HHS Public Access

Author manuscript

J Clin Epidemiol. Author manuscript; available in PMC 2024 July 01.

Published in final edited form as:

J Clin Epidemiol. 2023 July ; 159: 70–78. doi:10.1016/j.jclinepi.2023.05.011.

Scientists' Perception of Pilot Study Quality Was Influenced by Statistical Significance and Study Design

Lauren von Klinggraeff, MPH,
University of South Carolina, Department of Exercise Science, Columbia, South Carolina, USA

Sarah Burkart, MPH, PhD,
University of South Carolina, Department of Exercise Science, Columbia, South Carolina, USA

Christopher D. Pfledderer, MEd, PhD,
University of South Carolina, Department of Exercise Science, Columbia, South Carolina, USA

Md. Nasim Saba Nishat,
University of South Carolina, Department of Epidemiology and Biostatistics, Columbia, South Carolina, USA

Bridget Armstrong, PhD,
University of South Carolina, Department of Exercise Science, Columbia, South Carolina, USA

R. Glenn Weaver, MEd, PhD,
University of South Carolina, Department of Exercise Science, Columbia, South Carolina, USA

Alexander McLain, PhD,
University of South Carolina, Department of Epidemiology and Biostatistics, Columbia, South Carolina, USA

Michael W. Beets, MEd, MPH, PhD
University of South Carolina, Department of Exercise Science, Columbia, South Carolina, USA

Abstract

Objectives: Preliminary studies play a key role in developing large-scale interventions but may be held to higher or lower scientific standards during the peer review process because of their preliminary study status.

*Corresponding author: vonklinl@email.sc.edu, 921 Assembly Street, Columbia SC 29208 USA, USA (303) 908-3876.

Contributor Roles Taxonomy (CRediT) Author Statement

LvK: Conceptualization, Methodology, Validation, Formal analysis, Data Curation, Writing – Original Draft, Visualization, Funding acquisition. **SB**: Conceptualization, Methodology, Validation, Data Curation, Writing - Review & Editing, Visualization. **CDF:** Methodology, Validation, Data Curation, Writing - Review & Editing. **MNSN:** Methodology, Formal analysis. **BA:** Conceptualization, Writing - Review & Editing. **RGW**: Conceptualization, Writing - Review & Editing. **AM:** Methodology, Validation, Formal analysis, Writing - Review & Editing. **MWB:** Conceptualization, Methodology, Validation, Formal analysis, Writing - Review & Editing, Visualization, Supervision, Funding acquisition

Competing Interest

The authors have no competing interests or conflicts of interest to declare.

Study Design: Abstracts from five published obesity prevention preliminary studies were systematically modified to generate 16 variations of each abstract. Variations differed by four factors: sample size (n=20 vs. n=150), statistical significance (P<.05 vs. P>.05), study design (single group vs. randomized two group), and preliminary study status (presence/absence of pilot language). Using an online survey, behavioral scientists were provided with a randomly selected variation of each of the five abstracts and blinded to the existence of other variations. Respondents rated each abstract on aspects of study quality.

Results: Behavioral scientists (n=271, 79.7% female, median age 34 years) completed 1,355 abstract ratings. Preliminary study status was not associated with perceived study quality. Statistically significant effects were rated as more scientifically significant, rigorous, innovative, clearly written, warranted further testing, and had more meaningful results. Randomized designs were rated as more rigorous, innovative, and meaningful.

Conclusion: Findings suggest reviewers place a greater value on statistically significant findings and randomized control design and may overlook other important study characteristics.

Keywords

Pilot Projects; Feasibility Studies; Bias; Data Interpretation; Statistical; Peer Review; Research; Evaluation Studies

1. Introduction

The peer review process is intended to ensure the quality, credibility, and integrity of scientific research and allows scientists to collectively recommend research for dissemination or financial support.[1–3] For preliminary behavioral intervention studies (e.g., pilot/feasibility studies), the peer review process is often used to determine whether the intervention shows sufficient indications of promise to warrant the further testing of the intervention in a fully-powered, definitive study.[4,5] Favorable peer review may increase the probability a preliminary study can obtain funding for larger scale trials; thus, the peer-review of preliminary studies plays a pivotal role in the scale-up of behavioral interventions.

