacts in the pilot study compared with the number and length of contacts in the larger scale trial of the intervention?	7 sessions in 7 weeks in pilot/feasibility study vs. 4 sessions over 12 weeks in larger scale trial. 24 contacts per week for 12 weeks in pilot/feasibility vs. 2 contacts per week for 12 weeks in larger scale trial
Implementation support bias	… the amount of support provided to implement the intervention in the pilot study compared with the amount of support provided in the larger scale trial?	Any adherence issues noted were immediately addressed in ongoing supervision. At the end of each session, the researcher debriefed with the interventionist to discuss reasons for the variation in approach and to maintain standardization, integrity of implementation, and reliability among interventionists.
Intervention delivery agent bias	… the level of expertise of the individual(s) who delivered the intervention in the pilot study compared with who delivered the intervention in the larger scale trial	All intervention sessions were led by the first author. The interventionists were highly trained doctoral students or a postdoc.
Target audience bias	… the demographics of those who received the intervention in the pilot study compared with those who received the intervention in the larger scale trial	Affluent and educated background, mostly White non-Hispanic. Participants were predominately healthy and well-educated.
Intervention duration bias	… the length of the intervention provided in the pilot study compared with the length of the intervention in larger scale trial?	8-week intervention in pilot/feasibility study to 12-month intervention in larger scale trial.
Setting bias	… the type of setting where the intervention was delivered in the pilot study compared with the setting in the larger scale trial	A convenience sample of physicians, in one primary care office practice agreed to participate. They were approached because of a personal relationship with one of the investigators who also was responsible for providing physician training in the counseling intervention. The study was conducted at a university health science center.
Measurement bias	… the measures employed in the pilot study compared with the measures used in the larger scale trial of the intervention for primary/secondary outcomes?	Use of objective measures in pilot to self-report measures in larger scale trial.
Directional conclusions	Are the intervention effect(s) in the hypothesized direction for the pilot study compared with those in the larger scale trial?	Outcomes in the opposite direction (e.g., control group improved more so than treatment group).

Note: Based on definitions originally appearing in Beets et al.[11]

RGB outcome bias not coded because no analytical comparison for outcomes between a pilot and larger scale trial could be made on this RGB. RGBs were established by comparing the description provided in the pilot/feasibility study regarding the intensity of the intervention, the amount of support to implement, who delivered the intervention, to whom the intervention was delivered, the duration of the intervention, the locale of where the intervention was delivered, the types of measurements used to collect outcomes, and the direction of the findings. For example, a pilot/feasibility study may indicate all intervention sessions were led by the first author, whereas in the larger scale trial, the intervention was delivered by community health workers. In this instance, the pair would be coded as having the risk of delivery agent bias present in the pilot/feasibility study and not in the larger scale trial. Where discrepancies were encountered or clarifications were required, a third reviewer was brought in to assist in the final coding.

2.4.2 | Meta-analytical procedures

Standardized difference of means (SDM) effect sizes were calculated for each study across all reported outcomes. The steps outlined by Morris and DeShon[22] were used to create effect size estimates from studies using different designs across different interventions (independent groups pretest/posttest; repeated measures single group pretest/posttest) into a common metric. For each study, individual effect sizes and corresponding 95% confidence intervals (CI) were calculated for all outcome measures reported in the studies.

To ensure comparisons between pilot and larger scale pairs were based upon similar outcomes, we classified the outcomes reported across pairs (i.e., pilot and larger scale trial) into seven categories that represented all the data reported.[23] These were measures of body composition (e.g., body mass index [BMI], percent body fat, skinfolds), physical activity (e.g., moderate-to-vigorous physical activity), sedentary behaviors (e.g., TV viewing and sitting), psychosocial (e.g., self-efficacy and social support), diet (e.g., kcals and fruit/vegetable intake), physiological (e.g., high-density lipoprotein [HLD], low-density lipoprotein [LDL], and glucose), or sleep (e.g., duration, onset, and offset). Only outcomes within common categories represented across both the pilot and the larger scale trial were included in analyses. For instance, a study could have reported data related to body composition, physiological, and physical activity in both the pilot and larger scale trial, but also reported sedentary outcomes for the pilot only and psychosocial related outcomes for the larger scale only. In this scenario, only the body composition, physiological, and physical activity variables would be compared across the two studies within the pair. For studies that reported multiple outcomes within a given category, all outcomes were extracted, and the shared correlations among outcomes from the same trial were accounted for in the analytical models (see below for details).

All individual outcome measures reported within categories across pairs were extracted and entered into an Excel file. Once a pair was completely extracted and reported data transformed (e.g., standard errors transformed into standard deviations), data were transferred into Comprehensive Meta-Analysis (CMA) software (v3.3.07) to calculate a SDM for effects reported in a study. After all outcomes across all pairs were entered into CMA, the complete data file was exported as a comma separated file and uploaded into the R environment[24] for final data analyses to occur.

Consistent with our previous study[11] 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 change in the SDM from the pilot study to the larger scale trial for studies coded with and without a given bias present. All studies reported more than one outcome effect; therefore, summary effect sizes were calculated using a random-effects multilevel robust variance estimation meta-regression model,[25–27] with outcomes nested within studies nested within pairs in the R environment[24] using the package *metafor*.[27] This modeling procedure is distribution free and can handle the non-independence of the effects sizes from multiple outcomes reported within each of the seven categories reported within a single study. The difference in the SDM from the pilot and larger scale trial were quantified according to previously defined formulas for the scale-up penalty,[9,10] which is calculated as follows: the SDM of the larger scale trial was divided by the SDM of the pilot and then multiplied by 100. A value of 100% indicates identical SDMs in both the pilot and larger scale trial. A value of 50% indicates that the larger scale trial was half as effective as the pilot study; a value above 100% indicates the larger scale trial is more effective than the pilot, whereas a negative value indicating that the direction of the effect in the larger scale trial is opposite that of the pilot.

