



## *Comparative Effectiveness Review Disposition of Comments Report*

### ***Research Review Title:***

### ***Methods for Evaluating Natural Experiments in Obesity: A Systematic Evidence Review***

Draft review available for public comment from September 22, 2017 to October 26, 2017.

**Research Review Citation:** Bennet W, Cheskin LJ, Wilson RF, Zhang A, Tseng E, Shogbesan O, Knapp EA, Stuart E, Bass EB, Kharazzi H. Methods for Evaluating Natural Experiments in Obesity: A Systematic Evidence Review. Comparative Effectiveness Review No. 204. (Prepared by the Johns Hopkins University Evidence-based Practice Center under Contract No. 290-2012-00007-I). AHRQ Publication No. 18-EHC006-EF. Rockville, MD: Agency for Healthcare Research and Quality; December 2017.  
[www.effectivehealthcare.ahrq.gov/reports/final.cfm](http://www.effectivehealthcare.ahrq.gov/reports/final.cfm). DOI:  
<http://doi.org/10.23970/AHRQEPCCER204>.

## **Comments to Research Review**

The Effective Health Care (EHC) Program encourages the public to participate in the development of its research projects. Each research review is posted to the EHC Program Web site or AHRQ Web site in draft form for public comment for a 3-4-week period. Comments can be submitted via the Web site, mail, or email. At the conclusion of the public comment period, authors use the commentators' submissions and comments to revise the draft research review.

Comments on draft reviews and the authors' responses to the comments are posted for public viewing on the Web site approximately 3 months after the final research review is published. Comments are not edited for spelling, grammar, or other content errors. Each comment is listed with the name and affiliation of the commentator, if this information is provided. Commentators are not required to provide their names or affiliations in order to submit suggestions or comments.

The tables below include the responses by the authors of the review to each comment that was submitted for this draft review. The responses to comments in this disposition report are those of the authors, who are responsible for its contents, and do not necessarily represent the views of the Agency for Healthcare Research and Quality.



Commentator & Affiliation	Section	Comment	Response
Reviewer #1:	General	The key questions were mostly excellent but there were a few places where the report purpose and application (ie dissemination potential) seemed limited. For example, lines 52-55 state that "evaluation of the policies and programs... enhancing the evidence base for our ability to adapt, scale and disseminate those that are found effective." Yet, the report was designed to understand the data sources available and methods. It was not designed to evaluate comparative effectiveness, which is what is needed for scaling up and dissemination. Nonetheless, the KQs were clearly laid out and an excellent process was explained and well documented. It is not clear how clinically meaningful the results are, but again that was not really part of the KQs so a bit unfair to ask that question. As stated and inferred on page 58, lines 15-20, much work remains to be done to advance methodology (which currently has many weaknesses and high risk of bias) before population and clinical advances can be made.	We agree that the intention of the report was not to evaluate effectiveness, and worked to remove any reference to dissemination potential or effectiveness of policies and programs. In lines 52-55, 58 in the Intro, we clarified that the "evaluations" referred to are the published evaluations, and not our review.
Reviewer #2:	General	I understand that this report is intended to serve as a background document for the workshop on December 5 and 6. Overall, it is meaningful in that it provides a detailed catalogue of the relevant studies and their characteristics as prompted by the Key Questions. It is somewhat less meaningful in terms of information about the quality of what has been done.	The report's KQ1-2, 3 focused on description. KQ4-5 provided bias evaluations and more critique. The workshop and its report will likely lend itself for more comprehensive recommendations.
Reviewer #2:	General	I found the key questions to be somewhat unclear as to how much evaluation of current methods was in the scope of the review. It is largely descriptive with the exception of the application of the Risk of Bias assessments. Perhaps addition of a clear statement of what is NOT in the scope of this review would be helpful - not in terms of topic or study exclusion criteria but in terms of commenting on the appropriateness or limitations of various measures or designs or looking at the influence of study methods on the ability to determine effectiveness.	Thank you. We included a comment that the review does not address effectiveness.
Reviewer #3:	General	This report will be very beneficial for the field of obesity prevention and control, as we need more natural experiment studies to improve the state of knowledge, especially in terms of understanding causal relationships and/or dose-response relationships.	Thank for your comment.
Reviewer #3:	General	Some parts of the report can be better structured to make it more user friendly for researchers seeking information about potential data sources or methodological advances for their research. I will discuss some specific suggestions in the following sections.	Thank you.

Source: <https://effectivehealthcare.ahrq.gov/topics/obesity-research-methods/systematic-review>

Published Online: December 2017



Commentator & Affiliation	Section	Comment	Response
Reviewer #3:	General	Some terms need brief definitions for better clarity.	Thank you.
Reviewer #3:	General	Would suggest changing the report title to "Methods for Evaluating Natural Experiments in Obesity Prevention and Control: A Systematic Evidence Review".	Thank you for your comment—the report title is pre-defined by AHRQ and NIH.
Reviewer #3:	General	Some sections of the report are a bit difficult to navigate, especially when moving from the body text to appendices or references. Wonder if the structure can be improved for easier access. Hyperlinks will also help.	Thank you for your comment. AHRQ welcomes hyperlinks. We have added reference hyperlinks and will work with AHRQ to link Appendix call-outs to the appendices
Reviewer #3:	General	There are typos and grammatical errors needing to be fixed.	Thank you for your comment, these were addressed.
Reviewer #4:	General	I have attached a pdf of the document that contains my comments. There were quite a few typos and they are noted in comment bubbles. Many of the points below are also noted in the comment bubbles.	The typographical errors have been addressed. The more detailed comments appear in the comment table.
Reviewer #4:	General	The key questions are explicitly stated. I do not recall that the authors explicitly defined the target population or audience. That would be helpful	I am not sure if you mean the audience for the report. It is meant for a research audience.
Reviewer #5:	General	A key point is definition. What the authors refer to as 'natural experiments' are, in my opinion, a broader class of studies, most of which are not properly called experiments, but are better termed quasi-experiments. I would term natural experiments as studies in which subjects are randomly assigned to levels of the independent variable, but the levels chosen and/or randomization process is not controlled by the researchers (e.g., <a href="https://www.ncbi.nlm.nih.gov/pubmed/22329049">https://www.ncbi.nlm.nih.gov/pubmed/22329049</a> ). In contrast, a quasi-experiment is one in which assignment to levels of the independent variable is not random, but the study involves an intervention and a control condition (e.g., <a href="https://www.ncbi.nlm.nih.gov/pubmed/27070635">https://www.ncbi.nlm.nih.gov/pubmed/27070635</a> ). Clarifying this distinction is important.	We agree that there is a great deal of debate about the definition of natural experiments. We likely included both quasi-experimental and natural experiment studies using the MRC definition. For the purposes of this review, we kept the definition broad and then described the study designs in great detail in Key Questions 4 and 5.