Peer review does not always provide a rigorous evaluation of research and is known to be subject to systematic biases. Even when studies are identical in all factors, peer reviewers are known to favor papers from well-known authors and institutions,[6] rate treatments as more beneficial when positive secondary or subgroup findings are emphasized over nonsignificant primary outcomes (i.e., research spin),[7–10] and deem studies of orthodox interventions as more important than studies of unconventional interventions (i.e., confirmation bias).[11]

It is plausible the labeling of a study as preliminary (i.e., “pilot”, “feasibility”) could influence study appraisal by peer-reviewers. Pilot studies are often conducted in small samples, and may be underpowered to detect statistical significance, causing them to be overly or underly criticized. Further, trials with pilot-like characteristics (e.g., small sample sizes, testing new interventions) appear frequently, but may not be labeled by the authors as a preliminary study in the publication. Explicitly labeled preliminary studies and small trials without a preliminary study label may function similarly in their ability to inform larger, more well-powered trials. However, they may be viewed very differently in terms of

their rigor and what they can achieve in informing another, larger-scale intervention, simply because one is clearly identified as a preliminary study and the other is not.[12] Therefore, we aim to understand if there are differences in how pilot studies and small trials are viewed regarding evidence to scale up to a well-powered trial.

The primary aim of this study was to understand whether peer-reviewers evaluate pilot studies differently than identical studies not labeled as pilots. We hypothesized peer-reviewers would perceive intervention studies labeled as pilots as less developed (i.e., lower ratings on methodological rigor) and more promising (i.e., higher ratings on warrants further testing). Secondarily, we explored whether peer-reviewers differentiate based upon additional study characteristics, including sample size, study design, and statistical significance.

2. Materials and Methods

We designed a double-blind full factorial randomized control trial to evaluate the impact study characteristics have on peer-reviewers' assessments of preliminary studies. Abstracts from five published obesity prevention preliminary studies were systematically modified to generate 16 variations of each abstract which differed in sample size, statistical significance, study design, and preliminary study status. A small sample size was set to n=20, in a single group or n=40 with two group studies, with each group containing n=20. A large sample size was set to n=150 in single group studies with n=75 in each group for two-group studies. Within studies reporting RCT designs, control groups were provided with educational information. P-values for statistically significant studies were set between a range of p=0.02 to p=0.05, including "p<0.05". P-values for non-statistically significant studies were set between p=0.35 and p=0.07. P-values were provided in each abstract, with some abstracts presenting mean differences, confidence intervals and/or effect sizes. Within each abstract, the presented statistical information was identical between studies in different factor groups. In other words, all non-statistically significant variations within an abstract reported identical finding and all statistically significant variations with an abstract presented identical findings. All abstracts are presented in Supplement A.

Blind to the existence of abstract variations, behavioral scientists were randomly provided with one variation of each of five abstracts and surveyed on six measures: 1) significance of the scientific question, 2) methodological rigor, 3) study innovation, 4) writing clarity, 5) whether the findings warranted further testing, and 6) the probability of obtaining meaningful results in a subsequent larger-scale trial.

2.1 Participant Recruitment

We sought out behavioral intervention researchers with experience in behavioral interventions using several recruitment strategies. Direct emails were sent to behavioral intervention researchers using Special Interest Group (SIG) listservs managed by the Society of Behavioral Medicine (SBM) and the International Society of Behavioral Nutrition and Physical Activity (ISBNPA). Emails included an invitation and a link to a Qualtrics survey (Qualtrics, Seattle WA). We also asked the SIGs and conference organizers to post our survey in their newsletters and official Twitter accounts. Respondents were provided with

a $25 Amazon gift card for completing the survey. Survey completion was defined as completing all questions in the survey (i.e., selecting answers to all six Likert scales for all five randomly assigned abstracts). All procedures were approved by the university's Institutional Review Board (IRB Pro00101406).

2.2 Abstract Selection

Our group previously identified 153 pilot/feasibility behavioral intervention studies in the areas of child and adult obesity treatment/prevention that had been published in refereed journals and subsequently funded for larger, more well-powered trials.[13–15] From this collection, we randomly selected five pilot/feasibility study abstracts. The selected abstracts were published between 2004 and 2015 and targeted five different populations including peripartum weight (primary outcome: gestational weight gain in pounds), weight loss in adult men and women (primary outcome: weight loss in kilograms), health intervention for rural adults (primary outcome: changes in food and physical activity environment), school-based health programing for middle school students (primary outcome: BMI), physical activity for middle school girls (primary outcome: moderate to vigorous physical activity).

2.3 Constructing Abstract Variations

Selected abstracts were modified to generate 16 nearly identical variations of each abstract using a full factorial design inspired by a similar study on research spin.[8] This produced 80 total abstracts which are provided in the supplemental material (Supplement A). Abstract variations differed by four factors: sample size (n=20 vs. n=150), statistical significance (P<.05 vs. P>.05), study design (single group vs. randomized two group), and preliminary study designation (presence/absence of pilot/feasibility language). Reported direction and magnitude of change in outcome measures (e.g., effect sizes), intervention components, population, and duration were maintained across all variations within an abstract. An example of two variations of Abstract A are presented in Table1.

2.4 Survey Design

The survey was comprised of three parts. First, respondents provided active consent to participate in the study after reading an IRB-approved study description on the survey landing page. In the second section, respondents were asked to provide general demographic information. This included their age, country of residence, and whether they were a graduate student, post-doctoral fellow, or years since completing their terminal degree (from < 1 year/ graduate student' through '30+ years'). They were also asked to provide their professional demographics (e.g., title, areas of expertise; complete survey presented in Supplement B).

