THE
MILBANK QUARTERLY
A MULTIDISCIPLINARY JOURNAL OF POPULATION HEALTH AND HEALTH POLICY

Assessing the Evaluability of Complex Public Health Interventions: Five Questions for Researchers, Funders, and Policymakers

DAVID OGILVIE, STEVEN CUMMINS, MARK PETTICREW, MARTIN WHITE, ANDY JONES, and KATHRYN WHEELER

Medical Research Council Epidemiology Unit, Cambridge; University of London; Newcastle University; University of East Anglia

Context: Evidence to support government programs to improve public health often is weak. Recognition of this "knowledge gap" has led to calls for more and better evaluation, but decisions about priorities for evaluation also need to be addressed in regard to financial restraint.

Methods: Using England's Healthy Community Challenge Fund as a case study, this article presents a set of questions to stimulate and structure debate among researchers, funders, and policymakers and help make decisions about evaluation within and between complex public health interventions as they evolve from initial concept to dissemination of full-scale intervention packages.

Findings: This approach can be used to identify the types of knowledge that might be generated from any evaluation, given the strength of evidence available in response to each of five questions, and to support a more systematic consideration of resource allocation decisions, depending on the types of knowledge required.

Conclusions: The principles of this approach may be generalizable, and should be tested and refined for other complex public health and wider social interventions.

Keywords: Evaluation, complex interventions, public health, obesity.

Address correspondence to: David Ogilvie, MRC Epidemiology Unit and UKCRC Centre for Diet and Activity Research (CEDAR), Box 296, Institute of Public Health, Forvie Site, Robinson Way, Cambridge, CB2 0SR, UK (email: david.ogilvie@mrc-epid.cam.ac.uk).

The Milbank Quarterly, Vol. 89, No. 2, 2011 (pp. 206–225)

Government policies and programs to improve public health can often be regarded as complex interventions in that they typically involve the flexible or tailored implementation of multiple interacting activities in a variety of settings to change the population's behavior and improve its health (Craig et al. 2008). Evidence to support their development and implementation, however, frequently is weak. Recognition of this "knowledge gap" has led to repeated calls for more and better evaluation of these complex "natural experiments" (House of Commons Health Committee 2009; Sallis, Story, and Lou 2009; Wanless 2004). Such evaluation would be expected to provide more robust assessments of the overall impact, effectiveness, and cost-effectiveness of different approaches and to improve our understanding of how they work (or do not work) and how their effects are distributed within populations (Craig et al. 2008).

In the United Kingdom, revised guidance to support the development and evaluation of these kinds of complex interventions has recently been published by the Medical Research Council (MRC) (Craig et al. 2008). Although few experts may disagree in principle with the evaluative "call to arms," its implementation nonetheless raises a number of scientific, practical, and prioritization issues. The original MRC guidance, which drew on examples from public health and health services research, focused on the challenges both of developing and defining interventions composed of several components and of evaluating such interventions using randomized controlled trials (Medical Research Council 2000). The guidance's subsequent revision reflects attempts to address additional challenges, such as the fact that some interventions are driven largely by political rather than scientific considerations. These interventions may not be easily accommodated in the original guidance's framework, which describes a phased, theory-based approach to the development of interventions as well as to their evaluation (Craig et al. 2008; Medical Research Council 2000; Ogilvie et al. 2009). The current fiscal climate brings a further challenge, in that evaluative aspirations need to be tackled in light of the greater constraints on research funding and research capacity (Leviton et al. 2010). Writing about the existing evaluation guidance for health promotion, Windsor and colleagues noted the need to ensure that an evaluation is "realistic, prudent and efficient" because "every component of a program usually cannot and, in most cases, should not be evaluated" (Windsor et al. 2004, 18). This

observation emphasizes that the best way to deploy scarce resources for evaluation must be determined in order to maximize the new learning and evidence produced.

Researchers, policymakers, funders, and commissioners of research have a few simple tools to help them make these decisions. One such established tool is evaluability assessment (Wholey 1979), a systematic method of determining whether a given intervention program is ready for evaluation and how such an evaluation would help improve the program. Evaluability assessment also can be used to help plan and deliver interventions, for example, by providing feedback to staff on their implementation. The process of evaluability assessment has recently been reviewed elsewhere (Leviton et al. 2010).