Source: <https://effectivehealthcare.ahrq.gov/topics/obesity-research-methods/systematic-review>

Published Online: December 2017



Commentator & Affiliation	Section	Comment	Response
Reviewer #5:	General	The authors point out the poor quality of the methods used in most of the studies they analyzed. It would have been valuable to report on the funding sources (Govt; Foundation; Industry; none or unspecified) of those studies and whether some funding sources were associated with better quality than others. There is an emerging literature on this suggesting that government and industry funded studies tend to be strongest, followed by foundation studies, and then by unfunded studies.	We agree, but we did not collect the funding source of each study. This is more commonly done in drug trials and we did not consider it for these types of studies.
Reviewer #5:	General	There are substantial concerns about the use of 'spin' in studies in general (see work of I. Boutron). In this reviewer's experience, the most egregious 'spinners' are authors of papers evaluating childhood obesity community or school-based treatment or prevention programs. The extent to which the conclusions of the studies (and also the press releases about the studies) deviate from the results of the studies.	This is an interesting issue, but it was beyond our scope of work to critically assess the spin in the reviewed studies. We thought it was important to stay focused on an objective assessment of the risk of bias.
Reviewer #5:	General	The extent to which each of the studies considered were pre-registered (or registered at all) in a trial registry should be assessed and reported.	We did not assess whether the studies were pre-registered in clinicaltrials.gov or another registration system. That was beyond our scope of work. Registration would only be expected for clinical trials and not for many of the studies that were included in the review.
Reviewer #5:	General	Deviations between initial study protocols and registrations in terms of analytic plans and what is labeled primary outcomes are common (see: <a href="http://onlinelibrary.wiley.com/doi/10.1111/cob.12199/abstract">http://onlinelibrary.wiley.com/doi/10.1111/cob.12199/abstract</a> as well as letters to the editor in response). This should be assessed for the studies considered.	As indicated above, we did not assess whether studies were registered, and would not expect to find registered information for many of the studies included in the review.
Reviewer #5:	General	Inappropriate analyses of these types of studies which neglect or do not properly take clustering into account are common. See: <a href="https://experts.umn.edu/en/publications/best-but-of-forgotten-practices-designing-analyzing-and-reportin">https://experts.umn.edu/en/publications/best-but-of-forgotten-practices-designing-analyzing-and-reportin</a> and <a href="https://www.ncbi.nlm.nih.gov/pubmed/27028280">https://www.ncbi.nlm.nih.gov/pubmed/27028280</a> ). The extent to which this occurs should be assessed and described.	We did describe the clustering and analytic approaches in KQ5.
Reviewer #6:	General	Strengths of this report include: a) Overall organization and logic are excellent  b) Overall very clearly written with the exception of minor typographical errors  c) An especially thoughtful approach to Key Question 6 given the findings for key questions 1-5.	Thank you for your comments.

Source: <https://effectivehealthcare.ahrq.gov/topics/obesity-research-methods/systematic-review>

Published Online: December 2017



Commentator & Affiliation	Section	Comment	Response
Reviewer #6:	General	The key messages are exquisitely aligned with the stated purpose/aim. Of note, the term physical activity is standard but "dietary activity" is not. Consider using "diet change" or "diet intake" or "diet composition"?	We made some changes to be clear about the types of dietary behavior considered.
Reviewer #6:	General	With regards to the abstract, was there a search of gray literature at all?	Instead of conducting a laborious de novo gray literature review, we used a very recent review that conducted grey literature searches, and we used hand searching of references in eligible articles to look for relevant gray literature.
Reviewer #6:	General	Recommend that the term "data source" be used in the methods section only. In the results section, it should be introduced differently. For example, "We characterized studies of programs and policies in obesity prevention and control in terms of their data sources, data linkages, measures reported, study designs, and analytic approaches, and identify needed methodological advances. With regards to data sources ...geographic allocation."	We added this into the Results.
Reviewer #6:	General	Consider a concise mention of the subgroups considered (as listed in Figure 1).	Additional text was added in the Population section of the introduction.
Reviewer #6:	General	Minor comments, typos: "this" p1, line 3; "neighborhoods" p8, line 9 "interventions" p8, line 18	These has been corrected
Reviewer #7:	General	The key questions are clearly stated. The report does an excellent job of consolidating a large amount of information from the literature. The aim of the study was not to determine the effectiveness of different interventions, instead the review has identified different ways in which natural experiments have been conducted in the area of obesity. The identification of available data sources used in the studies and the detailed information that they provide about how they structured and implemented the literature review (including the entry form) will help further future research in this area and other areas.	Thank you.

Source: <https://effectivehealthcare.ahrq.gov/topics/obesity-research-methods/systematic-review>