A secondary evaluation of the impact of the biases was performed by examining the presence/absence of the biases on the occurrence of a nominally statistically significant (i.e., $p \leq 0.05$) outcome in the larger scale trials. These analyses were restricted to the p values for the individual outcomes in the larger scale trials, only. p values for each individual effect size were estimated within the Comprehensive Meta-Analysis software based upon the effect size and its associated standard error. p values from publications were not used because they were either not reported or reported as truncated values (e.g., $p < 0.01$). p values were dichotomized as $p > 0.05$ and $p \leq 0.05$ based on conventional behavioral intervention studies. Logistic regression models, using robust variance estimators, were used to examine the odds of a nominally statistically significant outcome in the larger scale trial based on the presence/absence of the biases. Across all models, we controlled for the influence of all other biases. Because it has been recently proposed that statistical significance thresholds should become $p < 0.005$ rather than $p < 0.05$,[28,29] we also explored these analyses using the $p < 0.005$ threshold.[28,29]

3 | RESULTS

A PRISMA diagram for the literature search is presented in Figure 2.

From the 114 pairs, a total of 1160 effects were extracted from the pilot studies (average 20 [*SD* ± 14] effects per study) and 1089 effects extracted from the larger scale trials (average 16 [*SD* ± 9] effects per trial). Studies were published between 1993 through 2020. Overall, the most commonly reported outcomes were body composition (54% of studies), physiological (51%), physical activity (50%), and psychosocial (29%), and diet (22%). The median sample size of the pilot studies was 44 (range 8 to 770, average 74) while the median sample size of the larger scale trials was 201 (range 45 to 5801, average 361). A total of 71% of the pilot studies utilized a randomized design while 92% of the larger scale trials utilized this design with the remaining 8% using either a single group pre/post design or two group non-randomized design.

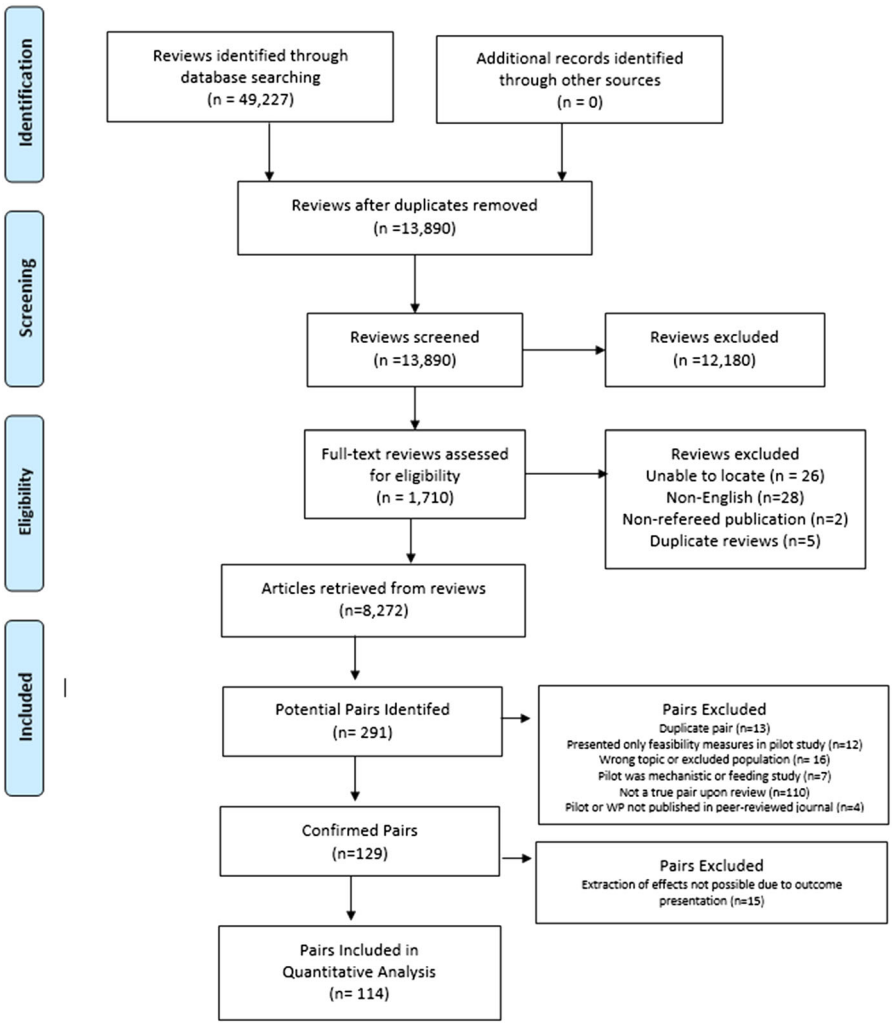

FIGURE 2 PRISMA diagram of systematic literature search and final studies included in analyses

The prevalence of the RGBs across the 114 pairs is reported in Figure 3. Overall, 25% of the pairs were coded as no presence of any biases, 30% containing 1 bias, 33% containing 2 biases, and 12% with 3 or 4 biases. The four most prevalent RGBs were duration (33%), delivery agent (30%), implementation support (23%), and target audience (22%) bias.