Evaluability assessment usually depends on articulating a specific theory of change used for the particular intervention under consideration. But specifying a theory may not always be possible when decisions about allocating evaluative resources need to be made, particularly when choosing between interventions or their components. We therefore will draw on the inspiration of evaluability assessment to develop a more general approach to considering what types of knowledge might be expected to be generated at different points in the evolution of a complex public health intervention and how these evaluative resources might best be deployed. Our approach focuses on the scientific principles to be considered when answering these questions rather than on the details of the processes involved.

As a case study, our article uses the Healthy Community Challenge Fund (HCCF), a government-funded program of interventions to modify the obesogenic environment in nine demonstration "Healthy Towns" in England (Department of Health 2008b). Although the HCCF is used as an example, our approach is likely to be generalizable and should therefore be capable of being tested, refined, and applied to other complex public health and social policy interventions. We begin by introducing the HCCF and the three general evaluative decisions to be made about complex interventions of this kind. Then we describe two components of the overall HCCF program to serve as worked examples. After setting the scene, we outline the three general principles underpinning our approach before presenting the core of our article, the "five questions" of its title. We then return to our worked examples and show how considering the five questions can help in making the three evaluative decisions required.

Given the variable use of terminology in this area, we should point out that in this article, we use the term *intervention* to refer to large-scale, complex packages of measures (such as the overall HCCF program) as well as to more discrete projects or components within those packages (such as the worked examples).

The Healthy Community Challenge Fund

The Healthy Community Challenge Fund (HCCF) is a £30 million (US$48 million) investment program funded by the Labour government (1997–2010) as part of its Healthy Weight, Healthy Lives obesity strategy for England (Department of Health 2008b). Local governments and Primary Care Trusts (local National Health Service [NHS] organizations responsible for public health) representing towns and cities of any size in England were invited to bid jointly for funds of up to £5 million (US$8 million) each "to build on existing work in their communities and test ideas on what further action needs to happen to make physical activity and healthy food choices easier for people" (Department of Health 2008a). On the basis of the proposals submitted, nine "Healthy Towns" were selected for funding: a London borough (Tower Hamlets), three cities (Manchester, Portsmouth, and Sheffield), three large towns (Dudley, Halifax, and Middlesbrough) and two smaller towns (Tewkesbury and Thetford).

The Healthy Weight, Healthy Lives strategy makes clear that the Healthy Towns program is intended to "test and validate holistic approaches to promoting physical activity and improving diet," inspired by the example of EPODE ("Ensemble prévenons l'obésité des enfants": "Together, let's prevent obesity in children"), a French program of community-based coalitions of education, health, retail, media, and other stakeholders to reduce childhood obesity (Department of Health 2008b; Westley 2007). The expectation of Healthy Weight, Healthy Lives was that Healthy Towns would "invest in infrastructure improvements that implement the lessons of a variety of programmes. . . combined with efforts to galvanise local members of the community to take action to change both food and activity habits" (Department of Health 2008b, 22). These nine towns interpreted this expectation in various ways, reflecting differences not only in specific local health concerns (such as the needs of particular immigrant communities, as in Tower

Hamlets or Thetford) but also in the modes of intervention emphasized in the proposals (such as individual incentives in Manchester or healthy urban planning in Sheffield). Each Healthy Town project team prepared a logic model summarizing how its overall program of activities was expected to work. The nine Healthy Towns specified well over two hundred individual interventions, some of which were expected to be initiated almost immediately. Because we had been commissioned to evaluate the Healthy Towns program from a national perspective, we needed a framework to guide three critical evaluative decisions within this complex intervention landscape: first, to identify which elements could be evaluated; second, to decide which elements should be evaluated first; and third, to clarify from what perspective those elements should be evaluated in order to produce the most useful new evidence for science and policy (Nutbeam 1998).

Examples of "Healthy Town" Interventions

Two examples based on real interventions in the Healthy Towns illustrate the utility of our approach.

Family Health Hubs

A key element of the program for the Dudley Healthy Town is the development of five "family health hubs": new buildings with exercise equipment and activities for families, located in parks, staffed by activity rangers, and accessed by improved infrastructure for walking and cycling from nearby local neighborhoods and schools (Dudley Healthy Towns 2010). The sites were selected on the basis of size, purpose, existing usage, geographical spread, and the feasibility of improving access by walking and cycling. The hubs provide a range of instructor-led fitness classes, access to low-maintenance outdoor gym equipment, and healthy eating and physical activity events organized by rangers in the local area (such as guided walks, dance classes, healthy-eating barbecues, and cooking demonstrations). The overall aim is to provide a family-friendly community "hub" for behavior change linked to obesity prevention in both adults and children. This intervention should provide opportunities to contribute new evidence on, for example, the impact

of making recreational facilities available closer to home, the effect of providing new infrastructure for walking and cycling, and the potential to change social norms concerning the use of structured opportunities for physical activity and healthy eating in children living in deprived neighborhoods.