Published Online: December 2017



Commentator & Affiliation	Section	Comment	Response
<b>Reviewer #1:</b>	Clarity and Usability	I don't think the report will provide much to policy or practice since this was a methods report and there was no comparative effectiveness evaluation of the types of interventions (both RCTs and natural experiments). Rather, it was designed to describe the types of data and data sources. It succeeded at answering the KQs. Report was extremely well organized and main results were clear. Tables and appendices were outstanding. Needs a stronger and more specific section on future research needs and opportunities.	Thank you. We worked to make the Future Research section focused on the aim of the review and not to have a broad scope on recommendations for all obesity policy/programs. We tried to make it more targeted and specific per your comments.
<b>Reviewer #2:</b>	Clarity and Usability	The report is well structured, following the order of the key questions, although the order of the key questions could be different (and I see that the presentation at the upcoming workshop actually takes them in a different order). Regarding the main points, below I include some comments on the key messages that appear on page ii. These points reflect my impression that the purely descriptive messages are not always sufficiently informative.	Thank you. We made edits in response to the specific comments about the key messages.
<b>Reviewer #2:</b>	Clarity and Usability	Finally, I think a few case studies in boxes within each chapter would be very useful. It is difficult to picture what is actually being done, or could be done, in some of these studies although there are examples in the text. More detail on some key natural experiments of experimental studies would allow illustration of some of the key points, especially where certain studies are considered to be exemplars.	We are working within a format specified by AHRQ and did not include separate boxes for case studies. Instead, we provided examples of 'exemplars' in each of the sections to allow the reader to "envision" what these studies are like.
<b>Reviewer #3:</b>	Clarity and Usability	This report provides a lot of useful information.	Thank you for your comment
<b>Reviewer #3:</b>	Clarity and Usability	It will be more helpful if all the concepts/terms have a brief explanation when they are first mentioned.	We are working to make the report as clear as possible.
<b>Reviewer #3:</b>	Clarity and Usability	It will be helpful to structure the report in a way that is easier to navigate for particular information (e.g. from the body text to one particular appendix, look for more details about one particular methodology in terms of definition and strength and weakness, or for one particular survey instrument in terms of reliability, validity and target population).	We are presenting the report in a different order for the P2P presentations, but the structure is dictated by AHRQ guidance.
<b>Reviewer #4:</b>	Clarity and Usability	The limitations and future research are described. The future research directions are described in a way that is easily translated into new research. The findings are described in relation to what is already known and the implications of the major findings are stated.	Thank you for your comment.

Source: <https://effectivehealthcare.ahrq.gov/topics/obesity-research-methods/systematic-review>

Published Online: December 2017



Commentator & Affiliation	Section	Comment	Response
Reviewer #4:	Clarity and Usability	The report is organized and the main points are clearly presented. The conclusions contribute new information and understanding. Overall, I think the report is well-done.	Thank you for your comment.
Reviewer #6:	Clarity and Usability	The logic and flow are excellent. Overall, the writing is clear. See comments above.	Thank you.
Reviewer #7:	Clarity and Usability	The report is well structured. The appendix tables provide detailed information that other researchers will be able to draw on in planning a study.	Thank you.
Reviewer #4:	Table of Contents	insert "activity"?	The error has been corrected.
Reviewer #4:	Table of Contents	spelling	The error has been corrected.
Reviewer #4:	Table of Contents	spelling	The error has been corrected.
Reviewer #2:	Introduction	The section on challenges in evaluating programs, policies etc. (page 2, lines 25 and following) could cite some of the articles or reports that describe the issues in evaluating complex health problems more generally.	We felt that obesity is uniquely complex and elected to focus on the articles that discuss obesity as a complex health problem with environmental and community influences. Other issues could include substance abuse and violence as public health problems that are complex, but it is difficult to draw connections to obesity for identifying solutions.
Reviewer #2:	Introduction	It would be good for readers to see that many of these issues are not unique to obesity and also that some of the recommendations that emanate from this report have a solid basis in other domains where natural experiments are important.	We added references to tie in other fields where natural experiment evaluations are common.
Reviewer #3:	Introduction	Page 14, lines 10-12: "The definition they proposed for natural experiment studies was, "methodological approaches to evaluating the impact on health or other outcomes of such events." This sentence does not make much sense to me, as whether a study is a 'natural experiment study' depends on the specific study design.	We included a wide range of study designs and did not define 'natural experiment' by design. There continues to be debate about the definition of a natural experiment so we broadly defined it. We worked to make this sentence clearer in the Introduction,
Reviewer #3:	Introduction	Page 15, lines 25-26: Having a brief explanation about "PICOTS typology" would be helpful.	We added clarification.

Source: <https://effectivehealthcare.ahrq.gov/topics/obesity-research-methods/systematic-review>

Published Online: December 2017



Commentator & Affiliation	Section	Comment	Response
Reviewer #3:	Introduction	Page 16, line 43: should "purchasing behavior" here and also in other places be changed to "food purchasing behavior"?	We agree and changed it to be clear.
Reviewer #3:	Introduction	Page 17, sections about "Outcomes and Measures": I am wondering if the team also searched for other obesity related outcomes such as body fat percentage, and whether relevant studies should also be included in the review	We did not include body fat percentage as an outcome of interest for this review. The rationale is that it is not a common outcome measure in the natural experiment studies we focused on.
Reviewer #3:	Introduction	Page 18, line 25: For "human development index", it would be helpful to have a brief explanation here. The report actually has an explanation later on, but it is better to explain it when it is first mentioned.	We included the explanation here as well.
Reviewer #6:	Introduction	This section is well written. On page 12, provide more information on the rationale for identifying key questions to inform the "Pathways to Prevention" Workshop. For example, will this systematic review be the foundational document on which their work is built?	Yes, we clarified the purpose.
Reviewer #6:	Introduction	For Table 1, when referring to PICOTS, use "comparator" or "comparison" consistently throughout the document.	We tried to be more consistent about this.
Reviewer #7:	introduction	The report includes a brief introduction to the challenges in conducting obesity research and the key questions that the review seeks to answer.	Thank you.
Reviewer #4:	Introduction	definition of a natural experiment The MRC definition given in quotes does not seem like a definition. Is that literally all there is to their definition? If so, I think someone could write a better one. For example, what does "such events" refer to?	We described how we operationalized the definition to make it clearer.
Reviewer #4:	Introduction	Is the next sentence a part of the definition? Because this adds some important ideas. In any case, I think this part could be written a bit clearer because the definition in the appendix is much clearer and this is really important for understanding the manuscript.	We changed the wording to make it clear.

