Lead Article

Best practices for using natural experiments to evaluate retail food and beverage policies and interventions

Lindsey Smith Taillie, Anna H. Grummon, Sheila Fleischhacker, Diana S. Grigsby-Toussaint, Lucia Leone, and Caitlin Eicher Caspi

Policy and programmatic change in the food retail setting, including excise taxes on beverages with added-caloric sweeteners, new supermarkets in food deserts, and voluntary corporate pledges, often require the use of natural experimental evaluation for impact evaluation when randomized controlled trials are not possible. Although natural experimental studies in the food retail setting provide important opportunities to test how nonrandomized interventions affect behavioral and health outcomes, researchers face several key challenges to maintaining strong internal and external validity when conducting these studies. Broadly, these challenges include 1) study design and analysis; 2) selection of participants, selection of measures, and obtainment of data; and 3) real-world considerations. This article addresses these challenges and different approaches to meeting them. Case studies are used to illustrate these approaches and to highlight advantages and disadvantages of each approach. If the trade-offs required to address these challenges are carefully considered, thoughtful natural experimental evaluations can minimize bias and provide critical information about the impacts of food retail interventions to a variety of stakeholders, including the affected population, policymakers, and food retailers.

INTRODUCTION

The 2015–2020 *Dietary Guidelines for Americans* argued that shifting to healthier eating patterns will require fostering partnerships between food producers, suppliers, and retailers to increase access to foods and beverages.[1] This strategy is informed by research indicating that low-income, certain racial/ethnic minority, and rural communities tend to have limited access to supermarkets but easier access to fast food restaurants and convenience stores.[2–4] In turn, these differences in access have been linked to dietary patterns[5] that increase the risk for poor health outcomes such as cardiometabolic disease.[6] To address disparities in access to healthy foods, a variety of policy and programmatic approaches have emerged at the local, state, tribal, and federal levels. For example, governmental approaches include enacting land use and zoning provisions that enable the presence of farmers' markets, offering healthy food financing incentives that support the construction or renovation of grocery stores, and implementing new methods for promoting healthy eating among

Affiliation: *L.S. Taillie* and *A. Grummon* are with the Carolina Population Center, Gillings School of Global Public Health, University of North Carolina – Chapel Hill, Chapel Hill, North Carolina, USA. *S. Fleischhacker* is with the Office of Nutrition Research, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, USA. *D. Grigsby-Toussaint* is with the Department of Kinesiology and Community Health, College of Applied Health Sciences, University of Illinois, Champaign, IL, USA. *L. Leone* is with the School of Public Health and Health Professions, Department of Community Health and Health Behavior, University at Buffalo, Buffalo, New York, USA. *C. Caspi* is with the Department of Family Medicine and Community Health, University of Minnesota, Minneapolis, Minnesota, USA

Correspondence: *C.E. Caspi*, Department of Family Medicine and Community Health, University of Minnesota – Minneapolis, 717 Delaware St SE, Minneapolis, MN 55414, USA. E-mail: cecaspi@umn.edu.

Key words: food retail, natural experiments, nutrition policy, policy research.

doi: 10.1093/nutrit/nux051

participants in federal food and nutrition assistance programs.[7–11] Such government-led initiatives have been accompanied by public–private partnerships and voluntary retailer commitments to improve the retail food environment, particularly in underserved communities.[12] However, knowledge gaps remain about whether, for whom, and how these intervening strategies impact dietary intake or disease outcomes. To address these gaps and develop a deeper understanding of the impacts of these diverse initiatives – within rapidly changing food and political landscapes – researchers need to leverage rigorous research methods while also grappling with the constraints of the policy setting that may preclude the use of "gold-standard" methodological approaches.[13,14]

Randomized controlled trials (RCTs) are often considered the best practice for estimating the causal impact of a change in exposure (ie, an intervention). Well-designed explanatory RCTs can test the efficacy of interventions or the degree to which interventions produce effects under tightly controlled, optimal conditions.[15,16] Identifying efficacious interventions can ensure that scarce resources are allocated toward those that are most promising. Unbiased estimates of intervention effect sizes can indicate whether interventions are achieving their intended goals and help identify what interventions should be scaled up or discontinued. However, randomly allocating participants to policy interventions is often not feasible. For example, it would be logistically and ethically challenging to randomly assign half of the households in a community to receive a new grocery store that was financed through a healthy food financing initiative and the other half to have no access to the store. Yet, researchers, practitioners, and policymakers alike would like to know whether the new grocery store meaningfully changes community members' dietary behaviors and health outcomes.

For this type of question, and many others in which random allocation of an intervention or policy is not possible or has not occurred, the use of natural and quasi-experimental studies is an important evaluation approach. Although the exact definition is debated, "natural experimental study" generally refers to a study that exploits a change or exposure that is not directly manipulated by the researcher; often, these changes are instead the result of a policy or programmatic intervention (see Appendix S1 in the Supporting Information for this article available online).[17] In addition to allowing researchers to understand the impact of interventions that were not randomly assigned, natural experimental studies test the effectiveness of interventions – that is, the degree to which interventions produce desired effects in real-world settings with heterogeneous populations.[15] Although "pragmatic" RCTs can also be used to study effectiveness,[16] natural experimental studies often offer greater generalizability than many RCTs and can also provide insight into the feasibility of implementing the intervention in situations when researchers lack control over these processes. Longitudinal observational designs also allow researchers to study effectiveness and offer an improvement over the cross-sectional designs common in food environment research. Causal inference in longitudinal designs may, however, be limited by unmeasured confounders and other threats to internal validity. For these reasons, funding agencies such as the National Institutes of Health are increasingly recognizing the utility of natural experimental evaluations to study changes in the retail food environment,[18] and public health researchers have called for greater use of these designs.[19] However, natural experimental studies are more susceptible to threats to internal validity than RCTs and come with unique challenges in study design, execution, and inference. Evaluating nonrandomized food retail interventions provides both opportunities for new knowledge about how these changes affect important behavioral and health outcomes and particular challenges for evaluators wishing to conduct internally and externally valid research.

This article describes key challenges facing researchers who wish to evaluate changes to the retail food environment with natural experimental studies and puts forth strategies for overcoming these challenges, presenting the advantages and disadvantages of these approaches. Table 1 provides an overview of the key concepts in this article. The challenges fall into 3 broad categories: 1) study design and analysis; 2) selection of participants, selection of measures, and obtainment of data; and 3) real-world considerations. To keep this article grounded in the real-world constraints that govern natural experimental studies, case studies of challenges and solutions are integrated throughout. The article concludes with information on how researchers prioritize evaluation plans and check the scope of their evaluation plan.

CHALLENGES IN STUDY DESIGN AND ANALYSIS

To conduct a natural experimental study in the retail food environment a study design must be established and appropriate analytic methods to carry out this design must be selected. Because researchers conducting natural experimental studies are typically interested in the causal impact of an intervention on some outcome, internal validity is a key concern. This section focuses on some critical methodological and analytic decisions researchers face when evaluating retail food interventions with natural experimental studies and how these

Table 1 **Challenges, approaches, and considerations in natural experimental studies**

Challenge	Description	Approaches and considerations
Study design and analytic methods		
Representing the counterfactual	Estimating causal impacts requires comparing the observed outcomes to some reasonable estimate of the counterfactual outcomes. Control groups can provide a guess about the counterfactual outcomes but only if they are adequately similar to the intervention group	Consider how the intervention units are defined (geography, product type, etc), and use this definition to select a control group of units that are similar to those expected for intervention status Project forward preexisting trends as a plausible counterfactual Triangulating results using multiple control groups may increase costs, but could improve confidence in findings
Preexisting differences in the outcome variable	If the intervention and control groups have different baseline levels of the outcome variable, postintervention differences in the outcome variable could be due to the intervention or to these preexisting differences	Difference-in-differences models control for preexisting differences in the outcome variable between groups but assume that both groups would follow parallel time trends in the absence of the intervention, an assumption that can be difficult to verify
Differences in underlying trends	Even if preexisting differences in the outcome between groups are controlled for, the intervention and control groups may follow different natural trajectories of the outcome. Observed differences at follow-up could then be due to the intervention or to these differences in underlying trends	Multiple interrupted time series (also called comparative interrupted time series) approaches collect data from multiple preintervention time points, allowing the researchers to control for both preexisting differences and differences in the underlying trends between groups
Confounding and selection bias	Because participation in naturally occurring interventions or policy changes is typically voluntary, factors related to or determining participation may influence outcomes (rather than the intervention itself changing the outcome). Researchers must understand and account for these potential confounding variables	Measure and include control variables in analyses Use propensity score matching or inverse probability weights to model selection into intervention Use instrumental variables to leverage only plausibly exogenous variation in intervention participation/exposure levels
Selecting participants and ensuring external validity	Resource and time constraints often mean that research teams can only collect data on a fraction of the total population of interest. Procedures to select units for data collection must balance these constraints with a desire to produce results representative of the target population	A census will capture data on all units in the population. If a census is not possible, random selection procedures will maximize representativeness, although logistical constraints may prevent random selection. When this is the case, consider strategies to ensure that participants represent the target population of interest
Data and measures		
Selecting outcome measures	Researchers must define the constructs they believe might be affected by the intervention and select appropriate, valid, and reliable data collection methods for studying these outcomes. Dietary behaviors are often of interest; researchers must select from among many options for how to define the particular behavior of interest and identify appropriate data collection tools to measure this behavior	Consider the evolution of the intervention and the time scale over which to measure the outcome (eg, is short- or long-term change feasible/desirable to capture?) Clarify the type of behavior or outcome that the intervention is expected to change. What changes would indicate that the intervention was successful? A precise definition of the target behaviors or outcomes can guide measure selection Identify validated and reliable measures of the outcome, or, if none are known, consider a validation sub-study Consider including multiple outcome measures to allow for capturing and explaining unintended effects

(continued)