The impact of the RGBs on trial-related outcomes are presented in Figure 4. For pairs where the pilot was coded as having an RGB present and in the larger scale trial, the RGB was no longer present, the SDM decreased by an average of ΔSDM −0.41, range −1.06 to 0.01. The largest reductions in the SDM were observed for implementation support (pairs = 5, ΔSDM −1.06, 95% CI [−2.26, 0.132]) and setting (pairs = 4, ΔSDM −1.01, 95% CI [−1.56, −0.46]), followed by target audience (pairs = 13, ΔSDM −0.42, 95% CI [−0.62, −0.22]), intervention intensity (pairs = 3, ΔSDM −0.28, 95% CI [−0.59, 0.04]), and delivery agent (pairs = 15, ΔSDM −0.25, 95% CI [−0.43, −0.06]) biases in comparison with pairs where these biases were not coded as present in the pilot or larger scale trial.

Four of the RGBs were coded as present in the pilot and larger scale trial: delivery agent (pairs = 19), implementation support (pairs = 21), target audience (pairs = 12), and setting (pairs = 7). Three of these four biases were associated with a smaller reduction in the SDM in comparison with pairs without the biases present. Implementation support and setting biases were associated with a reduction of −0.09 (−0.23 to 0.04) and −0.10 (−0.37 to 0.17), respectively, whereas target audience bias was associated with an increased effect in the larger scale trial (+0.10, −0.05 to 0.25). The presence of intervention duration, directional conclusions, and measurement biases were not associated with a larger reduction in the SDM compared with pairs without these biases present.

The scale-up penalties associated with the RGBs are presented in Figure 4. Overall, on average, pairs with biases present in the pilot and removed in the larger scale trial's effects were 33% (range −10% to 104%) of those reported in the pilot/feasibility study. This is compared with 63% (range 55% to 65%) for pairs without biases in either the pilot or larger scale trial. Further, for pairs where a bias was present in both the pilot and the larger scale trial, the larger scale trial effect was 86% (range 60% to 136%) of the effect observed in the pilot/feasibility study.

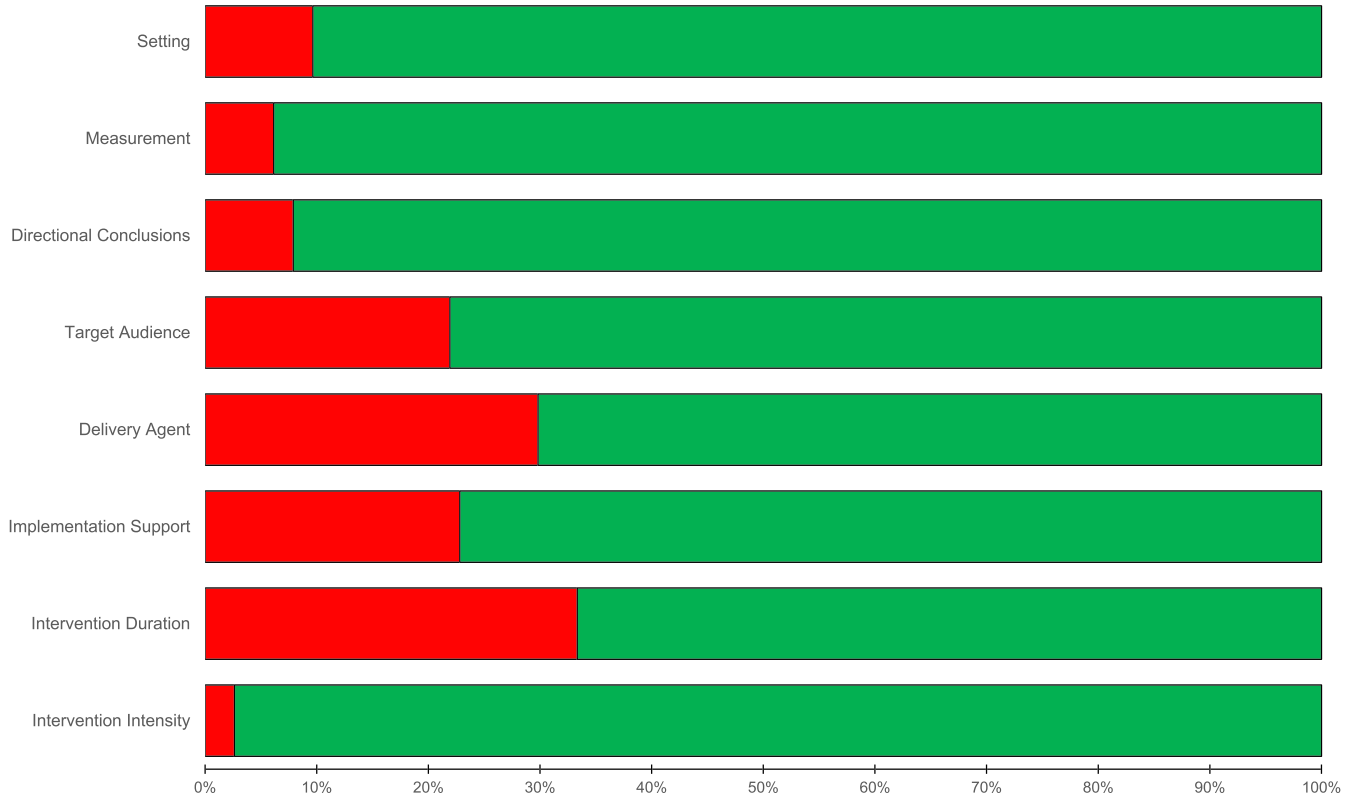

FIGURE 3 Classification of the presence (red circle) and absence (green circle) risk of generalizability biases across pilot and larger scale trial pairs