Healthy Urban Planning

A key element of the program for the Thetford Healthy Town is the redesign of the existing built environment to encourage health-promoting behaviors (Breckland Council 2009; Thetford Healthy Town 2010). Thetford has "growth-point" status, meaning that the town is expected to grow rapidly over the next twenty years, thus increasing the demand for new housing, transport, and community infrastructure. This intervention is therefore a long-term strategic activity to ensure that "health" is fully incorporated into the urban design and planning policy related to future growth and regeneration. Mechanisms by which this may occur include incorporating health in the strategic master plan, ensuring that active travel is reflected in the local transport plan, and embedding public health principles into the assessment of planning applications and the design of new neighborhoods and communities. As such, this is a long-term aspiration related to the gradual enlargement of the town (Moving Thetford Forward 2010), a timescale well beyond the horizons of most evaluation teams.

Three Underpinning Principles

Testing Theories Rather Than Interventions

The problem of how to evaluate complex social programs is not new. One major impediment highlighted in a recent report of the UK House of Commons Health Committee is the tendency for new initiatives to be introduced quickly, without sufficient opportunity to collect baseline "before" data from the populations served by the initiatives (House of Commons Health Committee 2009). The Healthy Towns program was specifically identified in that report as having been introduced in a way that "is likely to make [rigorous evaluation] impossible." Moreover,

without a national surveillance system able to track changes in dietary and physical activity behavior and anthropometric measures (other than height and weight in selected school-year groups) at the local level, the likelihood of being able to evaluate the overall outcomes of the Healthy Towns through a before-and-after study of the target populations appears remote. In addition, the Healthy Towns encompass multiple dimensions of contextual variation that may affect the outcome of a given project within the overall program. Taking as an example the construction of a new cycle path, these may include the epidemiological context (e.g., the local prevalence of cycling), the environmental context (e.g., the congeniality of the local topography and climate for cycling), the sociopolitical context (e.g., the local political balance of power between "green" and "pro-car" views), and the municipality's organizational context (e.g., whether cycling is regarded as falling under the jurisdiction of a roads department, a parks and recreation department, or a dedicated cycling unit). Drawing on the insights of realistic evaluation developed by Pawson and Tilley, we adopted the principle that rather than testing whether programs of this type "work" in an aggregate or generalizable sense, evaluative research in this area should test more general theories about how interventions work by aggregating the evidence for and against such theories across a range of situations or "context-mechanism-outcome" (CMO) configurations (Pawson and Tilley 1997).

Seeing the "Big Picture" of an Intervention Program

Second, the original concept of the CMO configuration implies a somewhat linear relationship among context, mechanism, and outcome, but in practice, these relationships are likely to be more complex. For example, in one town, certain actions (such as the appointment of a "healthy urban planning" officer) must take place first in order to change the organizational context, thereby preparing the ground for the mechanism of a second wave of actions (such as the adoption of new planning guidelines). But in another town, early adopters of an intervention (such as parents and children who accept the offer of a "walking bus," an escorted group of children walking to school) may help alter perceived social

norms about traveling to school in their community, thereby changing the context and catalyzing the mechanism by stimulating further adoption of the intervention by other families (Lorenc et al. 2008). It therefore may be unhelpful to regard any particular intervention project in isolation. In this article, we use the term *intervention* not only in the narrow sense of providing people with resources that might help change their behavior—such as a healthy-eating leaflet, an exercise referral scheme, or a new cycle path—but in the wider sense of "events in systems that either leave a lasting footprint or wash out depending on how well the dynamic properties of the system are harnessed" (Hawe, Shiell, and Riley 2009, 270). For obesity prevention, the latter definition might encompass less immediately tangible actions, such as efforts to change established ways of doing things in local statutory agencies, for example, by attempting to influence the priorities of a planning authority with respect to neighborhood design, or the priorities of an education authority with respect to physical education provision in schools.