Table 1 **Continued**

Challenge	Description	Approaches and considerations
Obtaining data: primary vs secondary data trade-offs	Researchers must obtain data on the outcome of interest, as well as on intervention exposure and covariates. Primary and secondary data have unique trade-offs that researchers must balance within their resource and time constraints	Primary data collection offers the most control over the data collection process – researchers can select the measures they desire, sample from the population of greatest interest, and collect data at the appropriate time points for their evaluation. However, primary data collection is often limited in size and scope by resource constraints, and fast-moving interventions are sometimes implemented before data collection can be organized and implemented When primary data collection is not feasible, researchers can also leverage secondary datasets, including publicly available resources such as BRFSS, as well proprietary, commercial, or privately obtained datasets such as household or retailer scanner data. Secondary datasets often offer more data (a larger sample size or more time points) for lower cost, although they may not contain all variables of interest
Real-world considerations		
Practical considerations: timing of data collection	Behaviors, particularly dietary behaviors, may naturally fluctuate through the year, and intervention assessments must account for seasonality in their data collection schedules. Other timing challenges arise when policy or intervention implementation is delayed, making it difficult to know when to collect baseline data, or rendering already-collected baseline data outdated by the time of actual policy implementation. On the other hand, rapidly moving policies may make it difficult to collect baseline data before an intervention begins	Researchers must use their substantive knowledge to determine the appropriate data collection schedule that accounts for seasonal or other fluctuations in the outcome of interest. Collecting data from a control group experiencing the same fluctuations can also help control for these variations Multiple preintervention data points can help ensure baseline data are collected close enough to implementation to be relevant Existing datasets may be necessary in situations in which appropriate preintervention data could not be collected (eg, because of time delays or rapid interventions)
Policy/intervention specifics may evolve over time	Policies and interventions may change as they are developed and implemented, impacting planned and completed data collection	Researchers may need to consider collecting additional data that align with changes in the policy language or intervention specifics
Variable, incomplete, or imperfect implementation	Natural experimental studies lack the experimental control of RCTs; implementation of policies and interventions may vary substantially from expectations or across units	Collecting process evaluation data – information on the fidelity, reach, and dose of the intervention's implementation – can provide course-corrective feedback for implementers, generate "lessons learned" for future implementation of similar interventions, and enable a more nuanced outcome evaluation
Stakeholder relevance	Healthy food retailer policies and interventions may have impacts beyond diet and health outcomes, and stakeholders in policy, business, and the community may wish for information on these effects. Studying a broader array of outcomes creates new challenges and requires additional resources but can help generate additional support for health initiatives and provide insights into barriers to passage or implementation of policies and interventions	A variety of nonhealth indicators are often of interest to policymakers, businesses, and the community. Incorporating such measures in evaluation plans can help the results speak to a wider array of stakeholders and help build the case for healthy food retail interventions. Some measures to consider include cost effectiveness, return on investment, and aggregate costs and benefits; business impacts such as changes in revenues or profits; impacts on local jobs (creation, losses); indicators of community and economic development

decisions help evaluators maximize their study's internal validity.

Study design: representing the counterfactual

When evaluating healthy food retail interventions, the primary goal is typically to understand what impact the intervention has on the outcome of interest. Implicit in this goal is a comparison with a counterfactual: there is interest not just in what happens after an intervention is implemented but in how this outcome differs from what would have happened had there been no intervention. This distinction is important because a simple comparison of outcomes before and after an intervention tells us only what has changed over time, not whether observed changes are attributable to the intervention or policy. Because only one potential outcome can be observed – that is, only what actually happens, not what might have been, can be seen – a key study design consideration is how to represent the counterfactual. Natural experimental evaluators have 2 main options for representing the counterfactual: finding a control group and projecting forward historical trends.

Representing the counterfactual with a control group. As in RCTs, one way to represent the intervention group's counterfactual is by finding a suitable control or comparison group. Causal inference then depends on the extent to which the outcomes in this control group can be assumed to represent what would have happened to the intervention group in the absence of the intervention.

A first step in selecting a control group is to articulate a clear definition of the intervention group. In an RCT, this is often relatively simple: the intervention group is the group of individuals (or stores, neighborhoods, communities, etc) randomized to the intervention arm of the trial. In the case of natural experimental studies in healthy food retailing, researchers do not manipulate the exposure or randomize participants; thus, specifying the intervention group is more complex. To illustrate, food retail interventions are implemented at various levels: the geopolitical level (exposure varies across local, state, tribal, or national boundaries); community level (exposure varies across neighborhoods); retailer level (exposure varies across stores); and product level (exposure varies across products) (Table 2). Thus the intervention group might be all residents living in a state where a new tax is implemented, all participants in a federal food and nutrition assistance program that offers a new subsidy, all stores in a municipality implementing new minimum stocking requirements, or all products affected by an industry-led reformulation effort. These levels are not necessarily mutually exclusive and will often intersect. For example, if a city were to implement a per-ounce excise tax on beverages with added caloric sweeteners, exposure may vary across products (some products will be taxed, others will not), across geopolitical boundaries (the tax will be implemented within the city but not necessarily outside the city), and even across retailers within the city (eg, if some retailers, such as small businesses, are exempt from the tax).

Once the intervention group is identified, researchers can then use the same characteristics that define the intervention group to define the comparison group. For example, if the intervention group is a neighborhood where a new grocery story is being built, a potential control group might be another neighborhood with similar characteristics that is not slated to have a new grocery store enter. Researchers should attempt to find a control group that is similar enough to the intervention group that any differences in outcomes can reasonably be attributed to the intervention rather than to existing differences between the groups or to differences in natural trends between the groups over time.

In some cases, there may be many possible control groups to choose from. For example, if one state increases its sales tax rate on "disfavored" items such as junk foods and carbonated soft drinks, a researcher might consider several options for a control group, including a nearby state with no change in tax rate as the comparison unit, a state that does not share a border with the state of interest (to avoid potential cross-state shopping concerns), or a state that is geographically more distant but that is matched to the intervention state on important demographic, behavioral, health, or economic characteristics. It may be difficult to determine which of these options represents the best comparison group. In these instances, some have advocated for using a "synthetic control" group design: a data-driven approach in which researchers construct a weighted combination of all potential comparison units (eg, all 50 states) based on how similar those units are to the treated unit on observable characteristics thought to predict the outcome (eg, demographics, employment) and, sometimes, on the preintervention outcome of interest.[20,21] Additionally, in some instances, multiple types of control groups can be used simultaneously, reflecting that the levels of intervention described above may intersect. For example, suppose an evaluator wishes to estimate the impact of a city-wide 1-cent-per-ounce excise tax on sugar-sweetened beverages (SSBs). One comparison group might be a nearby or similar city without such a tax (geopolitical level). Another group could be untaxed beverages such as diet sodas, which are similar to SSBs in some ways (eg, they may follow similar seasonal trends in consumption) but are

Table 2 **Levels of healthy retailer interventions, example case studies, and example intervention and control groups**

Level of intervention[a]	Example case study	Example intervention group	Example control group
Geopolitical	2014 Minneapolis Staple Foods Ordinance: requires grocery stores within Minneapolis, Minnesota, city limits to carry minimum amounts and varieties of specific categories of foods and beverages	Licensed grocery stores and their shoppers in Minneapolis, Minnesota	Licensed grocery stores in neighboring "Twin City" of St Paul, Minnesota
Community	Pittsburgh Healthy Food Financing Initiative (PHRESH study): as part of the Healthy Food Financing Initiative, the Hill neighborhood in Pittsburgh, Pennsylvania, received a new full-service grocery store	Residents of the Hill neighborhood in Pittsburgh, Pennsylvania	Residents of Homewood, a neighborhood in Pittsburgh that is sociodemographically and geographically similar to the Hill neighborhood but that was not scheduled to receive a new full-service grocery store during the study period
Retailer	Walmart Healthier Food Initiative: in 2011, Walmart pledged to eliminate trans-fat and reduce added sugar and sodium in products sold in their stores, among other changes	Packaged food and beverage purchases at Walmart before and after the pledge	Packaged food and beverage purchases at comparable chain retailers before and after Walmart's pledge; projected simulations of Walmart's prepledge trends
Product	Berkeley tax on sugar-sweetened beverages: beginning in March 2015, the city of Berkeley, California, began levying a 1-cent-per-ounce excise tax on the distribution of sugar-sweetened beverages within city limits	Sales or consumption of taxed products (ie, sugar-sweetened beverages as defined by city ordinance) in Berkeley, California	Sales or consumption of untaxed products in other cities (eg, Oakland or San Francisco, California) without a sugar-sweetened beverage tax

[a]Interventions may occur at more than one level simultaneously, and examples of case studies, intervention groups, and control groups at each level are meant to be illustrative, not exhaustive.

Abbreviations: BRFSS, Behavioral Risk Factor Surveillance System; PHRESH, Pittsburgh Hill/Homewood Research on Eating, Shopping, and Health; RCT, randomized, controlled trial.

not taxed under most beverage excise policies (product level) (see Cawley and Frisvold[22] for an example using these 2 comparison groups to evaluate the Berkeley tax on SSBs). Triangulating results using multiple comparison groups can strengthen causal inference.

Case study: selecting a control group in the 2104 Minneapolis Staple Foods Ordinance evaluation. A recent example of these considerations comes from an ongoing evaluation of the 2014 Minneapolis Staple Foods Ordinance (SFO), the first and only local policy in the United States that sets minimum stocking standards for all stores with grocery licenses.[9] The SFO requires grocery stores within the Minneapolis, Minnesota, city limits to carry minimum amounts and varieties of specific categories of foods and beverages (eg, fresh fruits and vegetables, whole-grain items). In selecting a comparison community, the research team selected the city of St Paul, Minnesota, which lies adjacent to Minneapolis (its "Twin City"). St Paul was in many ways comparable with Minneapolis and offered the practical advantage of proximity. The 2 cities are similar in terms of demographics and retailer landscape but have 2 distinct city and county governments; thus they are subject to different local policies. The proximity of St Paul allowed for the same team to collect data in both intervention and control sites within the same study period. Moreover, St Paul could capture secular changes (natural changes in the outcome measure that would occur over time even in the absence of a healthy retail intervention)[23] that might occur locally during the study period but might be difficult to measure and control for. This might include changing local norms related to food purchasing or changing perishable food distribution practices among small stores in the greater metropolitan area.