For the remainder of the survey, respondents were asked to read five abstracts (Abstracts A-E) and rate them on six distinct 10-point Likert scales. We designed the scales to mimic the conference abstract review process used by prominent professional societies in behavioral health intervention research including the SBM and ISBNPA. Past and current rating scales used by each society were obtained from conference organizers via email and used as templates for the scales in the current study. The rating scales and prompts used in this study are presented in Table 2.

2.5 Study Design

Using the Qualtrics survey random assignment feature, we were able to implement a double-blinded randomized full factorial design in which each respondent was randomly provided one of 16 variations for each of the five abstracts by the survey interface. The research team had no role in assigning abstracts to respondents. Participants were not aware (i.e., blinded) of the existence of abstract variations other than those they were shown. Respondents were instructed to "rate the quality, strength of the findings, and whether a future evaluation of the same or similar intervention is warranted based upon the findings" and were not aware that the purpose of the survey was to evaluate the main effect of abstract attributes on reviewers' ratings of study merit.

2.6 Analysis

In the a priori power analysis, effect of pilot label was assumed to have a medium effect size of 0.5. Alpha was set equal to 5%. Power was set equal to 0.8. This resulted in a required minimum sample size of 144 respondents or nine individual ratings per abstract variation. Power analysis was completed in R.

Using STATA 16, The primary analysis was performed with multilevel mixed effects models evaluating the main effect of abstract pilot/feasibility designation and all 2-way interactions with abstract pilot/feasibility designation. Presented models included only main effects.[16] A secondary analysis was preformed restricted to only abstracts presenting studies containing pilot language. Mixed effects linear models evaluated the effect of abstract attributes (e.g., sample size, statistical significance, study design, preliminary study status) on outcomes. Ratings of abstracts were nested within respondents. For all scales, higher Likert values represented more positive ratings and all models used the absence of a characteristic as the referent. Model residuals were examined for their adherence to the assumptions of linear mixed effects models; no violations were found.

3. Results

3.1 Survey Responses

The survey was distributed from September 14, 2021, to January 3, 2022, resulting in 3,126 survey entries. Of these, 2,780 entries were determined to be false, robotic entries (i.e., 'bots') using Qualtrics bot detection features and bot-indicative patterns (e.g., completing the survey in less than three minutes) and were excluded. Of the non-robotic responses, 75 entries were incomplete producing a final analytic sample of 271 entries (Figure 1)[17] A completed survey entry consisted of answering six questions for five abstracts. In total, each abstract variation was rated by a mean of 16.9 respondents (median 13, range 13–22; Table 3).

3.2 Demographics

Respondents mostly resided in the United States (n=249, 91.9%), were female (n=216, 79.7%) and served in an academic role (n=179, 66.1%). Respondents reported having authored or co-authored cross-sectional studies (n=217, 80.0%), qualitative studies (n=157, 57.9%), and cohort studies (n=145, 53.5%) within the last three years. The average age of

respondents was 36.3 years (range 22–74, median 34). Among respondents with a terminal degree (n=210, 77.4%) the mean duration since terminal degree was 7.5 years (scale stopped at '30+'). Graduate students comprised 22.5% (n=61) of respondents. Complete respondent characteristics are listed in Table 4.

3.4 Researchers' Interpretation of Abstracts

For the primary outcome (Table 5), there was no evidence the preliminary study label impacted reviewers' interpretations of studies' scientific significance, rigor, innovation, clarity, or meaningfulness. Evaluation of the impact of other study characteristics showed abstracts presenting statistically significant results were rated higher on all attributes. Abstracts with a randomized design were also rated higher on all attributes, but generally had smaller effect sizes compared to the impact statistical significance. Abstract variations with larger sample sizes (n=150) were not rated differently than abstracts with smaller sample sizes (n=20). Interclass correlations for between-rater variability ranged from 0.36 – 0.55.

Models restricted to only labeled preliminary studies (Table 6; Supplemental Material) showed similar patterns. Abstracts with statistically significant results were considered more scientifically significant, rigorous, more clearly written, warranted further testing, and had a higher likelihood of producing meaningful results in subsequent testing. Among abstracts labeled as pilot studies, randomized designs were also considered more scientifically significant, more rigorous, more clearly written, warranted further testing, and were more likely to produce meaningful results in subsequent testing.

4. Discussion

We hypothesized pilot-labeled abstracts would be perceived as lower quality by peer-reviewers. Contrary to our hypothesis, reviewers did not differentiate between studies labeled "preliminary" and identical studies not labeled "preliminary". Results indicated studies with statistically significant findings and RCT design were considered more meritorious, irrespective of preliminary study status, and sample size. Participants even rated the clarity of writing higher in studies with statistical significance, even though the writing quality was identical among all variations of the abstract. These findings indicate reviewers have a notable implicit bias toward abstracts reporting statistical significance and/or RCT designs, with similar patterns observed among abstracts labeled as preliminary studies.