Favoring Judgment over Checklists

Our third principle reflects the tension between an understandable desire by various evaluation stakeholders for criteria, checklists, and scoring systems, on the one hand, and the need for a more nuanced consideration of the scientific issues, on the other. Our approach builds on the Standard Evaluation Framework (SEF) produced by the National Obesity Observatory (NOO) for England (National Obesity Observatory 2009), which is made up mainly of a checklist of issues to be considered when evaluating weight management or weight loss interventions targeted at individuals. But our approach has a somewhat different purpose. We retained the fundamental objective of assessing "evaluability," but we also drew on Bradford Hill's concept of "viewpoints" (Bradford Hill 1965). Bradford Hill explicitly cautioned against any attempt to treat his principles—developed as an aid to appraising the evidence for putative causal associations in etiological epidemiology—as a checklist of necessary or sufficient criteria. We adopted the same position in developing our five questions. Although the questions need to be considered, the answers are not necessarily straightforward, and the list of questions does not replace the need for judgment regarding the weights

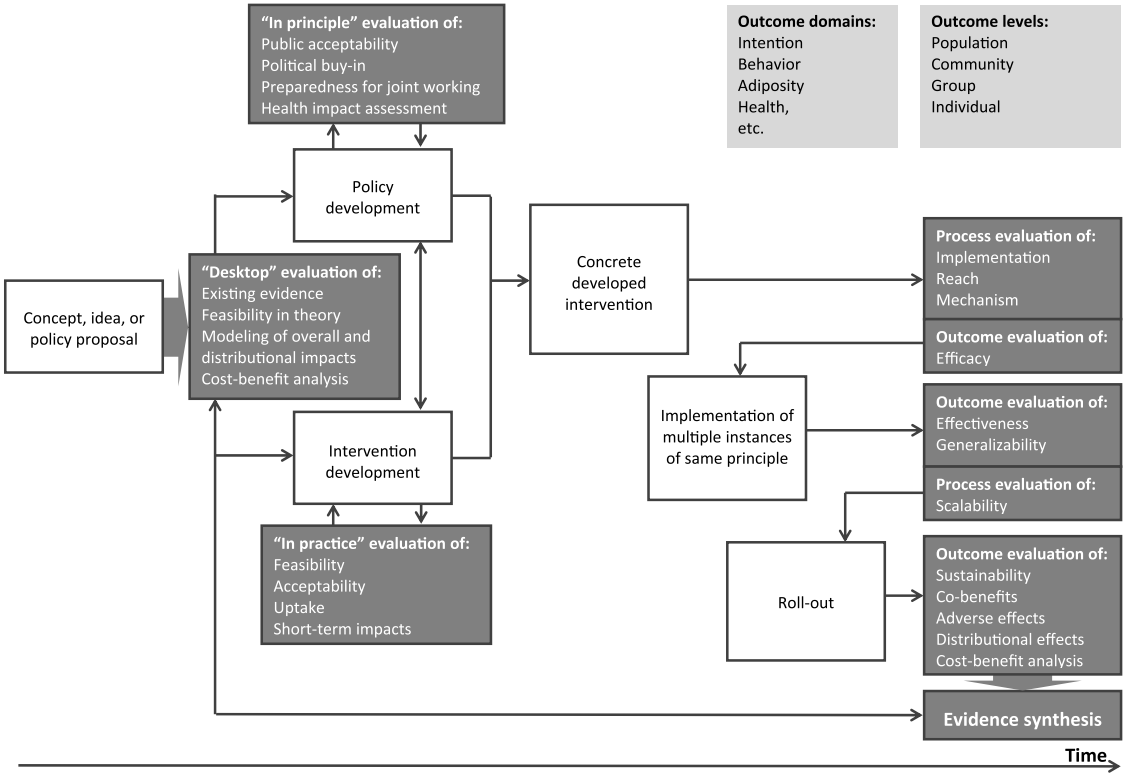
that should be applied to the various points. Nonetheless, the stronger the evidence is for each of the points, the stronger the case will be for evaluation.

Five Questions to Assess Evaluability

We now turn to the key questions and propose that decisions by researchers, funders, and policymakers regarding priorities for evaluation take into account the answers.

Where Is a Particular Intervention Situated in the Evolutionary Flowchart of an Overall Intervention Program?