Although St Paul was a practical choice as a comparison site, this decision also presented some risks. It would not have been inconceivable for a similar policy to be enacted in St Paul during the study period, even though the research team confirmed with St Paul authorities that such a policy was not under consideration during the study planning phase. An additional threat to validity in selecting St Paul was the possibility of contamination. Due to the geographic proximity of stores in the 2 cities, many customers might shop in both communities, yielding some overlap between the intervention and control communities. If the increase in healthy items in Minneapolis stores led to an increase in customer demand for healthy products, St Paul stores could change their inventory in response to the changing local norms resulting from the ordinance. All things considered, St Paul was an acceptable choice for a control setting. The research team is in the process of carrying out a prospective evaluation of the policy, collecting data pre- and postimplementation in both communities.

Representing the counterfactual with projected historical trends. In some cases, it may not be possible to choose a control group before the study begins – for example, if policy implementation has already begun. Additionally, in the case of national-level interventions (eg, initiatives by major national retail chains or policies implemented by the federal government), there may be no suitable control group. In these scenarios, rather than using a comparison group to represent the counterfactual, secondary data can be used to construct a counterfactual. Specifically, researchers first determine the historical time trends in the outcome for the intervention group. These trends are then projected forward as an estimate of what would be expected to occur in the absence of the intervention. Researchers then compare the observed postintervention change with the projected postintervention change; if the observed postintervention changes differ from this counterfactual and appropriate methods are used to control for secular trends, contextual factors, and individual and household characteristics, investigators may be able to demonstrate that postpolicy changes were attributable to the policy rather than preexisting trends. This approach has been used to evaluate corporate voluntary initiatives,[24,25] as well as national-level policies, such as Mexico's nationwide 8% excise tax on nonessential energy-dense foods and 1-peso-per-liter excise tax on nondairy, nonalcoholic SSBs.[26,27] One advantage to this method is that it can be used in conjunction with existing datasets to evaluate programs and policies retrospectively while ensuring a high level of statistical rigor. Nonetheless, a key limitation of this method is its reliance on the assumption that prepolicy trends would have continued into the future if the policy had not been enacted; however, other interventions or secular changes could invalidate this assumption.

Case study: representing the counterfactual for the 2011 Walmart Healthier Food Initiative. In 2011, Walmart introduced a Healthier Food Initiative (HFI), pledging to eliminate trans-fat and reduce added sugar and sodium in products sold in stores, reduce the price of healthier products, and place a front-of-package logo on private-label products meeting nutritional standards.[28] To evaluate the impact of this initiative, researchers examined changes over time in the nutritional profile of household purchases made at Walmart compared with purchases made at other chain grocers.[25,29,30] The treated units were purchases of foods made at Walmart, whereas control units were purchases made at other large retail food chains. Examining concurrent trends in packaged food and beverage purchases at both Walmart (intervention) and comparable chain retailers (control) allowed researchers to observe whether Walmart's initiative was associated with changes in purchases' nutritional profile above and beyond industry secular trends. In addition to using a control group, the research team also created counterfactual simulations projecting forward prepledge trends in the nutrient profile of Walmart packaged food purchases.[25] The researchers found that post-HFI shifts in nutrient density and percentage volume of key food groups were similar to or less than what would be expected had pre-HFI trends simply continued, highlighting the importance of accounting for how outcomes might naturally change even in the absence of specific interventions.

Analysis: accounting for key sources of bias

Because natural experimental studies involve interventions that are not randomly assigned nor under the control of the researcher, various sources of bias are possible, and appropriate analytic methods must be used to mitigate these potential problems. Selecting an appropriate counterfactual or control group addresses many sources of bias.[23] Yet, even thoughtfully selected control groups may differ from the intervention group in a variety of ways that undermine internal validity, including differing in preintervention outcomes, underlying time trends, or factors motivating participation in the intervention. Common analytic techniques to address these sources of bias are presented here.

Preexisting differences between groups. In RCTs, randomization ensures that the intervention and control group have similar levels of the outcome variable as a baseline. This ensures that any differences between groups after the intervention has taken place cannot be attributed to preexisting differences in the outcome variable. In natural experimental studies, the intervention and control groups may not have similar levels of the outcome variable at baseline. To overcome this problem, many natural experimental studies use difference-in-differences (DiD) estimation. In DiD, data are obtained from both groups before and after the intervention has been implemented. Then, the pre/post difference in the outcome variable is calculated for both groups, and the impact of the intervention is estimated as the difference in these pre/post differences. By comparing the change in the outcome variable over time in the intervention group with the same change in the control group, DiD methods control for any preexisting differences in the outcome variable between the groups. However, this method assumes that the intervention and its counterfactual would have followed parallel trends over time in the absence of the intervention, an assumption that can be difficult to verify or disprove.

Differences in underlying time trends. One way to improve upon a generic DiD design is to obtain data from many preintervention time points and apply a multiple interrupted time series (multiple ITS, also called comparative ITS) approach. In this design, the preintervention time trends in the outcome are modeled for both groups. Then, postintervention data are used to examine whether the intervention group deviates from its preintervention trend by a greater amount than does the control group. An advantage of using multiple ITS over DiD is that multiple ITS controls both for preexisting differences in the outcome variable and for differences in the underlying time trends between the intervention and control groups. Studies that represent the counterfactual with projected historical trends can also make use of ITS designs by comparing observed with projected time trends.

Selection bias and confounding. Two common challenges to natural experimental evaluation are selection bias and confounding. Confounding refers broadly to situations in which observed differences between the intervention and control groups can be explained by factors other than intervention status. Selection bias occurs when the observed differences between intervention and control groups are explained by factors that motivated or led the intervention group to participate in the intervention, rather than being explained by the intervention itself. Individuals, cities, companies, and stores that choose to participate in an intervention may differ in some important ways from those that do not. For example, a city that votes to implement a per-ounce excise tax on beverages with added caloric sweeteners may differ by education level, socioeconomic status, or underlying dietary preferences relative to a city that does not pass a soda tax. These differences might also lead cities with a tax to reduce their SSB consumption over time even if the tax does not itself reduce consumption. When the intervention and control groups have different distributions of key factors (potential confounders) influencing the outcome (including factors that also influence selection into the intervention), these differences must be controlled for or estimates of the intervention's effect will be biased.

To mitigate these issues, it is useful to create a conceptual model of key factors influencing the outcome. These models help the research team identify potential confounders and thus signal what variables need to be controlled for in an analysis (eg, with regression, matching, or stratification procedures, or, for confounders that are stable over time, with longitudinal models such as fixed-effects models). The conceptual model can also incorporate information on the process of selection into the intervention: how and why did the units under study come to be in either the intervention or control group? If factors predicting intervention status are also related to the outcome, researchers should attempt to collect or obtain data on these variables and either control for these factors directly, as in multiple regression, or use them to explicitly model and, therefore, account for selection into the intervention group, as with inverse probability weights[31,32] or propensity scores.[33] Alternatively, if some factor is known to strongly and exogenously influence participation in an intervention without otherwise influencing the outcome, instrumental variables estimation can exploit this exogenous variation in intervention status to provide an unbiased estimate of intervention impact.[34] Although each of these approaches has drawbacks, they can improve internal validity when their assumptions are met.

Systems science. Public health research and practice increasingly recognize health behaviors and outcomes, as well as interventions to improve these variables, as part of dynamic, complex systems.[35–38] Studying complex systems requires different methods than the traditional natural experimental designs described here, and future research should explore the potential for these methods (eg, network analysis, systems dynamics, agent-based modeling) to be fruitfully applied to understanding the impacts of food retail interventions.

Case study: accounting for selection bias in the 2011 Walmart Healthier Food Initiative. Selection bias is often relevant to evaluations involving retailer initiatives. Selection bias may be a concern if individuals who are more likely to shop at the intervention retail chain have different characteristics (eg, socioeconomic levels, underlying dietary preferences) than those who shop elsewhere and if these characteristics are also related to the outcome of interest. For example, low-income households tend to be more likely to shop at Walmart than higher-income households and may also purchase different types of foods and beverages. In the Walmart HFI described previously, the evaluation needed to account for the underlying differences in the types of consumers who do and do not shop at Walmart. In addition, selection bias might occur if Walmart's initiative led to changes in its consumer base (eg, by causing new, more health-conscious consumers to opt into shopping at Walmart), which could cause changes in the nutritional profile of purchases made at Walmart due to the shoppers who opted into participating in Walmart's initiatives rather than due to the intervention improving the nutritional profile of existing customers. To correct for the first type of selection bias (underlying fixed household characteristics related to the likelihood of shopping at Walmart), researchers used fixed-effects models, which control for the influence of stable (time-invariant) household characteristics.[25] To reduce the potential for bias from a changing consumer base, the researchers created inverse probability weights to model the probability of a household shopping at Walmart based on household size, income, race, household composition, and market-level covariates like the local unemployment rate. Weighting observations by the inverse probability of being a Walmart shopper create intervention (Walmart shoppers) and control groups (shoppers at other stores) that are more similar to one another, helping to reduce the likelihood of selection bias.