Source: <https://effectivehealthcare.ahrq.gov/topics/obesity-research-methods/systematic-review>

Published Online: December 2017



Commentator & Affiliation	Section	Comment	Response
Reviewer #4:	Introduction	<p>“The United Kingdom’s (UK) Medical Research Council (MRC) recently released guidance to assist researchers in conducting and evaluating the rigor of natural experiment designs.<sup>37, 38</sup> The definition they proposed for natural experiment studies was, “methodological approaches to evaluating the impact on health or other outcomes of such events.””</p> <p>This doesn't seem like a definition. Is that literally all there is to their definition? If so, I think someone could write a better one. For example, what does "such events" refer to?</p>	We removed this confusing phrase.
Reviewer #4:	Introduction	<p>“According to the MRC, the key features of a natural experiment are that (1) the intervention is not undertaken for the purposes of research, and 2) the variation in exposure and outcomes is analyzed using methods that attempt to make causal inferences.”<sup>37, 38</sup>”</p> <p>Is this part of the definition? Because this adds some important ideas. In any case, I think this part could be written a bit clearer because the definition is really important for understanding the manuscript.</p>	We worked to make it clearer.
Reviewer #2:	Introduction	Was % overweight or % obese included as an outcome?	Yes, these BMI categories were included as outcomes, and we clarified in Table 2.
Reviewer #2:	Introduction	Table 2, line 37 -- The line that says "specific dietary macronutrients related to obesity" then refers to food categories. Macronutrients are calories, protein, fat, and carbohydrate.	<p>We clarified that we defined macronutrients according to this reference: <i>Dietary Reference Intakes for Energy, Carbohydrate, Fiber, Fat, Fatty Acids, Cholesterol, Protein, and Amino Acids (2002/2005)</i>. This report may be accessed via <a href="http://www.nap.edu">www.nap.edu</a>.</p> <p>The macronutrients considered were carbohydrate, cholesterol, fiber, fat, fatty acids and protein.</p> <p>Total caloric intake was included as an outcome of interest and not considered as a macronutrient.</p>

Source: <https://effectivehealthcare.ahrq.gov/topics/obesity-research-methods/systematic-review>

Published Online: December 2017



Commentator & Affiliation	Section	Comment	Response
Reviewer #1:	Methods	Diet and PA measures were categorized consistent with NCCOR, which lends a good framework. However, was not sure why studies were excluded that measured access to healthy food and walkability (page 6, lines 51-57)	Access to healthy food and walkability are distal outcomes and cannot be directly correlated to changes in diet and physical activity.
Reviewer #2:	Methods	I thought the inclusion and exclusion criteria and the PICOTS specifications were appropriate. My main question about methods related to the definition of "built environment". On page 14, around line 14, it is described as referring to: transit, park, or outdoor spaces" in a way that implies a relationship to physical activity but not food environments. In later places in the report it becomes evident that food environments (e.g., new supermarkets) were also included (which is correct in my opinion). However, this should be clarified in the methods with respect to how the definitions of physical activity environment or food and beverage environments were distinguished from their respective built environments.	Thank you for pointing this out. In table 3, we have expanded the description of "built environment": Physical and built environment (e.g. transit, park, other outdoor spaces, farmers' markets, new supermarkets).
Reviewer #3:	Methods	Page 20, line 17: "Ten index articles". It is unclear to me what these 10 articles are and how they are used.	The text has been revised to read: "Ten index articles identified by internal experts as applicable to this study were used during the search development. Terms used in the titles and abstracts, as well as relevant MeSH headings were identified in the index articles and used to develop the main search strategy. After the search strategy was developed, it was tested to determine if the index articles were captured."
Reviewer #3:	Methods	Page 20, 2nd paragraph: it will be helpful to briefly describe the focus of the search and key words here.	We added the following wording to the Methods: "This search focused on identifying studies addressing obesity and behavioral changes impacting obesity."
Reviewer #3:	Methods	Page 21, line 28: 3rd item of the bulleted list, the "or" in the middle should be changed to "and".	We believe "or" is correct in this case..
Reviewer #3:	Methods	Page 21, line 49: Not sure why "Meeting abstract" is there since this is discussing reasons for excluding "full-text articles"	During the screening process (at the title/abstract level) it is not often clear if the abstract we are viewing is a meeting abstract; this is the reason for this exclusion criteria



Commentator & Affiliation	Section	Comment	Response
Reviewer #3:	Methods	Page 25, lines 6-15: The categories are a bit confusing to me, because the last item “Physical and built environment” can overlap with earlier items such as “Physical activity environment”, “Food and beverage environment, “Healthcare environment” or “Work environment”	We agree. When this information was abstracted from the included studies, the abstractors were instructed to identify all applicable environments in order to be inclusive and to not require a “one or the other” approach.
Reviewer #3:	Methods	Page 29, lines 10-11: I cannot find “Appendices I1 and I2”.	Appendix I is present in the appendix file.
Reviewer #4:	Methods	The inclusion/exclusion criteria are justifiable and the search strategies are explicitly stated and logical. The outcome measures are appropriate. There are not any statistical methods used beyond counts and percentages and that is appropriate.	Thank you for your comment.
Reviewer #6:	Methods	In the methods section, consider commenting on why no gray literature was searched. I don’t know that it was necessary but it is sometimes done in a systematic review of this nature. Was it discussed by the team?	Instead of conducting a laborious de novo gray literature review, we used a very recent review that conducted grey literature searches, and we used hand searching of references in eligible articles to look for relevant gray literature.
Reviewer #6:	Methods	In the subsection “Results from literature search” p26 What are “message environments”? Also applies to Table 4.	Text and footnotes were added to clarify the definition of message environments.



Commentator & Affiliation	Section	Comment	Response
Reviewer #6:	Methods	The inclusion and exclusion criteria are clearly stated and justifiable.  The search strategy is clearly presented.  Outcomes are defined clearly.	Thank you.
Reviewer #7:	Methods	This is a very large literature. The authors made thoughtful choices about inclusion and exclusion criteria that reduced the review to a clear focus and more manageable scope.	Thank you.
Reviewer #2:	Methods	line 17 –The sentence “additionally we identified 31 non-governmental programs was unclear. There were many numbers in the prior paragraph and I wasn’t sure what the 31 should be added to.	The general results section has been revised to be clear.
Reviewer #2:	Methods	line 24 –the sentence beginning “Some of the natural experiment studies included evaluations of programs that were not originally intended for research” was unclear to me. I think most natural experiments are evaluations of initiatives that are not intended for research. So, here, are you making the distinctions between policies and programs?	We revised this to read: “We included evaluations of programs that were not originally intended for research, such the CDC’s HealthMPowers program to improve physical activity and nutrition in the school setting and Jamie Oliver’s cooking skills program.”
Reviewer #4:	Methods	or?	The error has been corrected.
Reviewer #1:	Methods	The PICOTS framework was well laid out and well described. However, a few things that I thought were a little too strict/limited were excluding: school studies based in one school only; studies without a reference for "validation" of diet and PA (validation is a bit of a misnomer and sends the perhaps incorrect scientific message that just because an instrument is "validated" often vs. another form a self-report that it is somehow "OK is not really true; any studies using SOPARC/SOPLAY. Search strategies were excellent.	Thank you for your comment. We appreciate your concern regarding the strict inclusion/exclusion criteria, We added the following text: “We did not include studies using SOPARC/SOPLAY (System for Observing Play and Recreation in Communities/System for Observing Play and Leisure Activity in Youth). These studies typically only record numbers of individuals engaged in certain activities, and do not measure time and duration of any of the activities.”
Reviewer #1:	Results	The appendices and tables were outstanding. They were easy to follow and clearly laid out. Well labeled and organized. This certainly made the review easier!	Thank you for your comment
Reviewer #1:	Results	Key messages are excellent - brief and to the point. For the KQs identified, the messages are applicable to the KQs.	Thank you for your comment