To clarify the value and nature of a potential evaluative study of a given intervention, it may be useful to situate it in an evolutionary flowchart (see figure 1). Although this flowchart builds on previous flowcharts (Craig et al. 2008; Medical Research Council 2000; Nutbeam 1998), it also aims to better reflect the wider sociopolitical context in which complex public health programs take place, the importance of considering the degree to which an intervention is embedded in organizational contexts (Yin 1979), and the fact that the evaluative opportunities evolve in tandem with evolution of the intervention. Put simply, evaluating an intervention too early in its development risks reaching unhelpful or misleading conclusions—such as by concluding that no tangible effects have occurred—but this risk can be averted if the correct evaluative questions are asked (Nutbeam 1998). For example, the evaluation of the early stage of a new community gardening project might focus on understanding the process by characterizing the incentives for and barriers to participation, the socioeconomic profile of the participants and the success of their growing efforts, rather than measuring changes in behavioral or health-related outcomes such as eating habits or nutritional biomarkers, which are unlikely to be detected without both a larger sample of participants and a longer period of detailed follow-up.



Open boxes represent different stages in the evolution of an intervention. Shaded boxes summarize key evaluable constructs at different stages. Constructs listed down right-hand side could all apply to all stages from "concrete" to "roll-out," but the likelihood of being able to measure the items lower down the list, or generalize from the observations, increases as implementation proceeds from "concrete" through "multiple instances" to "roll-out" (Nutbeam 1998).

FIGURE 1. Evolutionary Flowchart for Typical Complex Public Health Intervention.

How Will an Evaluative Study of This Intervention Affect Policy Decisions?

This question reflects the initial assumption that "conducting evaluations of programs that are useful to decision makers is the hallmark of successful evaluation" (Trevisan and Huang 2003), implying that if an outcome of evaluative research has no identifiable "customer" who cares about the results, then it may not be worth conducting. A proposed evaluative study may appear to "fail" this test if (1) the results would have no bearing on any current or predictable policy question, (2) the key policy decisions that could have been informed by the results will be made before the results are known, or (3) the study is examining the health impacts of an intervention intended primarily to achieve a policy goal in another sector, such as education or transport. Closer engagement between the producers and the users of research, and between policymakers in different sectors, is likely to improve the utility of research in this regard by helping bring the streams of "problems, policies and politics" into alignment (Exworthy 2008; Ogilvie et al. 2009). In less clear-cut situations, a more nuanced assessment may be required to estimate the "policy prior" on the question (in other words, what policymakers are likely to do or recommend on the basis of what they already know) and the likelihood that those actions or recommendations will change as a result of a new evaluative study. It is important to note that this analysis should not be limited to an assessment of policymakers' potential reactions to new data on efficacy or effectiveness. Rather, in some situations, the key new data required to alter the "policy post" may relate to the acceptability, implementation, reach, uptake, mechanism, or dissemination of an intervention (Bond et al. 2010; Mackenzie et al. 2010) or to a cost-effectiveness or cost-benefit analysis (Wanless 2004).

Conversely, neither researchers nor funders should allow themselves to be shackled to an excessively instrumental or pragmatic view of the value of research, because that may sometimes conflict with the wider public interest. Aside from the potential (and sometimes serendipitous) value of "blue skies" research in general, even in a highly applied context, researchers should bear in mind that some studies may identify important evidence of adverse effects that is unwelcome to certain stakeholders, or they may provide answers to policy questions that are not currently being asked but turn out to become highly topical after the study is finished. Indeed, research with policymakers has shown that

some "historical" research findings can have a remarkably long shelf life in terms of their influence on their thinking (Petticrew et al. 2004).

What Are the Plausible Sizes and Distribution of the Intervention's Hypothesized Impacts?

Epidemiologists use the concept of population attributable risk to quantify the importance of a risk factor for the population at large. For instance, a particular risk factor may greatly increase the risk of disease. But if only a small proportion of the population is exposed to that risk factor, the overall impact on the population will be small, whereas another risk factor that confers only a modest increase in risk may have a greater overall impact on the population if a high proportion of the population is exposed to that factor (Last 2001). The same principle can be applied to selecting interventions for evaluation for public health purposes. All other things being equal, the interventions most worthy of evaluation would be those expected to have a large effect on a large number of people. To put it another way, the overall public health impact of, say, a small local community gardening scheme in which only twenty people take part will be minimal, however life changing its effects are on the individuals who do take part. It would, of course, be important to base such a "plausibility analysis" (Leviton et al. 2010) on a realistic-effect size, drawing, for example, on evidence from previous studies of effectiveness rather than efficacy, or from aggregated evidence for and against similar theory-based approaches to changing other health-related behaviors, rather than on the unsubstantiated claims or aspirations of the agents promoting a given intervention.