The impact of the RGBs on the probability of a statistically significant outcome are presented in Table 2, and the distribution of the *z* value is presented in Figure 5. Pairs coded as having either delivery agent, setting, intervention intensity, or directional conclusions biases in the pilot study and *not* in the larger scale trial exhibited a reduced odds of detecting a nominally statistically significant effect of $p < 0.05$ in the larger scale trials compared with pairs without an RGB present. Two (target audience and setting biases) of the four RGBs were coded as present in the pilot *and* larger scale trial; these larger scale trials were more likely to report a nominally statistically significant effect of $p < 0.05$ in comparison with pairs without the RGB present. For detecting a statistical effect at $p < 0.005$, pairs coded as having either delivery agent, setting, duration, or intensity bias in the pilot study and *not* in the larger scale trial exhibited a reduced odds of detecting a nominally statistically significant effect of $p < 0.005$ in the larger scale trials compared with pairs without an RGB present. Consistent with the analyses for $p < 0.05$, two (target audience and setting biases) of the four RGBs that were coded as present in the pilot *and* larger scale trial, these larger scale trials were more likely to report a statistically significant effect of $p < 0.005$ in comparison with pairs without the RGB present.

4 | DISCUSSION

Informative, early-stage pilot/feasibility studies may provide valuable information about whether an intervention is ready to be tested in a larger scale trial. When early-stage studies include features that result in incorrect conclusions about an intervention's viability, this can lead to premature scale-up and, ultimately, intervention failure when evaluated at scale. The purpose of this meta-epidemiological review was to examine the influence of a set of recently catalogued biases—the RGBs—that, when present in pilot/feasibility studies, can lead to reduced effectiveness in larger scale trials. As in a prior study focused on childhood obesity interventions,[11] the presence of an RGB in a pilot/feasibility study tended to be associated with reduced effects in the larger scale intervention and a reduced probability of detecting a nominally statistically significant effect. Further, the prevalence of the RGBs across the 114 pairs of published pilots and larger scale trials of the same intervention was high, with three out of four pairs containing at least one RGB. The most impactful RGBs were delivery agent, implementation support, target audience, setting, and intervention intensity bias, although data were limited for each RGB to allow a reliable ranking of the magnitude of these biases. These findings on the prevalence and impact of RGBs on the effectiveness of larger

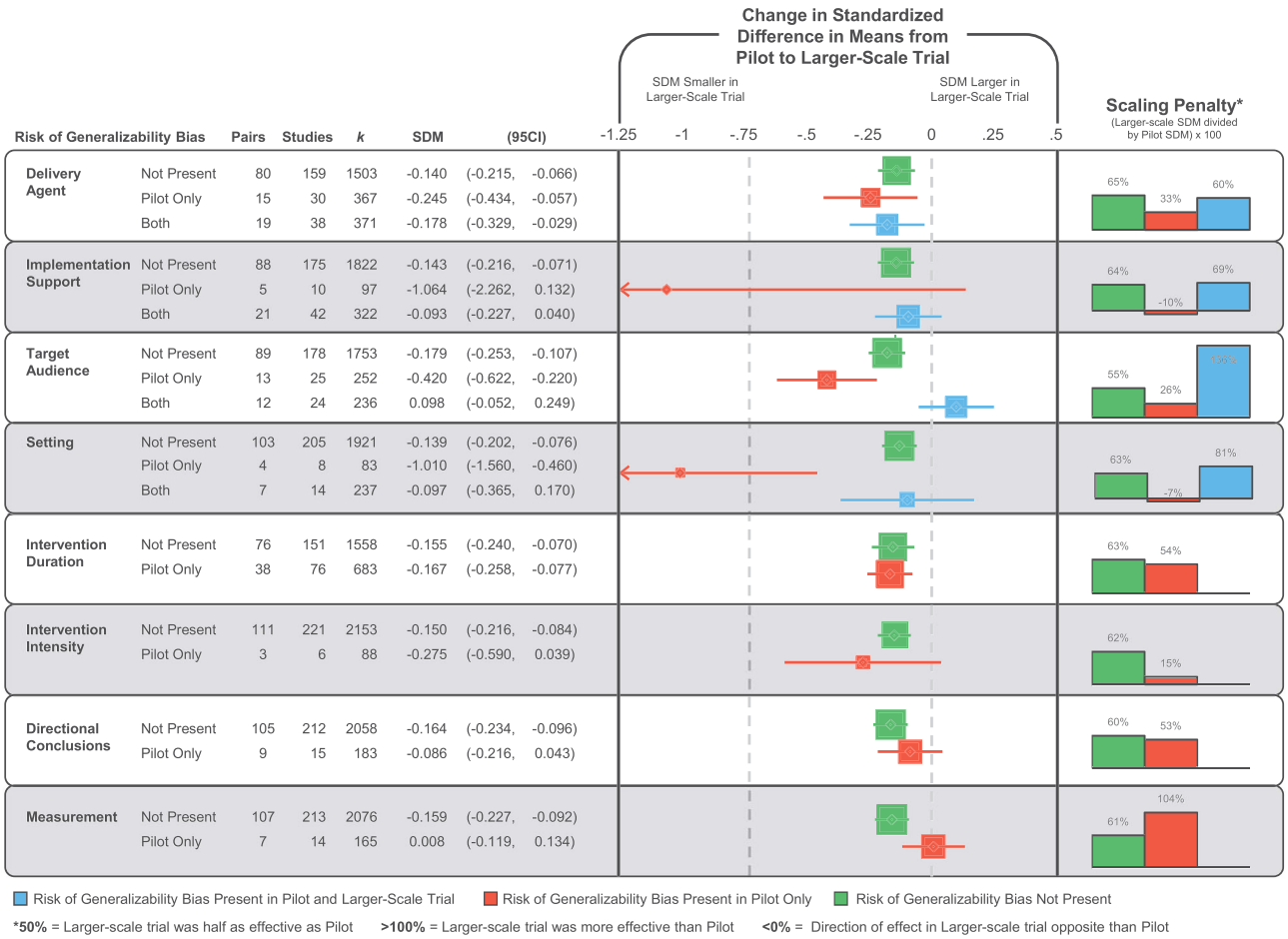

FIGURE 4 Forest plot of the change in the standardized difference in means (SDM) of the presence, absence, or carry forward of risk of generalizability biases from a pilot/feasibility study to a larger scale trial. No pairs contained directional conclusion bias in both the pilot and larger scale trial. Intervention duration, intervention intensity, and measurement describe differences between smaller and larger scale studies, so they cannot be present in both studies

scale trials have implications for the behavioral intervention field because early-stage studies with RGBs appear to produce misleading results about the readiness of an intervention for scaling.