Source: <https://effectivehealthcare.ahrq.gov/topics/obesity-research-methods/systematic-review>

Published Online: December 2017



Commentator & Affiliation	Section	Comment	Response
Reviewer #2:	Results	The results are quite detailed and systematically described for each key question. The Tables and Figures need clarification in some cases, e.g., with footnotes or other details that would enable them to stand alone. In some cases, I found the narrative hard to follow, although I was usually able to figure out what was being said by reference to the tables and figures.	The general results section has been revised to be clear.
Reviewer #2:	Results	I think the review would be more useful if each Chapter for the key questions would begin with a description of what the issue is that is being addressed by that key question -- a preamble to contextualize what will follow rather than going right into the abstraction approach. Readers will have variable knowledge of the issues to be considered in relation to each question and, therefore, will not necessarily understand why it is important to examine certain details or the import of the findings. This preamble would provide a reference point for recognizing the quality issues that are implied but not explicit in the results that are presented.	We have added a preamble to the beginning of each section that helps contextualize the chapter.
Reviewer #2:	Results	For example, some standards or reference points are implicit (e.g., page 28, line 42—saying that most studies had direct measurements implies that these are better than self-report, and most readers will know this. However, there are other issues where the authors are actually looking for certain elements for certain reasons that might be less obvious to readers.	This has been addressed in the added preamble
Reviewer #2:	Results	Of interest, no statements are made about about the quality or objectivity of diet or physical activity measures, whereas it is well known that high quality measures of these outcomes are quite difficult to obtain. Only counts of how many times each was measured are given for these. Also, no statements are made about the relative value or appropriateness of weight, BMI percentiles or BMI z scores for various purposes. It might be useful to add a statement that natural experiments have the same challenges that other obesity prevention and control studies have in terms of obtaining valid and reliable measures of dietary intake, physical activity, and weight status and should aim to find practical ways to obtain good quality measures (or that the assessments of these variables in ongoing data systems should be of the highest possible quality).	Thank you. We did not describe the hierarchy of quality or objectivity of the various diet and physical activity measures as it is outside the scope of this report (likewise for the weight and BMI scores).  We agree with your statement and have added it to the discussion section.

Source: <https://effectivehealthcare.ahrq.gov/topics/obesity-research-methods/systematic-review>

Published Online: December 2017



Commentator & Affiliation	Section	Comment	Response
Reviewer #3:	Results	Page 30-36 (Key question 1): Hoping to see some discussions about quality of data sources, even if it is difficult to come up with very comprehensive assessments.	As mentioned in the methods section, evaluating the quality of any of these data sources is out of scope of this project (page 31 of PDF); however, we have elaborated on the fact that data quality should be examined before use (page 32 of PDF).
Reviewer #3:	Results	Page 36, line 23: It is not very clear to me what the “objective approach” means here. A brief explanation and/or a list of samples would be helpful.	we added some wording to clarify this.
Reviewer #3:	Results	Pages 39-47 (Key question 3): When discussing different types of measures, it will be nice to have a summary for the strength and weakness of different measures. For survey instrument's, it will be nice to have a summary about their reliability, validity, target population, etc.	We agree that such a table would be very helpful. Unfortunately, this type of table is not readily available and putting this together is outside the scope of this report, requiring additional resources not at our disposal
Reviewer #3:	Results	Pages 48-58 (Key questions 4 & 5): When discussing study designs and research methods, it will be helpful to have a clear list of different methodologies (e.g. difference-in-differences, regression discontinuity, instrumental variables), including brief definitions, and strength and weakness of each method.	We have updated a table in the report (table 11) so that it clearly defines each method and describes the assumptions and risks of bias of interest.
Reviewer #3:	Results	Page 51, lines 36-37: Need a bit more information about the Newcastle-Ottawa Scale for better clarity.	More details about the Newcastle-Ottawa Scale (NOS) have been added to the “Alternative Risk of Bias Assessment” section.
Reviewer #3:	Results	Page 53, lines 36-37: Not sure why only a subset of, instead of all, natural experiment studies was analyzed using Newcastle-Ottawa Scale.	After completing EPHPP assessments for all studies, a reviewer suggested that we compare a subsample using the NOS assessment because this tool is mentioned in the MRC report on Natural Experiments as one that may be useful for evaluating natural experiments. However, as explained in our methods section, NOS was not an ideal tool to apply to all studies in this review (mainly it cannot be used on RCTs). Additionally, as we reviewed 295 studies for this reviewing, completing two risk of bias assessments on all studies was not feasible.
Reviewer #3:	Results	Page 55, Figure 8 and Page 57, Figure 9: Since 3 levels of bias (strong, moderate and weak) add up to 100%, stacking them vertically may make the image more readable and informative.	Thank you for this suggestion. We would prefer to keep the graphs as is because some find stacked bar graphs difficult to read.