CHALLENGES SELECTING PARTICIPANTS, SELECTING MEASURES, AND OBTAINING DATA

Once researchers have determined a study design and analytic approach for their natural experimental evaluation, next steps include selecting participants, selecting appropriate measures, and obtaining data on these measures from sampled participants. Natural experimental studies of food retail interventions generate unique challenges in these domains. Because the ideal data sources are not always readily available, many studies choose to include a variety of measures that can be triangulated to give a fuller picture of the effect of the program or policy being evaluated. Primary data collection (eg, store audits or dietary recalls) and secondary data (eg, retailer sales data or business databases) both have advantages and limitations but can complement each other for both exposure and outcomes assessment. Here, specific challenges are highlighted and options for addressing these issues in natural experimental studies of the retail food environment are suggested.

Selecting participants to maximize representativeness

The goal of participant selection is to choose units that represent the underlying group of interest, whether it be a group of individuals, stores, or products. Sometimes, it is possible to obtain data on all units in the population. For example, in an evaluation of the Berkeley 1-cent-per-ounce excise tax on the distribution of SSBs, Cawley and Frisvold[22] took a near-census of retailers in the city. This is also possible when the intervention has a relatively small target population, such as a mobile market program in a subsidized housing community. In the case of interventions meant to serve a community without well-defined borders (eg, new supermarkets may serve nearby residents as well as those who work in the area or who are willing to drive to grocery shop), it may be more difficult to determine exactly who the target population is and thus whom to include in the study. Some studies have chosen to approach a random sample of people within the community the intervention is meant to serve (eg, sampling from the neighborhood where a new supermarket is being built[39]) or everyone within a certain radius of the intervention site.[40] Another common approach is to conduct intercept surveys in the intervention community (and often also a comparison community not slated to receive any intervention), which some have found improves representativeness in hard-to-contact populations.[41–43] One complication with place-based sampling strategies in healthy food retail research is that many people do not shop at their closest grocery store,[44] and sampling in a particular location (as in street-intercept surveys) or from within a defined radius from a new store may mean that potential shoppers are missed and/or that some of those included will not be likely to shop at the store. This may not pose an issue if the sample is large enough and researchers have the resources to oversample to account for the fact that many people included in the data collection may not be exposed to the intervention. In other cases, such as during pilot work or initial efficacy studies, it may be more important that the data collection reach as many potential users of the intervention as possible. Because natural experiments harness real-world observations, there may be finite sample size limits that are not under the researcher's control. Nevertheless, it is always important to calculate and report power as in other quasi-experimental designs. Studies that rely on secondary data (and thus that may not be able to increase their sample size) may wish to conduct post hoc power calculations to understand the implications of the sample size on the ability to detect an effect.

Case study: participant selection in the Pittsburgh Hill/Homewood Research on Eating, Shopping, and Health Study. One example of participant selection comes from the Pittsburgh Hill/Homewood Research on Eating, Shopping, and Health (PHRESH) Study, a longitudinal quasi-experimental study of households in Pittsburgh, Pennsylvania, before and after the introduction of a new full-service grocery store in the Hill District of Pittsburgh.[45] The researchers are following households in neighborhoods in both the Hill District area (intervention neighborhoods) and in similar but geographically separate areas not receiving a new grocery store (control neighborhoods). To select households, the researchers first created a sampling frame of residential addresses in the intervention and control neighborhoods. Intervention households were randomly sampled within strata of increasing distance to the future grocery store site, with households closest to the new store oversampled. Control households were selected by simple random sample. All sampled households were approached, and, of those who were reached at home and were eligible, 87% agreed to participate. Within households, the primary shopper was interviewed.

This sampling strategy offered several advantages. Using random sampling helped the researchers achieve a study sample that reflected the communities as a whole.[45] In addition, the use of stratified random sampling within the intervention communities allowed for oversampling the households that may be most likely to use the new store.[45] Finally, interviewing the primary shopper in the household, rather than a randomly selected adult, meant that respondents could thoughtfully answer survey items about perceived access to healthy foods, food purchasing habits (eg, types of stores visited, the frequency of shopping, use of the new grocery store).[46] Despite these strengths, the in-person recruitment strategy may have skewed the sample toward the types of households most likely to have someone at home (eg, older, less likely to have children).[46] When possible, collecting data on nonrespondents can help researchers understand the characteristics of those who do and do not participate and correct for differences between the included sample and the target population.

Selecting outcomes

Before data can be obtained from sampled participants, evaluators need to decide what data are needed: what are the key independent, dependent, and control variables that need to be measured to conduct the outcome evaluation. Typically, the independent variable or exposure will be the intervention of interest (eg, a policy or ordinance change, tax, subsidy, label, marketing change). Although the ultimate goal of many natural experiments may be to change health outcomes, it is important to match the outcome of interest to the evolution of the intervention itself.[47] As noted above, like any intervention, natural experimental evaluations require a conceptual model that details expected changes, pathways to change, and potential factors that may affect outcomes. In this way, researchers can identify the most salient outcomes, mediators, process measures, and control variables to measure. Depending on the time available and the stage of the intervention, researchers may decide to focus on food environment outcomes, dietary behaviors, and/or health outcomes.

Food environment outcomes. In some cases, it may be more appropriate to look at food environment outcomes first before examining dietary measures or health outcomes. This may be the case if the natural experiment is unproven or in a pilot phase, the main focus of the evaluation is on understanding implementation, or the natural experiment is a policy designed to cause food environment change (eg, industry initiatives to remove calories from the food system). Prioritizing measuring environmental change may also make sense if there is not a strong indication that the natural experiment will change diet in the given timeframe.

In selecting food environment measures, it is important to consider that there are multiple dimensions to food access, including availability, affordability, accessibility, and acceptability.[48] Although ideally researchers would choose validated measures, most published measures of the food environment do not contain information on reliability or validity.[49,50] Those that do exist may need to be modified for specific contexts. If resources are available, researchers should consider conducting a validation substudy before or during the data collection process; ideally, validation occurs before the outcome evaluation begins so that a validated measure can be obtained prior to the beginning of data collection. The following case study highlights the importance of selecting culturally appropriate measures and collecting data on multiple food environment indicators as a change in factor can have detrimental effects on others.

Case study: evaluating the impact of the 2009 federal revisions to the Special Supplemental Nutrition Program for Women, Infants, and Children on the retailer environment. Efforts to evaluate the health impact of recent, deliberate changes to federal food assistance programs such as the Special Supplemental Nutrition Assistance Program for Women, Infants, and Children (WIC) highlight challenges with selecting and measuring appropriate outcomes in the retail food store environment. In 2009, federal revisions to WIC increased cash vouchers for fruit and vegetable purchases and updated cost containment, administrative, and WIC food packages.[51–53] The new policy, set forth by the US Department of Agriculture, sought to influence both the broader retail food environment and individual behavior. Consequently, researchers attempting to evaluate the impact of the new policy had to grapple with whether to focus on measures of the retail food environment (eg, availability, accessibility, and affordability of foods) or measures related to individual-level dietary consumption (eg, purchase and consumption of various foods). Lu et al.[54] recently undertook an evaluation of the WIC program in Texas and opted to examine the broader retail food environment. The researchers also considered geographic differences (ie, urban vs rural) in their evaluation due to urban–rural disparities in the resources of food stores.[55] Moreover, although examining the local retail food environment (eg, in a specific city) is important, WIC policy is driven by state guidelines,[55] so the state, rather than the city, was used as the geographic unit of analysis for the evaluation. One of the first challenges was finding a validated instrument that would capture foods that were culturally relevant for the area and reflected dietary patterns. In this instance, the authors first adapted and field-tested one of the popular validated store survey tools, the Nutrition Environment Measures Survey, for use in the study area. Conceptualizing and measuring the availability, accessibility, and affordability of foods also presented a challenge for the researchers. Availability measurements included the visibility and amount of shelf space allocated to each item, the variety of produce, the stocking and quality of products, and the availability of culturally specific (Hispanic) foods. Accessibility was defined as the visibility and labeling of WIC foods based on marketing principles, such as whether specific foods were at eye-level, and affordability was defined as the price of the least expensive brand item for a particular product. Lu et al.[54] found improved accessibility and availability of food items following the WIC policy update but did not find an improvement in affordability. It is possible that the addition of vouchers may have increased purchasing power among WIC participants but did not make food more affordable to the broader community, which presumably would include many other low-income families who were not eligible for WIC eligible. This underscores the importance of examining multiple outcomes in the evaluation of food and beverage policies to accurately assess the full impact on the retail food store environment.

Dietary behaviors. Public health evaluators will typically be most interested in whether a healthy food retail intervention changes dietary behaviors, but dietary behavior can itself be complex to measure and define, and which dietary behaviors are of interest will vary with the specifics of the intervention being evaluated. For example, a subsidy program for fruits and vegetables will likely be most interested in measuring fruit and vegetable consumption or nutrients related to these items (eg, fiber), whereas a menu-labeling policy might wish to examine caloric intake at restaurants subject to the policy change. An additional challenge is that even if an intervention is expected to change dietary behaviors in the short term, some have argued that a better outcome measure is whether healthy habits are sustained, rather than just initiated.[56]

Although the focus in this article is more on the types of outcomes researchers should consider measuring rather than the specific measurement tools, publications exist that debate the merits of different dietary intake tools.[57] Because collecting individual dietary intake data can be challenging, many food retail studies have chosen to focus instead on looking at changes in store purchases as a proxy for dietary change, relying on the assumption that if people are purchasing healthy food, they are likely to be eating healthier food. A benefit to this approach is that purchase data, when objectively collected, are less subject to desirability bias than self-reported dietary data. Pilot studies or evaluations with limited time to collect data prior to the start of a natural experiment may also favor purchase data. Challenges to using purchase data include limited or incomplete data from smaller markets, such as farmers' markets, corner stores, and mobile markets, because these venues often lack sophisticated point-of-sale systems to track customer purchases. Retail inventory records may be an alternative (ie, looking at trends in wholesale purchases of target items by the participating retailers) in this situation but still may not give a full picture of the effect of the program on diet. Using store-level data on food purchases may indicate that healthy food purchases are increasing but could also represent a change in the customer base at those retailers. To reduce the likelihood of this alternative explanation, researchers can track individuals' purchases over time by collecting individual-level purchase data in the form of customer receipts, personally identifiable purchase data from loyalty cards, or sales records that record customer identity (such as at some mobile markets).