When moving from smaller, early-stage studies to trials to progressively larger sample sizes, it is not uncommon nor unexpected to see a drop in the respective impact (i.e., voltage) of an intervention.[8–10] Our findings of reduced effects in the larger scale trials compared with the effects from the pilot/feasibility studies match this pattern. This pattern, however, becomes accentuated when RGBs are considered. Pairs where an RGB was present in the pilot/feasibility study (i.e., delivery agent, implementation support, target audience, setting, and intervention intensity) and not present in the larger scale trial resulted in a greater scaling penalty, compared with pairs where an RGB was not present. Conversely, two of the four RGBs (setting and target audience) that could be carried forward (i.e., present in both the pilot/feasibility and larger scale trial) showed less of a scaling penalty or demonstrated a greater effect in the larger scale trial. These findings, both quantitatively and conceptually, fit the patterns expected with the presence of an RGB and provide evidence of their impact on the outcomes reported in larger scale trials.

Comparing effect sizes generated from pilot/feasibility studies to their larger scale trial is not without limitations, given the lower precision attributed with effect size estimations from pilot/feasibility studies.[17,30–32] Thus, we recognize that effect sizes in pilot studies are estimated with large uncertainty; therefore, putting too much trust on point estimates may be misleading. To address this, we conducted analyses considering only the larger scale trial outcomes and the probability of detecting nominally statistically significant effects ($p < 0.05$ and $p < 0.005$). This approach eliminated the issues associated with using effects from the pilot/feasibility studies, instead relying solely upon those effects presented in the larger scale trial. Again, findings demonstrated the presence of RGBs have an impact on the statistical significance in larger scale trials. When larger scale trials without an RGB are informed by a pilot/feasibility study with an RGB, the large-scale study had a lower probability of detecting a nominally statistically significant effect compared with larger scale trials without an

TABLE 2 Odds for detecting a nominally statistically significant p value ($p < 0.05$ and $p < 0.005$) in a larger scale trial from the presence of a risk of generalizability bias

		Odds ratio for $p < 0.05$ in larger scale trial			Odds ratio for $p < 0.005$ in larger scale trial		
Risk of generalizability bias		**OR**	**(95% CI)**	**% Effects**	**OR**	**(95% CI)**	**% Effects**
Delivery agent	Not present	Reference		34%	Reference		23%
	Pilot only	**0.62**	**(0.43, 0.90)**	24%	**0.60**	**(0.39, 0.93)**	15%
	Both	1.39	(0.97, 2.00)	42%	0.96	(0.63, 1.47)	22%
Implementation support	Not present	Reference		34%	Reference		22%
	Pilot only	0.81	(0.42, 1.56)	30%	0.54	(0.22, 1.30)	13%
	Both	0.74	(0.51, 1.08)	30%	1.03	(0.67, 1.57)	22%
Target audience	Not present	Reference		32%	Reference		20%
	Pilot only	0.72	(0.47, 1.11)	26%	1.11	(0.70, 1.75)	21%
	Both	**2.05**	**(1.35, 3.12)**	49%	**2.51**	**(1.58, 3.98)**	34%
Setting	Not present	Reference		33%	Reference		21%
	Pilot only	**0.17**	**(0.06, 0.47)**	8%	**0.22**	**(0.07, 0.70)**	6%
	Both	**2.01**	**(1.29, 3.14)**	48%	**2.49**	**(1.54, 4.03)**	35%
Intervention duration	Not present	Reference		34%	Reference		23%
	Pilot only	0.83	(0.61, 1.12)	32%	**0.60**	**(0.42, 0.84)**	17%
Intervention intensity	Not present	Reference		34%	Reference		22%
	Pilot only	**0.06**	**(0.01, 0.46)**	3%	**0.12**	**(0.02, 0.90)**	3%
Measurement	Not present	Reference		35%	Reference		22%
	Pilot only	0.86	(0.52, 1.41)	15%	0.93	(0.52, 1.65)	20%
Directional conclusions	Not present	Reference		34%	Reference		23%
	Pilot only	**0.27**	**(0.13, 0.52)**	32%	**0.25**	**(0.10, 0.60)**	8%

Note: Bolded values 95% confidence intervals (CIs) do not cross 1.00. Reference group is pilot/feasibility study and larger scale trial pairs without a risk of generalizability bias present.

RGB in either the pilot/feasibility study or larger scale trial. An example of this is delivering an intervention in a university setting during the pilot/feasibility study versus delivering the intervention in a community-based setting in the larger scale trial (setting bias). A potential reduction in the odds of detecting a nominally statistically significant effect is observed across most of these comparisons and provides further support for the impact of RGBs on larger scale trial outcomes, above and beyond any concern about comparing effect sizes between pilot/feasibility studies and larger scale trials.