Source: <https://effectivehealthcare.ahrq.gov/topics/obesity-research-methods/systematic-review>

Published Online: December 2017



Commentator & Affiliation	Section	Comment	Response
Reviewer #3:	Results	Page 61, Key question 1: It will also be helpful for data sources with spatial attributes to use consistent spatial unit of analysis when possible, so that they can be easily linked later on. For example, census data and traffic data may come in with different spatial units, which makes linking data difficult.	Thanks for the comment. This can be addressed in our Table, but not in the universe of data sources except as a recommendation in this report. We have added this to the Discussion.
Reviewer #6:	Results	The authors are to be commended as the results section is clear and concise considering the amount of data analyzed. The key messages are explicit and applicable. The tables and figures are clear and informative. The studies included seem reasonable and exhaustive. I am unaware of any studies that should have been included that were not.	Thank you.
Reviewer #7:	Results	The authors keep the main results short and provide extensive appendices for those seeking more information. The key information is clear from the summaries, figures, and tables.	Thank you.
Reviewer #4:	Results, 28	Table 4, or worksite?	The error has been corrected
Reviewer #2:	Results, KQ1	line 19 (although 52% is technically a majority, I would say “about half” here.	Thank you. The wording has been changed to “About half of the studies.”
Reviewer #2:	Results, KQ1	line 41 – add the numbers to the percent in this bullet, as in the one below.	We have edited the text to match the other bullets.
Reviewer #2:	Results, KQ1	Figure 3 was a bit difficult to follow in the “sharable” column. I think I figured out that the first box seems to be a count of studies and the boxes below are about the data sources you found within those studies. That could be made clearer but revising the wording in the boxes.	We have the wording ‘Studies’ in the PRISMA diagram.
Reviewer #2:	Results, KQ1	line 23, I did not understand the footnote to Figure 4. Looking at the text I think it refers to the three lighter grey bars but not the darker bar at the top of each section. That could be clarified in the footnote.	Done. Added some wording.
Reviewer #2:	Results, KQ1	where would the NHANES data fall in Figure 5. Public health? A lot of research is done with the NHANES data, but I wasn’t sure if you classified it as designed for research. I looked for a categorization of the data systems to match those in Figure 5 but could not find it. If it is in the Appendix perhaps a specific reference to where in a footnote to the Figure could be added.	NHANES is categorized as “Public Health.” Note that the ‘original/primary’ purpose is depicted. We decided not to make all of the coding details available in an appendix.

Source: <https://effectivehealthcare.ahrq.gov/topics/obesity-research-methods/systematic-review>

Published Online: December 2017



Commentator & Affiliation	Section	Comment	Response
Reviewer #2:	Results, KQ1	I could not get a sense of whether the data systems are linked permanently or were only linked for a specific purpose at one point in time and would have to be linked by new users or for new rounds of data (e.g., linking future WIC participants to some of the ongoing data sets on child outcomes).	It depends on the study/project, but they were mostly linked temporarily for the purpose of the research.
Reviewer #2:	Results, KQ3	Table 6 –It seems that some studies had more than one type of weight measure. Perhaps that could be mentioned in the text, footnote, or both.	Thank you. We have clarified it in the footnotes of the tables.
Reviewer #2:	Results, KQ3	line 30 – I think that “of the 35 experimental studies....” Should say “of the 42 experimental studies”	In this context, the 35 studies refers to those that reported BMI-z score or BMI percentile only. In the table, the count also includes studies reporting other weight outcomes (e.g., BMI or just weight). We have added additional wording to make this clear: “Of the 39 experimental studies reporting on children’s BMI-z or BMI percentile.....”
Reviewer #2:	Results, KQ3	As noted above, this section would be more useful if there were more comments about the desirability (methodologically) of various measures. This become particularly important for the dietary section. On page 31, (and from the relevant appendix), I did not see a clear distinction between full food frequency questionnaires and “screeners” that focus on a particular set of foods and other brief dietary questionnaires. In the appendices I did see some descriptions of instruments in conjunction with entries for particular studies but these seemed to be incidental.	These measures are categorized based on the NCCOR Measures Registry. FFQs gather information about the frequency with which different foods and drinks are consumed over a certain period of time. While screeners may also ask about the frequency of intake of certain foods and beverages, they are brief and quick to complete. Definitions were added to the glossary (Appendix B).
Reviewer #2:	Results, KQ3	Table 7 – The categories for the columns could be more specific. I suggest differentiating full food frequency questionnaires from screeners and also indicating what the other types of questionnaires are.	Thank you. We took these categories from the NCCOR Measure Registry.
Reviewer #2:	Results, KQ3	There are many types of electronic monitors, with pedometers now being the low end approach. Some comment on the relative importance or utility of these measures would be helpful.	Thank you. We did not distinguish between the various electronic monitoring devices and did not describe the different rankings of the electronic monitoring devices as it is outside the scope of this report.



Commentator & Affiliation	Section	Comment	Response
Reviewer #2:	Results, KQ3	re Key Question 3 – “more detail about the recommendation to standardize approaches to collecting obesity-related outcomes” and “consistent use of validated outcomes”. Does this refer to weight and BMI measures alone or does it include the challenges related to dietary and physical activity assessments. As I already noted, quality issues related to diet and PA were not discussed in the review (although types of measures were described) but I think the quality issues of these measures, which are not unique to natural experiments but perhaps particularly challenging in natural experiments, really need attention in both primary and secondary data.	In the process of evaluating studies for this systematic review, we noticed the use of various measures for diet and PA, and hence feel that standardizing measures would be important. However, we did not delve into the details of quality issues among these measures because this was outside the scope of this review.
Reviewer #4:	Results, KQ3	<p>“The majority (n=133 studies) of the 171 US studies evaluated governmental programs or policies at the local, state/regional, or federal levels. Of these, we identified 111 unique policy or program evaluations (Appendices I1 and I2). Sixty of these are studies of named programs, and 40 are evaluations of policies.”</p> <p>Is the 111 out of the 133 or of the 171? This is getting really confusing. Similarly, is the 60 out of the 111?</p>	This section has been revised for clarity
Reviewer #4:	Results, KQ3	spelling	The error has been corrected
Reviewer #4:	Results, KQ3	Table 6 is confusing	Thank you for informing us. We have added more to the footnotes to help clarify the table headings.
Reviewer #4:	Results, KQ3	Experimental studies reporting BMI in children section - How come these numbers add up to 42 instead of 35? Did some studies measure in more than one way? This happened in the previous section too.	You are correct, several studies used more than one method to measure outcomes. We have clarified this in the Table footnotes
Reviewer #4:	Results, KQ3	Table 6, I find this table confusing.	We have added more to the footnotes to help clarify the table headings.