In addition, a final challenge is that researchers may be interested in understanding how the policy impacts total dietary intake because individuals consume many foods and beverages – not just the food(s) or beverage(s) targeted – as well as consume foods from a variety of sources (food retail outlets as well as away-from-home sources like restaurants, fast food outlets, or schools). Individuals may respond to policy change by substituting one food or beverage for another or shifting their allocation of in-store and away-from-home food purchases, but sales and food purchase data only capture in-store purchase and, if product categories are restricted, may not capture the full range of substitutions individuals might make. Supplementing purchase or sales data with dietary intake measures such as 24-hour recalls or food-frequency questionnaires may provide a better assessment of the total dietary change. With sales and purchase data, it is preferable to collect data on all products rather than only those targeted by the policy, and it is also useful to collect data from a variety of retail types (supermarket, convenience store, locally owned shop, or tienda) in order to understand potential shifts across retail outlets.

Health outcomes. Individual and population nutrition-related health outcomes (eg, obesity, diabetes, cardiovascular disease, and certain types of cancer) are also often of interest; however, literature looking at the effects of food environment interventions on health outcomes is generally limited to the collection of body mass index (BMI). Other potential health targets (eg, diabetes, blood pressure) generally require longer-term exposure and thus are not easily collected during the typical evaluation that lasts a year or less. Given the expensive and often invasive procedures required to measure these health outcomes, they would be most appropriate for longer-term evaluations of established interventions, which have already shown changes in purchasing and/or diet. In many cases, longer-term outcomes are studied using other methodologies, such as simulation modeling (see Basu et al.[58] and Wang et al.[59]).

Linking intervention exposure to outcomes. Considerations for defining and measuring the exposure in natural experimental studies have been detailed elsewhere[60]; here, it is noted that collecting measures of participants' exposure to the intervention can help researchers understand the mechanisms through which an intervention exerts its influences or the reasons why an intervention was not effective. Data on purchasing or store usage behaviors can be used as indicators of intervention exposure and might mediate the relationship between an intervention and a dietary outcome. Self-reported measures of exposure are another alternative. Many studies looking at changes in the food environment, including the case study that follows, have also included perceived access measures,[61–63] which can serve as both an outcome measure and an indicator of intervention exposure.

Case study: evaluating the impact of the Pittsburgh Healthy Food Financing Initiative on behavioral outcomes in the Pittsburgh Hill/Homewood Research on Eating, Shopping, and Health study. The PHRESH study also highlights the importance of measuring multiple behavioral outcomes to fully understand the effect of healthy food retailer interventions.[46] For their evaluation, the researchers measured dietary intake among a sample of residents in both the intervention and control communities using two 24-hour dietary recalls before and after the supermarket opened. On the follow-up survey, they also asked residents in the intervention community how often they had visited the new supermarket. The evaluation found that, compared with the control communities, participants in intervention communities decreased their consumption of kilocalories, added sugars and solid fats, alcohol, and added sugars. No change in BMI, fruit and vegetable intake, or whole-grain consumption was reported. The researchers also conducted a subgroup analysis comparing individuals in the intervention community who reported regularly using the new store to those who did not. These analyses revealed no significant associations between dietary variables and store usage, although regular shoppers in the intervention community did report increases in perceived access to healthy food compared with residents who did not shop regularly at the new supermarket. The addition of the perceived access measure proved useful in this case because it suggested mechanisms underlying the improvement in dietary variables seen in the intervention community. This study demonstrates the importance of linking intervention exposure to outcomes and of incorporating data on multiple behavioral measures to get a richer picture of how changes to the food retail environment affect individuals.

Obtaining data

Investigators have several options for gathering data to evaluate healthy retail interventions; the best option will depend on the outcomes of interest as well as time and resource constraints. For example, to measure dietary behaviors (eg, consumption of the products targeted by the intervention; overall dietary quality), evaluators can collect their own data using intercept surveys, random-digit dialing, or other methods of interviewing. But, primary data collection can be resource intensive. Another limitation may be insufficient lead time before a policy is implemented to effectively gather baseline (preintervention) data, particularly for rapidly moving interventions. If primary data collection is not feasible, another option for measuring dietary behaviors is to use publicly available datasets that contain consumption data that can be matched to intervention variables of interest. For example, researchers have modeled the impacts of sales taxes on carbonated soft drinks using surveillance data from the National Health and Nutrition Examination Survey, Behavioral Risk Factor Surveillance Survey, and Youth Risk Behavior Surveillance Survey and using longitudinal studies such as the Early Childhood Longitudinal Study.[64–66] Additionally, researchers can seek out electronic purchase data, such as store-based scanner data or household electronic purchase data. These datasets can be obtained through private agreements with retailers or, if resources are available, by purchasing them from commercial venders such as the Nielsen Corporation or Information Resources and have been used by some researchers to evaluate healthy food retail interventions.[67] Product- or store-level variables (eg, prices, environmental changes) can be assessed with store-based audits or retailer surveys and use of store-based scanner data (see below for case study examples) or, if available, public databases of prices (eg, the National Institute of Statistics and Geographic Consumer Price Index in Mexico).[68] Data on longer-term outcomes, such as health impacts and cost-effectiveness, may be more difficult to acquire, although each of the aforementioned public datasets contain some health measures (eg, BMI, diabetes).

Case study: 2015 Berkeley sugar-sweetened beverage tax. Recent and ongoing evaluations of taxes on SSBs implemented in Berkeley, California, highlight some of the challenges and opportunities for obtaining data. In March 2015, Berkeley implemented a 1-cent-per-ounce excise tax on the distribution of nonalcoholic, nondairy beverages with added caloric sweeteners. Several outcomes were of interest following implementation, including changes in consumer-facing prices of taxed beverages (ie, "pass-through" rate) as well as changes in purchases and consumption of taxed and untaxed beverages. Different research groups took slightly different approaches to examining these outcomes. As one example, to examine pass-through, 3 separate studies each conducted store-based surveys examining the price of taxed and untaxed beverages before and after the tax was implemented.[22,69,70] To collect data on price changes, researchers visited stores and recorded prices of taxed and untaxed products. In 2 studies,[22,69] data were also collected from stores in neighboring cities without an SSB tax. These methods allowed researchers to focus data collection efforts on stores of particular interest (eg, Falbe et al.[69] examined stores in low-income areas). In addition, these methods allowed researchers to examine stores that may not have access to or be willing to provide scanner data. Despite the benefits of primary data collection, these time- and resource-intensive approaches could potentially limit sample size in terms of the number of stores visited, the number of products assessed, and the number of time points of data collection. Ng et al.[70] complemented their survey-based store price data with detailed retailer scanner data from 2 chains with stores in both Berkeley and neighboring cities. The combination of these methods increased the number of products and time points examined and allowed for a detailed analysis of trends in prices over time.

To examine changes in beverage consumption, Falbe et al.[71] conducted intercept surveys of residents in Berkeley and 2 comparison cities. This data collection effort was resource intensive, limiting sample size. Additionally, to keep the survey brief, researchers asked participants to report their beverage consumption in broad categories (eg, integer servings per day, week, or month), rather than in quantitatively precise amounts (eg, calories/d) as could be estimated from a 24-hour recall or more intensive dietary assessment method.[57] Silver et al.[72] collected 24-hour beverage recalls from a random sample of Berkeley residents and also obtained point-of-sale data from supermarkets in Berkeley and comparison cities. Because sales data are collected on an ongoing basis, they could be acquired even after the intervention had begun, relieving some of the time pressure of primary data collection. However, purchase data did not contain information on the characteristics of the individuals making purchases.

REAL-WORLD CONSIDERATIONS

In addition to study design, analysis, and data challenges, natural experimental evaluations of healthy food retail interventions often face practical and logistical challenges that RCTs may not. Randomized controlled trials by definition involve a degree of control over intervention implementation, which natural experimental studies lack. The evaluation team must therefore be vigilant in monitoring the development and implementation of the intervention. Natural experimental studies also have real and immediate relevance to a range of stakeholders, including business, government agencies, shoppers, and taxpayers. Thus, successful evaluations require fostering partnerships with key stakeholders. Collecting measures of intervention impact (both costs and benefits) that are most relevant to these stakeholders could help enhance these partnerships.

Implementation and practical challenges

One practical challenge faced in many natural experimental studies is dealing with timing. For example, dietary behaviors often vary seasonally, and intervention assessment must account for seasonality when scheduling data collection. For example, evaluations of the WIC Farmers' Market Nutrition Program, a program to increase WIC participants' access to farmers' markets and community gardens, compared participants' fruit and vegetable intake to the previous season's intake to control for seasonality of fruit and vegetable consumption.[73] In addition to farmer's markets, many mobile market and corner store programs are increasingly relying on locally grown produce to stock their shelves. This poses a challenge to researchers conducting smaller pilot studies because produce availability in many areas can change dramatically throughout the year. Wherever possible, researchers should work to collect data from intervention and control participants contemporaneously and to collect pre- and postintervention data at the same time of the year. However, it may not be possible to wait a full year to complete follow-up measures within the same season or to perfectly time data collection across groups. Including questions about locally grown/seasonal products on food-frequency questionnaires can help ensure that highlighted products are captured in the data collection. Even in these cases, subgroup and sensitivity analyses may be necessary to examine the impact of the timing of data collection on food and beverage consumption.