We recognize that the list of RGBs evaluated herein may not fully capture the entirety of mechanisms that lead to successful or unsuccessful larger scale trials. Recent studies[4,5,33–35] describe a number of mechanisms linked to either the successful scale-up of behavioral interventions or that should be considered in the conduct of early-stage implementation studies of behavioral interventions. These include mechanisms associated with the intervention (e.g., credibility, relevance, compatibility), organization (e.g., perceived need for intervention), environment (e.g., policy context and bureaucracy), resource team (e.g., effective leadership), scale up strategy (e.g., advocacy strategies), and planning/management (e.g., strategic monitoring). These are clearly critical mechanisms associated with developing a behavioral intervention that survives the scaling process. The current list of RGBs can have important implications and potential interactions with studies evaluating these scaling mechanisms. For instance, the risk of setting bias, target audience bias, delivery agent bias, and intervention duration bias have the potential to influence assessments related to whether an environment can support the intervention, an interventions' adoption, adherence and dose received, and evaluations of whether the intervention is compatible and relevant. Although the list of RGBs, at this time, is likely incomplete, we believe the current list identifies features of a study that transcend the content of the intervention and focus on features of how it is conducted that can lead to a lower likelihood of success in a larger scale trial.

The introduction of one or more RGBs in pilot studies could be due to a lack of reporting and/or procedural guidelines for pilot studies that focus on topics related to RGBs. Recently, an extension to the CONSORT statement was developed for pilot/feasibility studies.[36] This statement focuses predominately on features of the research design and conduct associated with *internal* validity and does not provide guidance for factors affiliated with *external* validity, such as the RGBs examined herein. Other reporting guidelines, for instance PRESCI-2[37,38] and TiDieR,[14] incorporate elements of the RGBs and recommend they be detailed in scientific publications (e.g., clear description of who delivered the intervention) but do not provide a

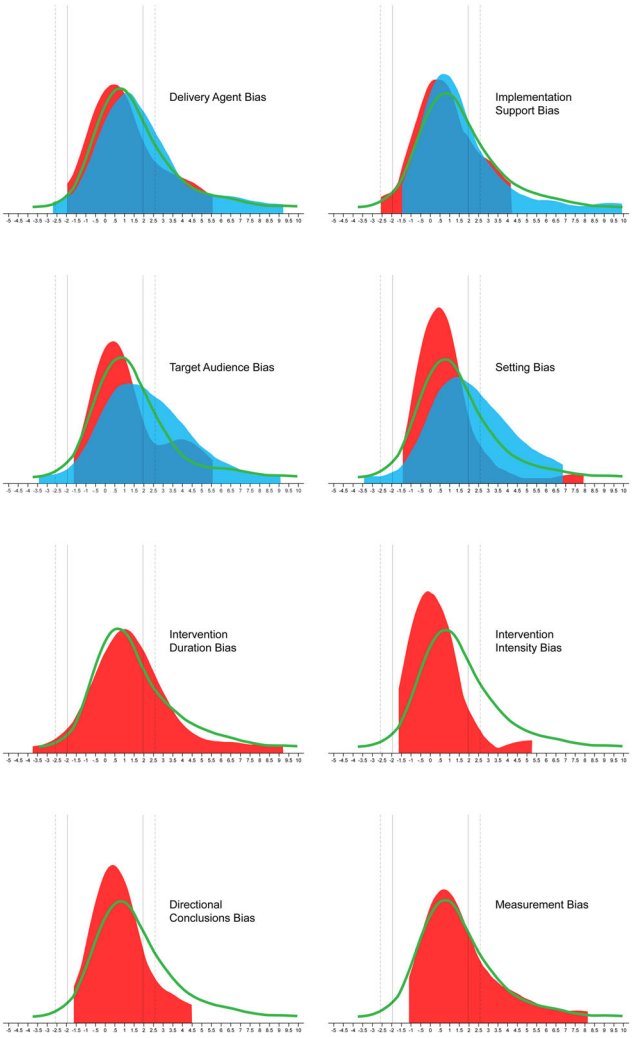

FIGURE 5 *Z* value distribution of outcomes in larger scale trials by the absence (green line) of risk of generalizability bias (RGB), RGB present in pilot and absent in larger scale trial (red distribution), and RGB present in both pilot and larger scale trial (blue distribution). Positive *z* values indicate that the intervention was better than the control group, and negative *z* values indicate that the control group was better than the intervention group. Solid vertical lines represent $z \pm 1.96$ ($p = 0.05$); dashed vertical lines represent $z \pm 2.58$ ($p = 0.005$)

rationale for why interventionists may want to consider not introducing an RGB into their early-stage study. A lack of guidelines may stem from the limited evidence, to date, demonstrating the potential impact of the RGBs on early-stage studies and their larger scale trial outcomes. Thus, the scientific field may be relatively unaware of the potential influence that the inclusion of RGBs in preliminary studies has on decisions related to scaling behavioral interventions.

Another potential reason for introducing RGBs is the need to demonstrate early success in a pilot study to receive funding for a larger scale trial. Large-scale trials require strong preliminary data. Embedding RGBs into early-stage studies may provide a means to this end. A researcher may unwittingly or unintentionally introduce RGBs within preliminary studies to enhance perceived scientific credibility of the evidence to support a larger scale trial. Given the hypercompetitive funding environment, producing strong preliminary data, which includes clearly demonstrating preliminary efficacy and promise of an intervention, may provide a logical rationale for testing an intervention at the early stages with one or more RGBs embedded within. Despite how justified and RGB introduction may be, our findings indicate that doing so has important ramifications for the outcomes of larger scale trials. With larger scale trials of interventions requiring some form of preliminary data, combined with the widespread introduction of RGBs within early pilot work, an environment is potentially created where awarding grants to support large-scale studies is high risk, because the pilot/feasibility work no longer provides reliable evidence of an intervention's likelihood of success.