Statistical significance had a consistent and significant influence on study ratings, such that respondents rated studies reporting statistical significance higher than studies with nonsignificant results, even though the content of the abstracts was identical. This held true even for studies labeled as "preliminary". These findings have important implications for study interpretation. While it may be appropriate to conduct hypothesis testing in a large-scale, adequately powered study, smaller, preliminary studies are more likely to produce false, exaggerated findings.[18,19] This is because statistical significance in smaller, earlier studies is less likely to represent the true effect of the intervention but rather a product of the variability in the sample[20–22] or an inflated effect associated with introduced intervention artifacts. This is problematic because the effects are unlikely to be maintained at-scale.[13–15]

Given larger trials take increased time and resources, chasing false positives which have low impact at-scale reduces the efficiency with which public health interventions are developed and widely disseminated.

Abstracts reporting results from randomized controlled trials (RCTs) were rated more positively than single group studies. RCTs were thought to be more rigorous than single group studies, as well as more innovative and more likely to produce meaningful results in a subsequent, large-scale trial. RCT designs allow researchers to support the internal validity of their study findings, so it is logical that reviewers perceived RCTs as more rigorous and more likely to produce meaningful results. It is less clear why RCTs had higher ratings of innovation, as the intervention and measures were identical between the single-group and RCT variations of each abstract. Even when constricting analyses to just those abstracts labeled as “preliminary”, RCTs were considered more clearly written, more rigorous and more likely to produce meaningful results in subsequent, larger studies compared to single group studies. RCTs are largely considered the gold standard of study design, though the quality of information they produce is related to other study factors such as sample size. Our study indicates reviewers struggle to delineate study design and study rigor. This may lead to biases towards publishing and funding RCTs, with single group studies less likely to be viewed favorably.

Given the strong implicit bias reviewers show towards statistical significance, it is plausible researchers seeking larger-scale funding may knowingly or unknowingly introduce artifacts into preliminary studies to increase the odds of producing statistically significant findings.[15] The Risk of Generalizability Biases (RGBs) are study/intervention artifacts known to artificially inflate the magnitude of effects and statistical significance of preliminary studies.[13,14] These artifacts are easily identified and largely avoidable. They include the delivery of the intervention by a principal investigator (or graduate students) or delivering an intervention to a readily accessible population different from the population to be targeted in a larger-scale study. Systematic prevention of non-scalable intervention artifacts (i.e., RGBs), in addition to minimizing hypothesis testing in underpowered, earlier studies, could improve the validity of early-stage scientific findings and ensure funding for larger-scale studies is directed towards truly promising interventions.

4. 1 Strengths and Limitations

Our study utilized a double blinded full factorial RCT design to evaluate the impact of study characteristics on reviewers’ interpretation of a study abstract’s scientific merit. To our knowledge, few RCTs have been conducted in the field of evidence evaluation, and none on the role of preliminary study designation. Second, we had a sufficient sample size comprised of current researchers from a variety of positions, with an even spread of professional experience ranging from current graduate students to faculty with over 30 years’ experience. This is important because recommendations for reviewing preliminary studies have shifted over the past twenty years[12,23,24] and surveying a wide range of ages allows for more realistic representation of the span in peer reviewer positions.

It should be noted that our study has some limitations. Respondents did not have access to the full-text articles to use in evaluating each study. However, abstract-only evaluation

is a common format of evaluation used in scientific conferences and initial manuscript reviews.[25] For journal articles not published under open access, an abstract may also be the only available report of study findings. Second, the original abstracts were chosen from the childhood and adult obesity literature and 20% of respondents did not select obesity, physical activity, nutrition, or diabetes as their area of expertise. This may have impacted their ability to evaluate some characteristics of the abstract. Though this may also increase the generalizability of our findings, as it is common for research grants to be reviewed by individuals who are not content experts. It should also be noted that our respondents were overwhelmingly from the United States (91.9%) though similar discussions about overemphasizing statistical significance and appropriate interpretation of preliminary intervention studies are ongoing worldwide.

5. Conclusion

Peer-reviewers perceive abstracts reporting statistically significant findings more favorably regardless of study design, sample size, or preliminary designation. This means that a single group, preliminary study with a sample size of 20 participants and statistically significant findings was considered more likely to produce meaningful results in a definitive trial than an identical intervention tested with 150 participants in an RCT design. Adhering to best scientific practices requires that statistical significance be interpreted alongside critical study features including sample size, effect size, sample distribution, and study design.[26] Given the crucial role peer-review plays in the evaluation and funding of behavioral science, efforts to improve the consistency of peer-evaluation, particularly for small, early studies reporting hypothesis testing, is warranted. Comprehensive evaluations of preliminary behavioral interventions may ensure interventions providing adequate evidence of promise are prioritized for further study, therein producing more effective at-scale interventions, and ultimately improving population health outcomes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the respondents for participating in our survey.