Source: <https://effectivehealthcare.ahrq.gov/topics/obesity-research-methods/systematic-review>

Published Online: December 2017



Commentator & Affiliation	Section	Comment	Response
Reviewer #4:	Results, KQ3	<p>“Of the 35 experimental studies reporting on children’s BMI, 35 studies reported the outcome based on direct measurement from trained staff, one study used an EHR, four studies used selfreported data, and two studies used other measures (i.e. body composition analyzer).”</p> <p>How come these numbers add up to 42 instead of 35? Did some studies measure in more than one way? This happened in the previous section too.</p>	You are correct, several studies used more than one method to measure outcomes. As stated in a previous reply, we have clarified this in the Table footnotes
Reviewer #4:	Results, KQ3	Did some studies use more than one measure of dietary behavior?	Yes, this is similar to the BMI weight sections and the physical activity measures. We have clarified this in the Table footnotes
Reviewer #4:	Results, KQ3	<p>“Of the 44 experimental studies (RCTs and non-randomized controlled trials) reporting on children’s diet, 15 reported the outcome based on 24-hour recall, 24 used a food frequency”</p> <p>Some studies must have used more than one of these? Same for the next section?</p>	Yes, this is similar to the BMI weight sections and the physical activity measures. We have clarified this in the Table footnotes
Reviewer #4:	Results, KQ3	Table 8 - why include the GIS column is there are no entries	We included GIS as it is a category listed in the NCCOR Registry from which we derived the other categories of measures for physical activity.
Reviewer #6:	Results, KQ3	With regards to Chapter 5, KQ3: P39, line 23 (and in other places), what is meant by fast food frequency? The frequency with which one visits a fast food restaurant or the frequency that fast food is consumed?	We have corrected this error. It should be described as “fast food intake” which refers to the consumption of fast food, which is defined differently by each study.
Reviewer #2:	Results, KQ4	Was there any consideration of the ability of study designs to ensure that those in the population assessed were exposed to the policy or program in natural experiments, and the level at which they were exposed. “Dose” may vary within a geographically defined population, and response rates for assessments might be influenced by selection factors associated with exposure.	This is an interesting question but we did not collect this level of information. EPHPP has some basic questions somewhat related to this point, but we do not feel that they would capture the relevant aspects of this for the sorts of policies and programs of interest in this review. .



Commentator & Affiliation	Section	Comment	Response
Reviewer #2:	Results, KQ4	Statements made earlier in this section referred to whether “all confounders” had been taken into account, and I was wondering how the set of relevant confounders would be known. This is addressed by the comment about having a “Table 1”, on lines 20 and following. Perhaps this can be explained earlier.	<p>We were not able to assess whether “all confounders” were taken into account, and apologize if this was not clear. Absent an agreed upon list of relevant confounders, the best risk of bias assessment is to assess what each study presents as the relevant confounders—often through a table 1—and then assess what proportion of these were accounted for during the analysis. This is the approach taken by EPHPP. Text has been edited to make it clear that we are referring to confounders identified within the study.</p> <p>A potential area for future work in the field of obesity research is to get a better handle on the likely confounders of different programs, and a consensus regarding what are the key factors to adjust for. Efforts like this are currently underway in some areas or education research (particularly for the What Works Clearinghouse).</p>
Reviewer #4:	Results, KQ4	I do not understand the definition of controlled clinical trial as it is written here and there is not a definition given in the appendix.	We used the standard definition from the EPHPP.
Reviewer #4:	Results, KQ4	<p>“According to the EPHPP, “controlled clinical trials” were defined as an “experimental study design where the method of allocating study subjects to intervention or control groups is open to individuals who are responsible for recruiting subjects or providing the intervention.””</p> <p>I don't understand. What does "open" mean?</p>	“Open” in this definition means under the control of. In controlled clinical trials. The researchers assign participants to exposed or unexposed groups based on some factor other than randomization.



Commentator & Affiliation	Section	Comment	Response
Reviewer #4:	Results, KQ4	<p>“Difference-in-difference studies also all had a single measure pre-intervention, and 88 percent had only a single pre- or post-intervention measure. We use the term difference-in-difference for studies with multiple time points postintervention rather than interrupted time series due to the small number of time points (maximum 6 points) and the lack of formal interrupted time series methods such as autoregressive integrated moving average.”</p> <p>delete "pre- or"? Already stated that they all had only one pre-measure.</p>	Thank you for pointing out this typo - “pre- or” has been deleted.
Reviewer #4:	Results, KQ4	In general, some of the details are repeated several times, whereas other details (those mentioned above) are not entirely clear.	We feel that some details are helpful to repeat in multiple places, and have endeavored to clarify confusing aspects, as addressed in other responses to these comments.
Reviewer #4:	Results, KQ4	The characteristics of the studies are described (and presented in the appendices in more detail). Unless noted above, the figures and tables are clear.	Thank you for this comment.
Reviewer #6:	Results, KQ4	With regards to Chapter 6, KQ4: P50, line 36, should be 38% not 39%	Thank you for pointing out this discrepancy. We updated our literature review prior to completing this final version, which changed all data points in the report. The table now matches the text for this section.
Reviewer #4:	Results, KQ5	<p>From pdf comment: “As shown in Figure 10, the 18 studies (n=7 percent) in this category did not provide sufficient detail to determine whether the research team was in control of the intervention, whether the intervention was originally intended to be research, or otherwise would be included as either a natural experiment or experiment.”</p> <p>Is this n=7 or 7 percent (or neither)?</p>	7 percent. We have corrected this in the text.
Reviewer #6:	Results, KQ5	With regards to Chapter 7, KQ5: The risk of bias issues are nicely described. Consider summarizing the implications on the overall interpretation of this body of work when the natural experiments are included in the data analysis.	We have summarized the implications of this work in the discussion section.
Reviewer #4:	Results, KQ6,	spelling	This has been corrected.