Every occurrence of an RGB, however, is not inappropriate. On the contrary, RGBs within "first run" interventions can prove useful for refining intervention components. Investigators who deliver interventions themselves might gain insight about how participants react to content and whether changes are necessary. Employing RGBs may be necessary during the very early-stages of testing, but only if such studies lead to another pilot without RGBs present. The problem arises when "first run" interventions with RGBs go directly to a larger scale trial, without conducting an intermediary trial that more closely mimics the conditions of the anticipated larger scale trial. A sequence of studies that goes from a first run intervention to another, progressively larger, pilot/feasibility trial then onto a larger scale more well-powered trial, may be ideal.[1,2] Conducting multiple iterations of an intervention's content, refining the content and then re-piloting could assist in identifying the correct ingredients for an intervention. This sequence implies funding is available to support a sequential, iterative process of intervention development, refinement, and testing.

The following are recommendations regarding the RGBs and early-stage preliminary studies. First, we suggest interventionists avoid introducing RGBs. Where RGBs are introduced this needs to be supported with rationale, such as first run interventions, and discussion whether their presence impacts outcomes. Existing guidelines (e.g., CONSORT) should include the RGBs in the required reporting of pilot/feasibility studies and their subsequent larger scale trial. In published pilot/feasibility studies, details should be provided as to the anticipated conditions under which a larger scale trial may be conducted and what changes, if any, to RGB-related items (e.g., who delivers the intervention, the length of the intervention, the setting where the intervention is conducted) may occur and how these may influence the larger trial results. Finally, clear linkages should be made among studies used to inform a larger scale trial. Pilot/feasibility studies need to be clearly identified and published and these should be clearly referred to in a subsequent larger scale trial they informed.

There are several limitations in this study. First, while this study represents the largest pairing of pilot and large-scale trials to date, there are, undoubtedly, more pairs, which exist but were not captured due to our stringent inclusion criteria. We predicated our search

strategy on the premise that larger scale trials would reference published pilot studies if they existed. Thus, our search strategies likely resulted in a comprehensive list of pairings that could be made, albeit small in overall number compared with the number of larger scale trials that exist. There are instances where a larger scale trial was identified that did not explicitly reference prior pilot/feasibility studies, and there were instances where a pilot/feasibility study was identified but no larger scale trial could be found. Second, the coding of the RGBs relied upon authors to clearly provide information upon which to judge their presence/absence. A recent study found only 12 of 200 randomized controlled trials (RCT) provided sufficient information about who delivered the intervention.[39] Other aspects of intervention reporting, such as the location of where the intervention occurred, are also poorly detailed in published RCTs. The ambiguousness or absence of this information makes identification of the RGBs difficult. The analyses conducted herein likely contain some studies that have an RGB but were coded as not having due to inadequate reporting of these important features. Hence, the impact of the RGBs could be greater or lesser, depending on the outcomes of studies that have unclear reporting. Third, not all of the RGBs demonstrated an impact. This is not entirely unexpected and could be due to the issues raised above or the reliance upon only 114 pairs. Uncertainty about the exact magnitude of the effects of RGBs is substantial. Finally, comparisons using pilot/feasibility effect sizes comes with issues of inaccurate and inflated effect sizes often observed in early-stage work. Exact replication of an effect from a pilot/feasibility study to a larger scale trial is not the purpose of pilot/feasibility testing. Thus, analyses comparing reported effect sizes in pilot/feasibility study to those observed in the larger scale trial are inherently limited. We agree effect sizes in pilot/feasibility studies can be imprecise and inflated and that they should not be used to inform power analyses of a larger scale trial. Yet, it is common practice for meta-analyses to include pilot/feasibility studies in thereby giving scientific credibility to the effects they report. To address this, the analyses conducted herein include both a comparison the pilot/feasibility to larger scale effect sizes as well as solely focusing on the effects reported in the larger scale trials. These findings were consistent across these analyses demonstrating the impact RGBs have in trial outcomes. Fourth, even larger scale trials may be biased or inaccurate for many other reasons not captured by RBGs; therefore, their treatment effects should not be seen as an absolute gold standard.

In conclusion, the RGBs demonstrated moderate to strong support of their influence on the success of larger scale trials. Consideration of the RGBs and how they can potentially misinform decisions about whether an intervention is ready for scale is critical given the time and resources required to conduct larger scale trials. Future preliminary, early-stage work needs to consider whether the introduction of one or more RGBs is justifiable and if their presence will lead to incorrect decisions regarding the viability of an intervention.

ACKNOWLEDGMENTS

We would like to thank all the authors of the published studies included in this review.

FUNDING INFORMATION

Research reported in this abstract was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health (under award number R01HL149141 [Beets], F31HL158016 [von Klinggraeff], and F32HL154530 [Burkart]) as well as by the Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health (under award number P20GM130420) for the Research Center for Child Well-Being. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Dr Lubans is supported by a National Health and Medical Research Council Senior Research Fellowship (APP1154507). Dr Okely is supported by a National Health and Medical Research Council Investigator Grant (APP1176858). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The work of John Ioannidis is supported by an unrestricted gift from Sue and Bob O'Donnell.

CONFLICT OF INTEREST

No conflict of interest statement.