Funding Information

Research reported in this manuscript was supported by The National Heart, Lung, and Blood Institute of the National Institutes of Health under award number R01HL149141 (Beets), F31HL158016 (von Klinggraeff), 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.

Availability of data and materials

Access to the data will be made available upon reasonable request to the corresponding author. Preliminary results were presented orally at the International Society of Behavioral Nutrition and Physical Activity 2022 Annual Meeting.

References

1. Neumann N Imperfect but important: a fellow's perspective on journal peer review. Journal of Medical Toxicology. 2020;16(1):1–2. [PubMed: 31853737]
2. Tamblyn R, Girard N, Qian CJ, Hanley J. Assessment of potential bias in research grant peer review in Canada. Canadian Medical Association Journal. 2018;190(16):E489. [PubMed: 29685909]
3. Tennant JP, Ross-Hellauer T. The limitations to our understanding of peer review. Research Integrity and Peer Review. 2020;5(1):6. [PubMed: 32368354]
4. Recio-Saucedo A, Crane K, Meadmore K, et al. What works for peer review and decision-making in research funding: a realist synthesis. Research Integrity and Peer Review. 2022;7(1):2. [PubMed: 35246264]
5. Health NIo. About Grants Peer Review. Grants & Funding: NIH Central Resource for Grants and Funding Information Web site. https://grants.nih.gov/grants/peer-review.htm. Accessed.
6. Okike K, Hug KT, Kocher MS, Leopold SS. Single-blind vs double-blind peer review in the setting of author prestige. Jama. 2016;316(12):1315–1316. [PubMed: 27673310]
7. Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and Interpretation of Randomized Controlled Trials With Statistically Nonsignificant Results for Primary Outcomes. JAMA. 2010;303(20):2058–2064. [PubMed: 20501928]
8. Boutron I, Altman DG, Hopewell S, Vera-Badillo F, Tannock I, Ravaud P. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. J Clin Oncol. 2014;32(36):4120–4126. [PubMed: 25403215]
9. Jankowski S, Boutron I, Clarke M. Influence of the statistical significance of results and spin on readers' interpretation of the results in an abstract for a hypothetical clinical trial: a randomised trial. BMJ Open. 2022;12(4):e056503.
10. Wegwarth O, Schwartz LM, Woloshin S, Gaissmaier W, Gigerenzer G. Do physicians understand cancer screening statistics? A national survey of primary care physicians in the United States. Ann Intern Med. 2012;156(5):340–349. [PubMed: 22393129]
11. Kaptchuk TJ. Effect of interpretive bias on research evidence. Bmj. 2003;326(7404):1453–1455. [PubMed: 12829562]
12. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ. 2016;355:i5239. [PubMed: 27777223]
13. 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. International Journal of Behavioral Nutrition and Physical Activity. 2020;17(1):19. [PubMed: 32046735]
14. 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. Obes Rev. 2022;23(2):e13369. [PubMed: 34779122]
15. von Klinggraeff L, Dugger R, Okely AD, et al. Early-stage studies to larger-scale trials: investigators' perspectives on scaling-up childhood obesity interventions. Pilot and Feasibility Studies. 2022;8(1):31. [PubMed: 35130976]
16. Kugler KC, Dziak JJ, Trail J. Coding and Interpretation of Effects in Analysis of Data from a Factorial Experiment. In: Collins LM, Kugler KC, eds. Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: Advanced Topics. Cham: Springer International Publishing; 2018:175–205.
17. Sharma A, Minh Duc NT, Luu Lam Thang T, et al. A Consensus-Based Checklist for Reporting of Survey Studies (CROSS). J Gen Intern Med. 2021;36(10):3179–3187. [PubMed: 33886027]
18. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2(8):e124. [PubMed: 16060722]
19. Ioannidis JP. Scientific inbreeding and same-team replication: type D personality as an example. J Psychosom Res. 2012;73(6):408–410. [PubMed: 23148806]
20. Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience. 2013;14(5):365–376. [PubMed: 23571845]

21. Slavin R, Smith D. The Relationship between Sample Sizes and Effect Sizes in Systematic Reviews in Education. Educational Evaluation and Policy Analysis. 2009;31(4):500–506.
22. Sullivan GM, Feinn R. Using Effect Size-or Why the P Value Is Not Enough. J Grad Med Educ. 2012;4(3):279–282. [PubMed: 23997866]
23. Bowen DJ, Kreuter M, Spring B, et al. How we design feasibility studies. Am J Prev Med. 2009;36(5):452–457. [PubMed: 19362699]
24. Pearson N, Naylor P-J, Ashe MC, Fernandez M, Yoong SL, Wolfenden L. Guidance for conducting feasibility and pilot studies for implementation trials. Pilot and feasibility studies. 2020;6(1):167–167. [PubMed: 33292770]
25. Scherer RW, Meerpohl JJ, Pfeifer N, Schmucker C, Schwarzer G, von Elm E. Full publication of results initially presented in abstracts. Cochrane Database Syst Rev. 2018;11(11):Mr000005. [PubMed: 30480762]
26. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed). 1986;292(6522):746–750.