Another practical challenge is that local policies may be delayed for political reasons or because competing priorities emerge.[74] Unanticipated delays make it challenging to allocate evaluation resources efficiently (eg, hiring and training staff). Additionally, if too much time passes between baseline data collection and intervention implementation, preintervention data can become irrelevant. Although potentially expensive, collecting or obtaining data from multiple preintervention time points can mitigate these issues and also allow for special analytic techniques, such as interrupted time series designs and counterfactual simulations (see above for an example of estimating counterfactual trends from preintervention trend data). Conversely, some interventions move rapidly, making it difficult to obtain data before implementation begins. In these situations, researchers may need to rely on existing datasets, such as household food purchases or store sales data, which can typically be acquired at any time because data collection occurs continuously. In addition, it is not uncommon for organizations to initiate changes in products prior to the formal beginning of a program or policy, as demonstrated by recent school lunch and menu-labeling initiatives and corporate voluntary initiatives.[24,75,76] This makes the implementation date ambiguous and complicates analyses. When the true starting date is unknown and continuous data across a time period are available, techniques such as switching regressions[77] can allow investigators to identify the potential starting point that best fits the data.

Another potential challenge is that policies and other initiatives can evolve during their development and implementation. Researchers must keep current with the specifics of the policy language and implementation. When specific policy requirements change, it may be necessary to seek out alternate sources of data or collect additional preimplementation data.

Conducting evaluations in real-world settings requires careful planning and collecting of process measures to address implementation challenges. As described by Moore et al.[78] and demonstrated in the case studies above, process measures have a number of functions, including revealing how contextual factors shape the findings, reflecting on intervention implementation, and illuminating mechanisms of impact. Process evaluation can be useful across all stages of evaluation. In the formative stages, collecting measures on feasibility can inform optimal strategies for implementation or provide helpful feedback to implementers while there may still be an opportunity for course correction. Upon beginning implementation, key process measures include fidelity (whether the intervention was delivered as planned), reach (whether the intervention reaches the intended audience), and dose (the intensity of the intervention).[78] Policymakers and other stakeholder might be particularly interested in process measures pertaining to the use of resources, such as implementation costs, staff training, and communications.[78] Process measures also include barriers and facilitators for implementation. For example, in the evaluation of the Berkeley SSB tax, researchers conducted qualitative interviews with retailers, distributors, and city officials regarding the challenges and successes of implementation.[79] Such information could be used by the city to understand how to better communicate aspects of the tax with the public and business community. Moreover, understanding how the local context might have affected results is important to assess generalizability and may allow other communities to better predict how successful implementation might be given their own local context.

Process evaluation data can also enable a more nuanced outcome evaluation. For example, when an intervention fails to demonstrate a significant effect, researchers can use process evaluation data to assess whether the intervention was truly ineffective, or whether it was simply poorly implemented and therefore unlikely to have had an effect. This may be particularly important in evaluations of nonrandomized, noncontrolled interventions because policy adoption and compliance can be a slow process. In addition, researchers can use process data to examine whether implementation success is related to changes in the outcome of interest (eg, did stores with the best implementation of a retailer intervention also exhibit the largest changes in purchases?) or whether implementation was different across settings. Finally, qualitative assessments can provide valuable insight into the success and barriers faced by stakeholders, as has been demonstrated in new retailer evaluations.[80,81] In these ways, process measures can help understand the success of shortcomings and unintended consequences of policy implementation.

Case study: practical challenges in evaluating the Minneapolis Staple Foods Ordinance. Under the original proposal of the Minneapolis SFO, which set minimum stocking requirements for 10 categories of food for all licensed grocery stores in the city, retailers were required to carry 15 gallons of low-fat milk and 12 boxes of whole-grain cereal at all times. After seeking feedback from local businesses and business associations, the requirements were reduced to 5 gallons and 4 boxes, respectively, just weeks before the ordinance passed, in order to reduce the burden to stores. This last-minute change presented a challenge in collecting the appropriate data: compliance with the proposed law was defined differently when data collection instruments were created than when the policy was passed. Maintaining a strong collaboration with city partners (eg, representatives from the Health Department and City Council) through all phases of development and passage of the policy ensured that the researchers were informed about decisions to modify the ordinance requirements.

Measuring implementation challenges in the Minneapolis SFO proved important in understanding the policy process. A year after the official policy implementation date, store compliance was still low, likely because enforcement (fines, citations) did not begin until an additional year after implementation. This necessitated a long follow-up period of observations. To examine the implementation challenges faced by stores, the evaluation team conducted interviews at multiple time points with store managers to ascertain what changes managers had made in supply sources and stocking patterns, decision-making around stocking, and perceived changes in item-specific sales.[82] The evaluators plan to conduct growth mixture modeling to examine compliance over time and to determine whether compliance classes exist (eg, immediate compliers, delayed compliers, noncompliers). This method can incorporate categorical latent variables that represent trajectory classes to better understand reasons for successful implementation and challenges at the store level. These analyses will be particularly relevant to stakeholders, including city government officials who are charged with making sure the ordinance and its enforcement are appropriate uses of city resources, and to help develop additional supports for stores who are identified as "high-risk" through compliance trajectory modeling.

Analyses will also evaluate whether elements of implementation differed across the city – for instance, whether low-income areas of the city experience more challenges in implementation or whether prices for staple foods changed as a result of the policy (and changed differentially across neighborhoods). Such measures can help determine whether policies like the SFO are successful in decreasing disparities in healthy food access, as they were intended to do.

Stakeholder relevance

Healthy retail policies and interventions can have impacts beyond diet and health outcomes, and these other outcomes may also be of interest, especially for certain stakeholders in policy, business, and the community. These stakeholders are often interested in economic features of an intervention, including cost-effectiveness, return on investment, and aggregate costs and benefits,[13] as well as an intervention's acceptability, implementation, reach, and uptake.[47] Independent evaluators can often provide estimates of these measures.

Although many implementation measures require primary data collection to assess (eg, facilitators and barriers from store owners in implementing a new policy), other measures can be estimated by models based on a review of high-quality research. For example, cost-effectiveness studies can be useful in identifying the most economic policies to achieve a particular policy goal, such as reducing obesity. For example, Gortmaker et al.[83] estimated the cost-effectiveness of a hypothetical national excise tax of 1 cent per ounce on all beverages with added caloric sweeteners. They estimated that over 10 years, the tax would result in a net savings of more than $14 million, as the costs of implementation (eg, tax agents, auditors) are easily offset by healthcare savings. Such estimates may resonate with policymakers, but businesses may be more interested in local jobs and economic performance. When industry stakeholders have argued that targeted taxes on foods or beverages will cause job losses, researchers have responded with estimates of the employment impact of these taxes.[84,85]

At the local level, the framework developed to evaluate the Healthy Food Financing Initiative (HFFI)[13] is useful in building the case for collecting smaller-scale metrics of community development, economic development, and job creation alongside metrics of health. In the case of the HFFI and other government-financed initiatives, these metrics are important in demonstrating the impact of investing public funds in the community. More generally, researchers conducting a natural experimental evaluation may wish to align their evaluation with existing policy questions to increase the likelihood the results are used by decision makers.[47]

Case study: Veggie Van mobile market. The Veggie Van is a mobile market program designed to deliver reduced-cost locally grown produce and education to lower-income individuals or communities with limited access to fresh fruits and vegetables. To pilot test the effectiveness of the program, the nonprofit organization that developed Veggie Van partnered with researchers at the University of North Carolina, Chapel Hill. The evaluation was structured to benefit both the development and implementation of the Veggie Van program and to assess its effectiveness in increasing fruit and vegetable access and consumption in the target population of low-income and/or food-insecure individuals.[62]

Specifically, the Veggie Van team needed to determine where and when the mobile markets should be open, and the research team wanted to assess the impact the mobile markets had on shoppers. The evaluation met both of these needs by using a multistage recruitment strategy[86] and thoughtful data collection procedures. First, partnerships were developed with community organizations such as health clinics, community colleges, and low-income housing developments that could potentially serve as host sties for the mobile markets. A memorandum of understanding was created between Veggie Van, the host site, and the university. The Veggie Van team had determined that sites would be viable if they had at least 25 people who were interested in participating; thus the research team planned their study recruitment to help estimate interest at each site. A coordinator within each partner organization recruited individuals who frequented the location on a regular basis and asked them to fill out a short questionnaire indicating their potential interest in using the Veggie Van if it were to come to the host site; the Veggie Van team used participants' responses to determine whether there was sufficient interest to support a Veggie Van at the site. The researchers built upon this existing data collection, adding new items to the questionnaire about the participant's ability to access fresh produce, their receipt of government assistance, and their willingness to be contacted by researchers. In addition, to accommodate the evaluation, the coordinator helped identify 30 interested customers to be recruited to participate in the evaluation study. The research team used these 2 pools of potential shoppers to recruit a baseline sample of likely Veggie Van users for the pre/post evaluation of the new markets. To ensure that the target population was represented, the researchers used a tiered recruitment strategy, first calling individuals who were receiving government assistance, then those who reported barriers to accessing fresh fruits and vegetables. These individuals were resurveyed 2–3 months after the Veggie Vans opened to assess changes in their fruit and vegetable consumption, health, and ability to purchase fruits and vegetables. This recruitment and data collection strategy helped meet the needs of e both the research team and the nonprofit operating the mobile markets.

CONCLUSION

Although there are many challenges to overcome in the use of natural experimental approaches to evaluate programs and policies in the food retail setting, designs that are based on thoughtful conceptual models can minimize bias and provide critical information to stakeholders, including the target population, policymakers, and food retailers. These stakeholders may seek information about whether the policy is associated with the desired outcomes in the short and long term, including shifts in food availability, purchases, and intake, as well as key process and compliance indicators. Evaluations can also capture policy and intervention pitfalls that can provide insight into strategic elements that may need to be strengthened.

As demonstrated in the case examples provided, planning evaluations require a number of trade-offs, so addressing a challenge in one area may require less capability to address challenges in other areas. For example, sales data acquired retroactively can give researchers access to baseline conditions over a long period of time, data they may not have to collect prospectively. But, because data were not collected for the purposes of the specific study, retrospective data may not contain the most precise or useful measures, which may require researchers to compromise some degree of measure validity.