AUTHOR CONTRIBUTIONS

M. B. 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.

ORCID

Lauren von Klinggraeff https://orcid.org/0000-0002-4417-0701
David Lubans https://orcid.org/0000-0002-0204-8257

REFERENCES

1. Czajkowski SM, Powell LH, Adler N, et al. From ideas to efficacy: the ORBIT model for developing behavioral treatments for chronic diseases. *Health Psychol*. 2015;34(10):971-982.
2. Onken LS, Carroll KM, Shoham V, Cuthbert BN, Riddle M. Reenvisioning clinical science: unifying the discipline to Improve the public health. *Clin Psychol Sci*. 2014;2(1):22-34.
3. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases. Small R01s for clinical trials targeting Diseases within the Mission of NIDDK (R01 Clinical Trial Required). National Institutes of Health. https://grants.nih.gov/grants/guide/pa-files/pas-20-160.html. Published 2020. Accessed April 1st, 2021.
4. Pearson N, Naylor PJ, Ashe MC, Fernandez M, Yoong SL, Wolfenden L. Guidance for conducting feasibility and pilot studies for implementation trials. *Pilot Feasibility Stud*. 2020;6(1):167.
5. McCrabb S, Mooney K, Elton B, Grady A, Yoong SL, Wolfenden L. How to optimise public health interventions: a scoping review of guidance from optimisation process frameworks. *BMC Public Health*. 2020;20(1):1849.
6. Chambers DA, Glasgow RE, Stange KC. The dynamic sustainability framework: addressing the paradox of sustainment amid ongoing change. *Implement Sci*. 2013;8(1):117.
7. 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.
8. Kilbourne AM, Neumann MS, Pincus HA, Bauer MS, Stall R. Implementing evidence-based interventions in health care: application of the replicating effective programs framework. *Implement Sci*. 2007; 2(1):42.
9. Lane C, McCrabb S, Nathan N, et al. How effective are physical activity interventions when they are scaled-up: a systematic review. *Int J Behav Nutr Phys Act*. 2021;18(1):16.
10. McCrabb S, Lane C, Hall A, et al. 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):964-982.
11. Beets MW, Weaver RG, Ioannidis JPA, et al. Identification and evaluation of risk of generalizability biases in pilot versus efficacy/effectiveness trials: a systematic review and meta-analysis. *Int J Behav Nutr Phys Act*. 2020;17(1):19.
12. The Cochrane Collaboration. The Cochrane Handbook for Systematic Reviews of Interventions: Handbook is 5.1 [updated March 2011]. http://handbook.cochrane.org. Published 2011. Accessed.
13. 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(1):8-34.
14. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ*. 2014;348(mar07 3):g1687.
15. Araújo-Soares V, Hankonen N, Presseau J, Rodrigues A, Sniehotta FF. Developing behavior change interventions for self-management in chronic illness: an integrative overview. *Eur Psychol*. 2019;24(1): 7-25.
16. Sallis JF, Owen N, Fisher E. Ecological models of health behavior. In: *Health Behavior: Theory, Research, and Practice*. Vol.5; 2015:43-64.
17. Leon AC, Davis LL, Kraemer HC. The role and interpretation of pilot studies in clinical research. *J Psychiatr Res*. 2011;45(5):626-629.
18. Arain M, Campbell MJ, Cooper CL, Lancaster GA. What is a pilot or feasibility study? A review of current practice and editorial policy. *BMC Med Res Methodol*. 2010;10(1):67.
19. Stevens J, Taber DR, Murray DM, Ward DS. Advances and controversies in the design of obesity prevention trials. *Obesity*. 2007;15(9): 2163-2170.
20. Eldridge SM, Lancaster GA, Campbell MJ, et al. Defining feasibility and pilot studies in preparation for randomised controlled trials: development of a conceptual framework. *PLoS ONE*. 2016;11(3): e0150205.
21. Puljak L, Makaric ZL, Buljan I, Pieper D. What is a meta-epidemiological study? Analysis of published literature indicated heterogeneous study designs and definitions. *J Comp Eff Res*. 2020;9(7): 497-508.
22. Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychol Methods*. 2002;7(1):105-125.
23. Waters E, de Silva-Sanigorski A, Hall BJ, et al. Interventions for preventing obesity in children. *Cochrane Db Syst Rev*. 2011;12.
24. R: a language and environment for statistical computing. R Foundation for Statistical Computing. [computer program]. 2015.
25. Tipton E. Small sample adjustments for robust variance estimation with meta-regression. *Psychol Methods*. 2015;20(3):375-393.
26. Konstantopoulos S. Fixed effects and variance components estimation in three-level meta-analysis. *Res Synth Methods*. 2011;2(1):61-76.
27. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3):1-48.
28. Ioannidis JPA. The proposal to lower *P* value thresholds to .005. *JAMA*. 2018;319(14):1429-1430.
29. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. *Nat Hum Behav*. 2018;2(1):6-10.
30. Freedland KE. Pilot trials in health-related behavioral intervention research: problems, solutions, and recommendations. *Health Psychol*. 2020;39(10):851-862.
31. Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. *J Eval Clin Pract*. 2004; 10(2):307-312.
32. Thabane L, Ma J, Chu R, et al. A tutorial on pilot studies: the what, why and how. *BMC Med Res Methodol*. 2010;10(1).
33. 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(3):238.
34. 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(3):285-298.
35. Koorts H, Cassar S, Salmon J, Lawrence M, Salmon P, Dorling H. Mechanisms of scaling up: combining a realist perspective and systems analysis to understand successfully scaled interventions. *Int J Behav Nutr Phys Act*. 2021;18(1):42.
36. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. *BMJ*. 2016; 355:i5239.
37. 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(may08 1):h2147.
38. 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;84:27-29.
39. Rauh SL, Turner D, Jellison S, et al. Completeness of Intervention reporting of clinical trials published in highly ranked obesity journals. *Obesity*. 2021;29(2):285-293.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Beets MW, von Klinggraeff L, Burkart S, et al. Impact of risk of generalizability biases in adult obesity interventions: A meta-epidemiological review and meta-analysis. *Obesity Reviews*. 2021;e13369. doi:10.1111/obr.13369