What's New

- Bias in scientific peer-review is well documented but has not been studied specifically in preliminary (e.g., pilot, feasibility) studies.
- Peer-reviewers perceive abstracts of behavioral interventions reporting statistically significant findings more favorably than those reporting non-significant findings, regardless of sample size or preliminary study designation.
- Small, early studies are known to produce unreliable results.
- Gatekeeping institutions arranging scientific peer-review should consider creating clear and enforceable guidance for reviewing preliminary behavioral interventions in which the interpretation of hypothesis testing (i.e., statistical significance) in underpowered studies is discouraged.

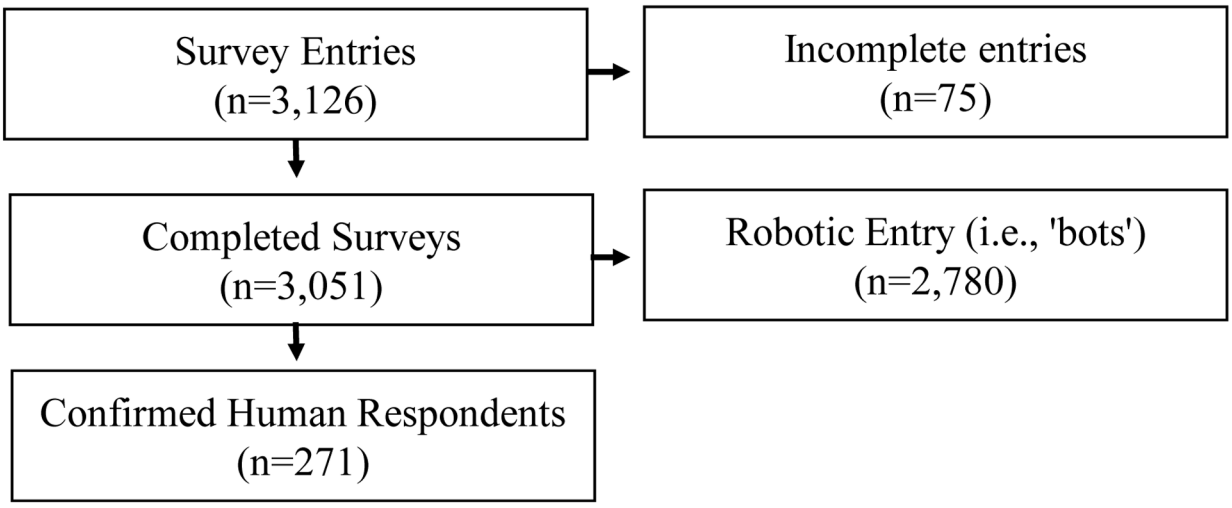

Figure 1.

CROSS[a] flow diagram of respondents. [a] Checklist for Reporting of Survey Studies

[b] Incomplete entries were define as surveys in which respondents exited the survey prior to the survey end, or who left the survey unfinished after more than 24 hours.

[c] To be considered a robotic entry, the responses had to occur during “blast” of a large number of responses, be linked to a Twitter recruitment post, and have a survey completion time of less than three minutes.

Table 1:

Two examples of Abstract A Variations. Abstract rewritten as a small, single group pilot study with statistically significant results (Variation 1) and an abstract rewritten as a large, randomized controlled trial (RCT), without statistical significance (Variation 8)