Making trade-offs also often means balancing methodological rigor and practical considerations. For example, complete dietary assessments yield rich outcome data but are extremely resource-intensive to collect. Outcomes that are perhaps the most clinically relevant (eg, obesity) can be difficult to measure and may not change during a relatively short evaluation period; proxy measures that are meaningful and also modifiable within the study period must be identified.

Trade-offs are also relevant to the scope of an evaluation. There are myriad process and outcome measures that may be relevant, all of which may compete for resources. Thus, the scope of an evaluation framework will depend largely on cost, logistics, the degree of stakeholder engagement and collaboration, and the research team's capacity. Different types of measures are important in early, intermediate, and late stages of implementation. Thus, careful planning, including the development of conceptual diagrams that are created with input from (or consideration of) other stakeholders, will help identify priorities and define the scope of the evaluation, given the resources of the research team. Engaging stakeholders throughout the process will help foster an evaluation plan that is "meaningful, measurable and manageable."[13]

Finally, trade-offs may also manifest in the balance of internal and external validity of the study. When evaluating the impact of a new policy on a health impact, preserving internal validity (by representing an appropriate counterfactual and minimizing sources of bias) is critical for understanding whether the policy works. Natural experimental studies, which are always conducted in real-world settings, generally compromise some degree of internal validity when compared with studies in tightly controlled laboratory settings. More rigorous demonstration of external validity would require evaluations to be replicated across heterogeneous regions or among a range of subpopulations.[15] For example, HFFIs have been evaluated in a range of studies, but these evaluations typically focus on a single store in a single community. Although it can be tempting to assume that conclusions about the effectiveness of HFFI from these studies will apply to other communities, such generalizations may not be warranted.

Considering the continued interest in using the food retail setting as a lever to improve access, affordability, and selection of healthier foods, the increasing number of voluntary corporate pledges, and the proliferation of SSB taxes and other local, state, and federal regulatory options to improve diet and reduce obesity, the use of natural experimental studies to evaluate these policies is likely to increase over time. Because of their inherent design, natural experimental evaluations will almost certainly never be free from methodological concerns, especially as they relate to causality assessment. Most decisions regarding trade-offs are context-specific, meaning that there is no "right" study design decision. Careful consideration of key methodological challenges outlined in this article can help ensure that scholars, policymakers, and other key stakeholders are able to rigorously assess the impact of policies and programs on food access, availability, and intake and inform decisions about priorities and investments for the future.

Acknowledgments

The authors thank the Healthy Eating Research program of the Robert Wood Johnson Foundation, and especially the Healthy Food Retail working group, for their support. The authors are thankful for the Carolina Population Center for training support (T32 HD007168) and general support (P2C HD050924).

Author contributions. LTS and CEC co-led the Healthy Food Retail Methods work group that produced this manuscript. AG led the literature review on which the manuscript is based. All authors (LTS, AG, SF, DGT, LL, CEC) contributed to the conceptualization, drafting, content editing, and final review of this manuscript.

Funding. This work was supported by the Healthy Eating Research program of the Robert Wood Johnson Foundation, specifically the Healthy Food Retail work group. This work was also supported by the Carolina Population Center (T32 HD007168 and P2C HD050924).

Declaration of interest. The authors have no relevant interests to declare.

Supporting Information

The following Supporting Information is available through the online version of this article at the publisher's website.

Appendix S1 **Glossary of terms**

REFERENCES

1. United States Department of Health and Human Services, United States Department of Agriculture. Dietary Guidelines for Americans 2015–2020. 8th ed. 2015. http://health.gov/dietaryguidelines/2015/guidelines/. Accessed November 10, 2016.
2. Larson NI, Story MT, Nelson MC. Neighborhood environments: disparities in access to healthy foods in the US. Am J Prev Med. 2009;36:74–81.
3. Odoms-Young A, Singleton CR, Springfield S, et al. Retail environments as a venue for obesity prevention. Curr Obes Rep. 2016;5:184–191.
4. United States Department of Agriculture Economic Research Service. Access to Affordable and Nutritious Food: Measuring and Understanding Food Deserts and Their Consequences. 2009. https://www.ers.usda.gov/webdocs/publications/42711/12716_ap036_1_.pdf?v=41055. Accessed November 15, 2017.
5. Moore LV, Diez Roux AV, Nettleton JA, et al. Associations of the local food environment with diet quality—a comparison of assessments based on surveys and geographic information systems the multi-ethnic study of atherosclerosis. Am J Epidemiol. 2008;167:917–924.
6. Nettleton JA, Steffen LM, Mayer-Davis EJ, et al. Dietary patterns are associated with biochemical markers of inflammation and endothelial activation in the Multi-Ethnic Study of Atherosclerosis (MESA). Am J Clin Nutr. 2006;83:1369–1379.
7. United States Department of Agriculture Food and Nutrition Services. Enhancing Retailer Standards in the Supplemental Nutrition Assistance Program (SNAP). http://www.fns.usda.gov/snap/enhancing-retailer-standards-supplemental-nutrition-assistance-program-snap. Accessed December 15, 2016.
8. United States Department of Agriculture Food and Nutrition Services. Final Rule: Revisions in the WIC Food Packages. Women, Infants and Children (WIC). http://www.fns.usda.gov/wic/final-rule-revisions-wic-food-packages. Accessed December 15, 2016.
9. Minneapolis Health Department. Staple Foods Ordinance. 2008. http://www.minneapolismn.gov/health/living/eating/staple-foods. Accessed November 15, 2017.
10. Public Health Law Center. Creating Tribal Laws and Policies to Promote Healthy Eating. 2015. http://www.nihb.org/docs/phs_2015/Public%20Health%20Law%20and%20Policy/Creating%20Tribal%20Laws%20and%20Policies%20to%20Promote%20Healthy%20Eating.pdf. Accessed December 15, 2016.
11. White House Task Force on Childhood Obesity. Solving the Problem of Childhood Obesity Within a Generation. 2010. https://letsmove.obamawhitehouse.archives.gov/sites/letsmove.gov/files/TaskForce_on_Childhood_Obesity_May2010_FullReport.pdf. Accessed November 15, 2017.
12. Partnership for a Healthier America. Partnership for a Healthier America: Making the Healthy Choice the Easy Choice. http://ahealthieramerica.org/. Accessed January 18, 2017.
13. Fleischhacker SE, Flournoy R, Moore LV. Meaningful, measurable, and manageable approaches to evaluating healthy food financing initiatives: an overview of resources and approaches. J Public Health Manag Pract. 2013;19:541–549.
14. Dubowitz T, Ghosh-Dastidar MB, Collins R, et al. Food policy research: we need better measurement, better study designs, and reasonable and measured actions based on the available evidence. Obesity. 2013;21:5–6.
15. Marchand E, Stice E, Rohde P, et al. Moving from efficacy to effectiveness trials in prevention research. Behav Res Ther. 2011;49:32–41.
16. Patsopoulos NA. A pragmatic view on pragmatic trials. Dialogues Clin Neurosci. 2011;13:217–224.
17. Craig P, Cooper C, Gunnell D, et al. Using natural experiments to evaluate population health interventions: New Medical Research Council guidance. J Epidemiol Community Health. 2012;66:1182–1186.
18. Hunter CM, McKinnon RA, Esposito L. News from the NIH: research to evaluate "natural experiments" related to obesity and diabetes. Transl Behav Med. 2014;4:127.
19. Petticrew M, Cummins S, Ferrell C, et al. Natural experiments: an underused tool for public health? Public Health. 2005;119:751–757.