Nonetheless, small-scale, novel, or untested interventions should not be automatically ruled out of the evaluative court on the basis of a small estimated health impact on a population. Indeed, the case for investing evaluative resources in such situations may be strengthened if they are predicted to have important adverse effects, benefits in other policy sectors, or differential effects that may contribute to widening or narrowing health inequalities; if they are scalable to widespread implementation; or if they contain a particularly novel and promising germ of an idea that could lead to an entirely new class of intervention. In the last case, however, the novel approach should be cross-referenced with existing theory as a means of distinguishing the brilliant insight from the ludicrous scheme before plunging in, in much the same way that Bradford Hill

highlighted the importance of identifying a plausible mechanism before treating evidence of an association between a disease and a putative risk factor with excessive respect (Bradford Hill 1965).

How Will the Findings of an Evaluative Study Add Value to the Existing Scientific Evidence?

The mere fact that a new intervention is being introduced does not in itself mean that this intervention must be evaluated, at least not in the scientific sense. Scientifically evaluating a project is completely different from auditing its execution and financial management, which may well be required by the funders of an intervention but contributes little to scientific knowledge. In many cases, it may be sufficient merely to record that the project has taken place and to describe and characterize its content as part of the overall package of measures. In fact, conducting further evaluative research on already well-researched interventions in the absence of genuine equipoise or uncertainty concerning their effectiveness may also be considered unethical in principle (Freedman 1987).

Having said that, even if the effectiveness of an intervention is unequivocally established with respect to one particular outcome, critical uncertainty may remain regarding its effects in settings or on populations different from those in which it was previously studied; its effects on other outcomes; the mechanisms by which its effects are achieved; or its scalability, sustainability, generalizability, or distributional effects. For example, interventions that improve health in an aggregate sense may widen inequalities in health if their uptake or effectiveness is socially patterned (Thomson et al. 2004; White, Adams, and Heywood 2009). Such considerations may be particularly applicable to interventions whose impacts could span multiple health and nonhealth domains or policy sectors. The added scientific value of one more study (and the interpretation of its results) therefore should be assessed in the context of the accumulated existing evidence pertaining to the same question (Ogilvie et al. 2009; Wilson et al. 2008), whether that evidence is aggregated across multiple instances within one program (such as Healthy Towns) or across a more general set of intervention studies. This in turn may permit more generalized causal inference from each of a group of studies specific to its local context. It is important to note that this is

not the same as assessing the value of a new study for informing a policy decision, since those making policy decisions may be influenced more by certain key findings of research than by the weight and methodological rigor of scientific evidence as synthesized in a systematic review (Petticrew et al. 2004).

To return to the example of a community gardening scheme, the intervention's overt focus may be on its direct effects on the participants' food supply or dietary behavior, but it may also produce benefits for other health outcomes (e.g., by increasing physical activity through gardening) and in other policy sectors (e.g., by contributing to environmental sustainability by reducing the frequency of car-based food-shopping trips). The opportunity to evaluate wider impacts of this kind may tip the balance in favor of investing evaluative resources, even if the answers to the more obvious evaluative questions are deemed to be sufficiently clear already.

Is It Practicable to Evaluate the Intervention in the Time Available?

Particularly when the design, allocation, or delivery of interventions is outside the researchers' control, some interventions may be introduced too early to permit a meaningful outcome-oriented evaluation (e.g., by leaving insufficient time to collect baseline data) or too late (e.g., by leaving insufficient time to collect and analyze data before the end of the researchers' contracts). In some cases, it may be possible to bypass such limitations by using routinely collected surveillance data. Indeed, this may be a more feasible and more affordable solution than collecting original data, although for many health-related behaviors the surveillance data available may be of unknown validity, insufficiently precise to detect change, or subject to other major limitations in their interpretation (Stamatakis, Ekelund, and Wareham 2007).

But even if sufficient time and resources are available for a meaningful evaluation of an intervention's long-term effects, it does not necessarily follow that anyone will be interested in the results by the time the intervention has been fully implemented and evaluated. Another risk of long-term evaluation is that the "signal" of the effects of one specific intervention may become lost in the "noise" of everything else that has happened in the meantime, particularly if the implementation of the

intervention or the study design does not provide for a counterfactual or control group.