Rewritten Abstract A.1 **Small Single-Group Statistically Significant Pilot Study**	**Rewritten Abstract A.8** **Large RCT Study without Statistical Significance**
Background: The Physical Activity Guidelines call for youth to engage in at least one hour of moderate-to-vigorous physical activity each day. Less than 20% of adolescent girls meet recommendations with even fewer meeting recommendations as they transition into adulthood. School-based programs are a popular intervention to increase adolescent physical activity, though after-school, girls-only programs remain understudied. This study examined the feasibility and preliminary efficacy of a girls-only physical activity intervention delivered after school.	**Background**: The Physical Activity Guidelines call for youth to engage in at least one hour of moderate-to-vigorous physical activity each day. Less than 20% of adolescent girls meet recommendations with even fewer meeting recommendations as they transition into adulthood. School-based programs are a popular intervention to increase adolescent physical activity, though after-school, girls-only programs remain understudied. This study examined the impact of a girls-only physical activity intervention delivered after school.
Participants and Methods: This 6-month pilot study employed a single-group design. Girls (n=20) attended a 60-minute after-school program five days a week along with two face-to-face counseling session with their registered school nurse. At each session, the nurse utilized tenets of Motivational Interviewing and had girls set physical activity goals. At pre and post, accelerometer-measured physical activity and cardiovascular fitness were the primary outcomes.	**Participants and Methods**: This 6-month study employed a two-group randomized control design. Girls in the intervention (n=75) attended a 60- minute after-school program five days a week along with two face-to-face counseling session with their registered school nurse. At each session, the nurse utilized tenets of Motivational Interviewing and had girls set physical activity goals. Girls in the control group (n=75) received a 60-minute after-school workshop once a month for six months and attended face-to-face sessions with a registered school nurse two times to discuss topics covered in workshops. At pre and post, accelerometer-measured physical activity and cardiovascular fitness were the primary outcomes.
Results: Controlling for baseline measures, regression analyses showed statistically significant changes with improvement for minutes of moderate to vigorous physical activity ($p < 0.01$, Cohen's d = 0.43) and cardiovascular fitness ($p = 0.05$, d = 0.27).	**Results**: Controlling for baseline measures, regression analyses showed no statistically significant group differences with trends towards significance of greater intervention group improvement for minutes of moderate to vigorous physical activity ($p = 0.21$ Cohen's d = 0.43) and cardiovascular fitness ($p = 0.32$, d = 0.27).
Conclusion: This pilot study demonstrated the feasibility of implementing a girls-only, physical activity program after school. Research on after school programs for girls appears promising and findings from this study warrant further investigation.	**Conclusion**: This study demonstrated the impact of implementing a girls-only, physical activity program after school, with results suggesting an improvement on activity and fitness levels. Research on after school programs for girls appears promising and findings from this study warrant further investigation.

Note: Underlined text has been added for emphasis and was not present during participants' reading.

All abstract variations are available in the online supplemental material (Supplement A).

Table 2:

Rating scales and prompts presented to participants with each abstract.

Significance	Rate the significance of the crucial scientific question, knowledge gap, or area of importance from not at all significant (1) to very significant (10).
Methodological Rigor	Was the overall design, methodology, measures, and analyses appropriate and rigorous on a scale from not at all rigorous (1) to very rigorous (10)?
Innovation	Is the study novel or creative in a way that moves behavioral science forward on a scale from not at all innovative (1) to very innovative (10)?
Further Testing	Do these findings warrant the further testing of this intervention in a fully powered, definitive study from very low probability (1) to very high probability (10)?
Subsequent Meaningful Results	Based on these findings, what is the probability of obtaining meaningful results in a subsequent fully powered, definitive study of this intervention from not at all warranted (1) to highly warranted (10)?
Clarity	Rate the clarity and quality of the writing from not at all clear (1) to very clear (10).

Table 3:

Number of respondents randomly assigned to each abstract variation.

	Variation 1	Variation 2	Variation 3	Variation 4	Variation 5	Variation 6	Variation 7	Variation 8	Variation 9	Variation 10	Variation 11	Variation 12	Variation 13	Variation 14	Variation 15	Variation 16
	Small	Small	Small	Small	Large	Large	Large	Large	Small	Small	Small	Small	Large	Large	Large	Large
	Sig.	Sig.	Non-Sig.	Non-Sig.	Sig.	Sig.	Non-Sig.	Non-Sig.	Sig.	Sig.	Non-Sig.	Non-Sig.	Sig.	Sig.	Non-Sig.	Non-Sig.
	RCT	Single Group	RCT	Single Group	RCT	Single Group	RCT	Single Group	RCT	Single Group	RCT	Single Group	RCT	Single Group	RCT	Single Group
	Pilot	Pilot	Pilot	Pilot	Pilot	Pilot	Pilot	Pilot	Non-pilot	Non-pilot	Non-pilot	Non-pilot	Non-pilot	Non-pilot	Non-pilot	Non-pilot
Rating Mean (range)	17.4 (16–20)	17.0 (13–19)	17.8 (16–19)	16.0 (15–18)	17.0 (16–18)	17.2 (15–20)	17.2 (15–29)	17.0 (16–19)	18.2 (17–19)	16.6 (15–18)	17.0 (15–19)	16.4 (16–17)	16.4 (15–18)	16.8 (14–22)	17.2 (16–19)	15.8 (14–17)
Total n	87	85	89	80	85	86	86	85	91	83	85	82	82	84	86	79

Small = total sample size equal to 20 in single group design and 40 in the randomized controlled trial; Large = total sample size equal to 150 in single group design and randomized controlled trial; Sig. = Statistical significance (e.g., $p < 0.05$); RCT = Randomized control trial; Non-pilot = abstract did not contain the words “pilot”, “feasibility”, or “preliminary”.