20. Abadie A, Diamond A, Hainmueller J. Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. J Am Stat Assoc. 2010;105.
21. Abadie A, Diamond A, Hainmueller J. Comparative politics and the synthetic control method. Am J Polit Sci. 2015;59:495–505.
22. Cawley J, Frisvold DE. The pass-through of taxes on sugar-sweetened beverages to retail prices: the case of Berkeley, California. J Policy Anal Manage. 2017;36:303–326.
23. Shadish WR, Cook TD, Campbell DT. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Belmont, CA: Wadsworth; 2002.
24. Ng SW, Slining MM, Popkin BM. The Healthy Weight Commitment Foundation Pledge: calories sold from U.S. consumer packaged goods, 2007–2012. Am J Prev Med. 2014;47:508–519.
25. Taillie LS, Ng SW, Popkin BM. Gains made by Walmart's Healthier Food Initiative mirror preexisting trends. Health Aff (Millwood). 2015;34:1869–1876.
26. Batis C, Rivera JA, Popkin BM, et al. First-year evaluation of Mexico's tax on nonessential energy-dense foods: an observational study. PLoS Med. 2016;13:e1002057.
27. Colchero M, Guerrero-López CM, Molina M, et al. Beverages sales in Mexico before and after implementation of a sugar sweetened beverage tax. PloS One. 2016;11:e0163463.
28. Wal-Mart Stores, Inc. Making Healthier Food a Reality for all. Our Commitments. http://corporate.walmart.com/global-responsibility/hunger-nutrition/our-commitments. Accessed December 15, 2016.
29. Taillie LS, Ng SW, Popkin BM. Global growth of "big box" stores and the potential impact on human health and nutrition. Nutr Rev. 2016;74:83–97.
30. Taillie LS, Ng SW, Popkin BM. Walmart and other food retail chains: trends and disparities in the nutritional profile of packaged food purchases. Am J Prev Med. 2016;50:171–179.
31. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. Am J Epidemiol. 2008;168:656–664.
32. Hogan JW, Lancaster T. Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. Stat Methods Med Res. 2004;13:17–48.
33. Dehejia RH, Wahba S. Propensity score-matching methods for nonexperimental causal studies. Rev Econ Stat. 2002;84:151–161.
34. Murnane RJ, Willett JB. Methods Matter: Improving Causal Inference in Educational and Social Science Research. Oxford University Press, New York; 2010.
35. Luke DA, Stamatakis KA. Systems science methods in public health: dynamics, networks, and agents. Annu Rev Public Health. 2012;33:357.
36. Homer JB, Hirsch GB. System dynamics modeling for public health: background and opportunities. Am J Public Health. 2006;96:452–458.
37. Leischow SJ, Milstein B. Systems thinking and modeling for public health practice. Am J Public Health. 2006;96:403–405.
38. Interagency Committee on Human Nutrition Research. National Nutrition Research Roadmap 2016-2021: Advancing Nutrition Research to Improve and Sustain Health. Washington, DC: Interagency Committee on Human Nutrition Research; 2016.
39. Dubowitz T, Zenk SN, Ghosh-Dastidar B, et al. Healthy food access for urban food desert residents: examination of the food environment, food purchasing practices, diet and BMI. Public Health Nutr. 2015;18:2220–2230.
40. Evans AE, Jennings R, Smiley AW, et al. Introduction of farm stands in low-income communities increases fruit and vegetable among community residents. Health Place. 2012;18:1137–1143.
41. Ompad DC, Galea S, Marshall G, et al. Sampling and recruitment in multilevel studies among marginalized urban populations: the IMPACT studies. J Urban Health. 2008;85:268–280.
42. Miller KW, Wilder LB, Stillman FA, et al. The feasibility of a street-intercept survey method in an African-American community. Am J Public Health. 1997;87:655–658.
43. Elbel B, Moran A, Dixon LB, et al. Assessment of a government-subsidized supermarket in a high-need area on household food availability and children's dietary intakes. Public Health Nutr. 2015;18:2881–2890.
44. Ver Ploeg M, Mancino L, Todd JE, et al. Where Do Americans Usually Shop for Food and How Do They Travel to Get There? Initial Findings from the National Household Food Acquisition and Purchase Survey. Washington, DC: US Department of Agriculture; 2015.
45. Dubowitz T, Ncube C, Leuschner K, et al. A natural experiment opportunity in two low-income urban food desert communities: research design, community engagement methods, and baseline results. Health Educ Behav. 2015;42(1 suppl): 87S–96S.
46. Dubowitz T, Ghosh-Dastidar M, Cohen DA, et al. Diet and perceptions change with supermarket introduction in a food desert, but not because of supermarket use. Health Aff (Millwood). 2015;34:1858–1868.
47. Ogilvie D, Cummins S, Petticrew M, et al. Assessing the evaluability of complex public health interventions: Five questions for researchers, funders, and policymakers. Milbank Q. 2011;89:206–225.
48. Caspi CE, Sorensen G, Subramanian SV, et al. The local food environment and diet: a systematic review. Health Place. 2012;18:1172–1187.
49. Lytle L, Sokol RL. Measures of the food environment: a systematic review of the field, 2007–2015. Health Place. 2017;44:18–34.
50. McKinnon RA, Reedy J, Morrissette MA, et al. Measures of the food environment: a compilation of the literature, 1990–2007. Am J Prev Med. 2009;36: S124–S133.
51. Andreyeva T, Luedicke J. Incentivizing fruit and vegetable purchases among participants in the Special Supplemental Nutrition Program for Women, Infants, and Children. Public Health Nutr. 2015;18:33–41.
52. Andreyeva T. Effects of the revised food packages for Women, Infants, and Children (WIC) in Connecticut. Choices. 2012;27:1–6.
53. Zenk SN, Odoms-Young A, Powell LM, et al. Fruit and vegetable availability and selection: rederal food package revisions, 2009. Am J Prev Med. 2012;43:423–428.
54. Lu W, McKyer ELJ, Dowdy D, et al. Evaluating the influence of the revised Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) food allocation package on healthy food availability, accessibility, and affordability in Texas. J Acad Nutr Diet. 2016;116:292–301.
55. Tisone CA, Guerra SA, Lu W, et al. Food-shopping environment disparities in Texas WIC vendors: a pilot study. Am J Health Behav. 2014;38:726–736.
56. Black AP, Brimblecombe J, Eyles H, et al. Food subsidy programs and the health and nutritional status of disadvantaged families in high income countries: a systematic review. BMC Public Health. 2012;12:1.
57. Thompson FE, Subar AF. Dietary assessment methodology. In: Coulston A, Boushey C, eds. Nutrition in the Prevention and Treatment of Disease. Vol 2. Academic Press. 2008:3–39.
58. Basu S, Seligman HK, Gardner C, et al. Ending SNAP subsidies for sugar-sweetened beverages could reduce obesity and type 2 diabetes. Health Aff (Millwood). 2014;33:1032–1039.
59. Wang Y, Coxson P, Shen Y-M, et al. A penny-per-ounce tax on sugar-sweetened beverages would cut health and cost burdens of diabetes. Health Aff (Millwood). 2012;31:199–207.
60. Humphreys DK, Panter J, Sahlqvist S, et al. Changing the environment to improve population health: a framework for considering exposure in natural experimental studies. J Community Health. 2016:70:941–946.
61. Cummins S, Flint E, Matthews SA. New neighborhood grocery store increased awareness of food access but did not alter dietary habits or obesity. Health Aff (Millwood). 2014;33:283–291.
62. Leone LA, Haynes-Maslow L, Ammerman AS. Veggie Van pilot study: impact of a mobile produce market for underserved communities on fruit and vegetable access and intake. J Hunger Environ Nutr. 2017;12:89–100.
63. Kegler MC, Alcantara I, Veluswamy J, et al. Results from an intervention to improve rural home food and physical activity environments. Prog Community Health Partnersh Res Educ Action. 2012;6:265.
64. Fletcher JM, Frisvold DE, Tefft N. The effects of soft drink taxes on child and adolescent consumption and weight outcomes. J Public Econ. 2010;94:967–974.
65. Fletcher JM, Frisvold DE, Tefft N. Non-linear effects of soda taxes on consumption and weight outcomes. Health Econ. 2015;24:566–582.
66. Sturm R, Powell LM, Chriqui JF, et al. Soda taxes, soft drink consumption, and children's body mass index. Health Aff (Millwood). 2010;29:1052–1058.
67. Finkelstein EA, Zhen C, Nonnemaker J, et al. Impact of targeted beverage taxes on higher- and lower-income households. Arch Intern Med. 2010;170: 2028–2034.
68. Colchero MA, Salgado JC, Unar-Munguía M, et al. Changes in prices after an excise tax to sweetened sugar beverages was implemented in Mexico: evidence from urban areas. PLoS One. 2015;10:e0144408.
69. Falbe J, Rojas N, Grummon AH, et al. Higher retail prices of sugar-sweetened beverages 3 months after implementation of an excise tax in Berkeley, California. Am J Public Health. 2015;105:2194–2201.
70. Ng S, Silver L, Ryan-Ibarra S, et al. Six-month evaluation of the Berkeley sugar sweetened beverage tax: prices, purchases and store revenues. 2016.
71. Falbe J, Thompson HR, Becker CM, et al. Impact of the Berkeley excise tax on sugar-sweetened beverage consumption. Am J Public Health. 2016;106:1865–1871.
72. Silver LD, Ng SW, Ryan-Ibarra S, et al. Changes in prices, sales, consumer spending, and beverage consumption one year after a tax on sugar-sweetened beverages in Berkeley, California, US: A before-and-after study. PLoS Med. 2017 Apr 18;14:e1002283.
73. McCormack LA, Laska MN, Larson NI, et al. Review of the nutritional implications of farmers' markets and community gardens: a call for evaluation and research efforts. J Am Diet Assoc. 2010;110:399–408.
74. Ulmer VM, Rathert AR, Rose D. Understanding policy enactment: the New Orleans Fresh Food Retailer Initiative. Am J Prev Med. 2012;43(3 suppl 2):S116–S122.
75. Bassett MT, Dumanovsky T, Huang C, et al. Purchasing behavior and calorie information at fast-food chains in New York City, 2007. Am J Public Health. 2008;98:1457–1459.
76. Ohri-Vachaspati P, Turner L, Chaloupka FJ. Fresh fruit and vegetable program participation in elementary schools in the United States and availability of fruits and vegetables in school lunch meals. J Acad Nutr Diet. 2012;112:921–926.

77. Akin JS, Guilkey DK, Popkin BM. The school lunch program and nutrient intake: a switching regression analysis. Am J Agric Econ. 1983;65:477–485.
78. Moore GF, Audrey S, Barker M, et al. Process evaluation of complex interventions: Medical Research Council guidance. BMJ. 2015;350:h1258.
79. Falbe J, Grummon A, Rojas N, et al. Implementation of a sugar-sweetened beverage excise tax in Berkeley, California and lessons learned. 2016. https://apha.confex.com/apha/144am/meetingapp.cgi/Paper/360266. Accessed November 15, 2017.
80. Chrisinger B. A mixed-method assessment of a new supermarket in a food desert: contributions to everyday life and health. J Urban Health Bull N Y Acad Med. 2016;93:425–437.
81. Morland KB. An evaluation of a neighborhood-level intervention to a local food environment. Am J Prev Med. 2010;39:e31–e38.
82. Caspi CE, Pelletier JE, Harnack L, et al. Differences in healthy food supply and stocking practices between small grocery stores, gas-marts, pharmacies and dollar stores. Public Health Nutr. 2016;19:540–547.
83. Gortmaker SL, Wang YC, Long MW, et al. Three interventions that reduce childhood obesity are projected to save more than they cost to implement. Health Aff Proj Hope. 2015;34:1932–1939.
84. Powell LM, Wada R, Persky JJ, et al. Employment impact of sugar-sweetened beverage taxes. Am J Public Health. 2014;104:672–677.
85. Wada R. Employment Impacts of Alcohol Taxes. 2014. https://apha.confex.com/apha/142am/webprogram/Paper306417.html. Accessed May 17, 2017.
86. Tripicchio GL, Grady Smith J, Armstrong-Brown J, et al. Recruiting community partners for Veggie Van: strategies and lessons learned from a mobile market intervention in North Carolina, 2012–2015. Prev Chronic Dis. 2017;14:E36.