Source: <https://effectivehealthcare.ahrq.gov/topics/obesity-research-methods/systematic-review>

Published Online: December 2017



Commentator & Affiliation	Section	Comment	Response
Reviewer #4:	Discussion/ Conclusion	beyond	Thank you.
Reviewer #4:	Discussion/ Conclusion	From pdf comment: “We presented the findings of this review by “natural experiment”, experimental designs, and their non-experimental designs, in order to compare within and across study design approaches, and also to avoid.....”  should it be other instead of their?	Yes, we corrected it.
Reviewer #1:	Discussion/ Conclusion	The needs for future research could emphasize more how we need objective measures of diet and PA and also measures of body composition (not just BMI) to better understand the important questions surrounding population based obesity.	Thank you. We included this observation in the Discussion as a limitation to the Body of Evidence.
Reviewer #1:	Discussion/ Conclusion	There should also be more emphasis on the need to have more diverse populations studies - many did not report on race/ethnicity and of those that did. many have combined groups that should not be combined (ie, combining all Hispanic and all Asian). There were very few studies with Native Americans. This is a research and public health need.	Thank you. We included this observation in the Applicability section of the Discussion. We also added that few studies included rural communities.
Reviewer #1:	Discussion/ Conclusion	The future research section does note the need for longer follow-up (giving Shape up Somerville as an excellent example) but should note the need for multi-level and multi-dimensional interventions. Natural experiments would also benefit from methods development since there is likely substantial confounding and that type of study design will continue to be subject to high risk of bias.	Thank you. Because we focused the future research section on the methods, we did not feel qualified to critique the intervention designs, as this was not the goal of the review. Many of these interventions were described as multi-dimensional and had multiple targets. We did highlight the need for more systems science (see below).
Reviewer #3:	Discussion/ Conclusion	Readers will benefit from a more user friendly format that directly points them to useful references in relation to data sources and/or possible methodological advances, etc.	Thank you. We will work to make it clearer in the Discussion. The P2P meeting may also provide further recommendations.

Source: <https://effectivehealthcare.ahrq.gov/topics/obesity-research-methods/systematic-review>

Published Online: December 2017



Commentator & Affiliation	Section	Comment	Response
Reviewer #3:	Discussion/ Conclusion	As mentioned above, I feel it would be helpful to provide information about (1) quality of data sources and specific sources that may be helpful for certain populations and/or geographic areas, (2) strength and weakness of different measurement methods, study designs, and methodologies.	<p>Thank you. We did comment that few studies used 24-hour recalls as an example of a standard measure of dietary recall. We did not have a process for ranking the quality of the data collection measure, outside of the bias assessments in KQ4-5. I am not aware of a tool for this. We did describe which studies used validated measures as a proxy for quality.</p> <p>Regarding the quality of data sources, we did provide an Appendix with the data sources that are highest quality – i.e. meeting criteria for a data system—and are most useful for researchers. (Appendix J)</p>
Reviewer #4:	Discussion/ Conclusion	In general, the discussion section primarily repeats what has already been stated in the chapters specific to each key question.	We tried to reduce duplication.
Reviewer #6:	Discussion/ Conclusion	The results are summarized well, and the meaning of the results are clearly stated in most cases. One recommendation is to add a summary of the implications on the overall interpretation of this body of work when the natural experiments are included in the data analysis.	We worked to be thoughtful about the implications and added more about systems science approaches.
Reviewer #6:	Discussion/ Conclusion	The future research section is clear and can guide new research. The inclusion of exemplars is good.	Thank you.
Reviewer #7:	Discussion/ Conclusion	Summary of the review and findings related to different questions are clearly stated.	Thank you.



Commentator & Affiliation	Section	Comment	Response
Reviewer #2:	Discussion/ Conclusion	lines 11 to 19. I thought it was positive that you listed several reviews of policies and programs. It seems that you are comparing your review with those reviews to show the value added by your review as being much more comprehensive than reviews that focus on single topics. However, I think those reviews looked at effectiveness, highlighting the fact that your review does not address the ways that the methodological issues you cover influence the ability to assess effectiveness. I think this is implicit in your review in that you talk about having better data and better designs but it would be great for you to connect these dots directly.	We clarified that these reviews assessed effectiveness, but we described them here because they used similar search strategies and evaluated similar articles, and some did risk of bias assessments.

Source: <https://effectivehealthcare.ahrq.gov/topics/obesity-research-methods/systematic-review>

Published Online: December 2017



Commentator & Affiliation	Section	Comment	Response
Reviewer #2:	Discussion/ Conclusion	<p>On Limitations of the Current Evidence, although you acknowledge that your review prioritized internal validity, some comments on research needs in the domain of external validity would be very helpful for what I believe is the overall purpose of the review. This might include mention of the importance of policy implementation for its ultimate effect and the need for data sets that address this.</p> <p>Finally, although you mention measurements at more than one point in time as an important advance, I also wondered if the ability to time measurements to get good baseline data before policy implementation is something that needs to be considered in data sets. From my discussions with people who do policy related natural experiments, this is a huge challenge, especially when funds are needed to collect the data and given that funding decisions might be dependent on assurance that a policy has been passed and when it will take effect.</p>	<p>These are helpful observations. We added in the need for policy implementation and generalizability (external validity) to multiple different, diverse populations.</p> <p>We added in the need for baseline assessments.</p>
Reviewer #2:	Discussion/ Conclusion	Mention of systems science tools would be useful as well. I don't think you included that.	We added a reference to systems science tools and the need for assessments of study methods to include systems science principles.
Reviewer #4:	Discussion/ Conclusion	“Our systematic review identified a heterogeneous large of natural experiment studies and data sources that have been used to estimate the effect of programs, policies, or built environment changes on obesity prevention and control.”	We have corrected the typo
Reviewer #2:	Appendices	Some appendices could benefit for footnoting that helps to understand what the cell entries mean. For example, I could not understand the race/ethnicity columns in Appendices H2 a, b, and C. There were ranges in that column (and sometime other columns) but the notation was too cryptic for me to figure out what the ranges meant.	Footnotes were added to help clarify the table contents (including data ranges). In addition, the tables were edited to be more uniform and easier to understand.



Commentator & Affiliation	Section	Comment	Response
Reviewer #2:	Appendices	I also could not understand what “standard measure” meant in H3 and H4. I suggest reviewing all appendices to make sure that they can stand alone as understandable. Footnoting can address these issues	“Standard measure” refers to the standardized method for calculating BMI, BMI-z score, and BMI-percentile as stated by the CDC. An explanation and citation to these definitions are added as footnotes in the appendices.

Source: <https://effectivehealthcare.ahrq.gov/topics/obesity-research-methods/systematic-review>

Published Online: December 2017