Table 4.

Demographic and other characteristics of survey respondents (n=271)

	M	SD
Age, years	36.3	8.9
	n	**%**
Female[1]	216	79.7
Grant Reviewer[2]	78	28.7
Location		
Asia	2	0.7
Australia	4	1.5
Europe	5	1.8
North America		
United States	249	91.9
Canada	7	2.6
South America	1	0.4
Years from terminal degree [3]		
< 1 year/graduate student	75	27.7
1–5	101	37.3
6–10	43	15.9
11+	52	19.9
Professional Title		
Doctoral Student	60	22.1
Post-doctoral Fellow	53	19.6
Assistant Professor	69	25.5
Associate Professor	36	13.3
Professor	21	7.7
Clinical Psychologist	5	1.8
Leadership-focused role[4]	6	2.2
Research-focused role[5]	15	5.5
Other[6]	6	2.2
Published Study Designs [7]		
Cross-sectional	217	80.1
Qualitative or Focus Group	157	57.9
Longitudinal or Cohort	145	53.5
Feasibility or Pilot	140	51.7
Randomized Controlled Trial	130	48.0
Systematic Review or Meta-Analysis	109	40.2
Commentary	82	30.3
Methods or Quantitative	65	24.0
Protocol	72	26.6
Quasi-experimental	42	15.5

	M	SD
No experience selected	4	1.5

[1] 0.74% (n=2) respondents identified as non-binary/prefer not to respond.

[2] Respondents were asked if they had served as a reviewer for competitive federal or government grants in the past three years.

[3] Mean and SD in table, and Median in manuscript included graduate students (calculated as “0”) and less than one year from terminal degree (calculated as “0.5”). Scale stopped at 30+ (calculated as “30”).

[4] Leadership roles encompass titles such as “director.”

[5] Research roles encompass titles such as “research coordinator.”

[6] Other encompasses titles such as “faculty” or “grant writer.”

[7] Respondents were asked to select the types of articles they had published (primary author or co-author) in the last 3 years.

Table 5:

mmary of mixed effects linear models predicting effects of study characteristics on perceived study quality

erception of Study Quality	RCT					Statistical Significance					Sample Size					Pilot Designation				
	Percent Difference	**Estimate**	**SE**	**95%CI**	**ES**	**Percent Difference**	**Estimate**	**SE**	**95%CI**	**ES**	**Percent Difference**	**Estimate**	**SE**	**95%CI**	**ES**	**Percent Difference**	**Estimate**	**SE**	**95%CI**	**ES**
gnificance	**1.7**	**0.17**	**0.08**	**(0.,0.33)**	**0.07**	**3.5**	**0.35**	**0.08**	**(0.19,0.51)**	**0.14**	−0.4	−0.04	0.08	(−0.2,0.12)	0.02	0.8	0.08	0.08	(−0.08,0.24)	0.03
gor	**11.2**	**1.12**	**0.09**	**(0.94,1.30)**	**0.44**	**5.4**	**0.54**	**0.09**	**(0.36,0.72)**	**0.22**	−1.1	−0.11	0.09	(−0.29,0.07)	0.04	0.7	0.07	0.09	(−0.11,0.25)	0.03
novation	**3.3**	**0.33**	**0.09**	**(0.15,0.52)**	**0.13**	**4.1**	**0.41**	**0.09**	**(0.23,0.60)**	**0.16**	−0.2	0.02	0.09	(−0.17,0.20)	0.01	0.6	0.06	0.09	(−0.12,0.25)	0.03
arity	**3**	**0.30**	**0.10**	**(0.10,0.49)**	**0.11**	**5.7**	**0.57**	**0.10**	**(0.37,0.76)**	**0.21**	−0.7	−0.07	0.10	(−0.27,0.13)	0.03	1.2	0.12	0.10	(−0.08,0.31)	0.04
rther esting	**3.9**	**0.39**	**0.10**	**(0.20,0.59)**	**0.13**	**11.0**	**1.10**	**0.10**	**(0.90,1.30)**	**0.37**	−1.2	−0.12	0.10	(−0.32,0.08)	0.04	0.2	0.02	0.10	(−0.18,0.22)	0.04
eaningful esults	**4.9**	**0.49**	**0.10**	**(0.30,0.69)**	**0.17**	**10.8**	**1.08**	**0.10**	**(0.89,1.28)**	**0.38**	0.1	0.01	0.10	(−0.19,0.20)	0.00	−0.1	−0.01	0.10	(−0.20,0.19)	0.00

d font indicates statistical significance p<0.05

reviations: RCT = Randomized Control Trial, SE = standard error, CI = confidence interval, ES = effect size, Cohen's D

1,355 observations across 271 individuals

J Clin Epidemiol. Author manuscript; available in PMC 2024 July 01.