In all these situations, researchers may need to consider reconfiguring apparently simple evaluative research questions to suit the practical realities of the situation. For example, if the content of an intervention and the contexts in which it operates change over time (Hawe, Shiell, and Riley 2009; Shepperd et al. 2009), the focus of evaluation may shift from comparing the effects of "intervention" versus "control" to comparing the effects of different "doses" of intervention, or the effects of the intervention in different contexts.

Two Worked Examples

We now return to our worked examples to illustrate how the approach we have outlined may help guide evaluative decisions. Although these examples are based on real interventions, the working is hypothetical.

Family Health Hubs

At first glance, this kind of intervention provides the opportunity to evaluate the physical activity benefits of providing recreational and other facilities in parks and an improved infrastructure for walking and cycling in local neighborhoods. The research recommendations published by the National Institute for Health and Clinical Excellence (NICE) identified both of these as important *potential* means of improving health (question 3: plausible effects) for which little robust evidence currently exists (question 4: contribution to scientific evidence) (National Institute for Health and Clinical Excellence 2008). However, reflection prompted by question 1 (evolution of the intervention) may indicate that this is a new and untried type of intervention currently located at the "intervention development" stage of the evolutionary flowchart, and careful investigation of the local context prompted by question 2 (relevance to policy decisions) may show that concerns about improving community safety and reducing antisocial behavior take precedence over concerns about the hubs' impact on physical activity. It therefore may follow that a qualitative evaluation of the users' and nonusers' safety perceptions of facilities in parks with and without one of the new hubs might provide

the single most influential piece of evidence to justify future investments in parks and recreational services.

Further "proof of concept" evidence from early, small-scale implementations of the "hub" concept—demonstrating at least an acceptable level of adoption and usage by the local population, as shown in the "in practice" evaluation box in the figure—may be sufficient to justify the funding of further roll-out. In turn, this would lead, in time, to an opportunity to mount a more considered, outcome-oriented evaluation (question 5, time available) examining the influence of the hub's location, content, and level of usage on changes in families' physical activity behavior, perhaps stratified by household or area-level socioeconomic status. It may then be possible to combine evidence derived from the family health hubs with evidence from evaluations of other types of local physical activity provision to test more general theories about the role of the physical environment in promoting active living.

Healthy Urban Planning

The long-term health impact on a population of a major redesign of the built environment is both unknown (question 4, contribution to scientific evidence) and potentially substantial (question 3, plausible effects), but very difficult to quantify using short-term empirical data (question 5, time available). In the early years (question 1, evolutionary flowchart), it may therefore be more appropriate to evaluate the degree to which the principles of "healthy urban planning" espoused in the original proposal turn out to be expressed and embedded in organizational practice (Hawe, Shiell, and Riley 2009; Nutbeam 1998; Yin 1979) and to develop a theory of how this type of intervention is expected to work in the longer term. Short-term research questions might therefore include how widely across the municipality the principles have become evident (which might be addressed by seeking evidence of partnership working between education, health, planning, and transport departments), or how deeply the principles have penetrated routine practice (which might be addressed through an ethnographic study of the proceedings of planning committees or by auditing planning decisions against evidence-based guidance such as that issued by NICE (National Institute for Health and Clinical Excellence 2008). Evidence of this kind may be crucial to building support for maintaining the general policy direction or promoting its more widespread adoption (question 2, relevance to policy decisions).

Conclusions

We developed a set of questions to help guide decisions about the use of evaluative resources within and between complex public health interventions. We assumed that the purpose of evaluation is to contribute to the scientific understanding of the process and impacts of interventions rather than merely to justify their existence. Unlike some previous approaches to assessing evaluability, our approach involves neither a checklist of criteria (National Obesity Observatory 2009) nor a structured linear process (Trevisan and Huang 2003). Instead, it is a set of questions to be considered, debated, and reiterated (Leviton et al. 2010) by researchers, funders, and policymakers as a complex intervention evolves from its initial concept through the developmental and pilot stages to the testing, refinement, and dissemination of a concrete intervention program. Although our approach was developed in one particular context—that of England's Healthy Community Challenge Fund—the principles may be generalizable and could therefore be tested and refined for other complex public health and wider social intervention programs.

Our questions are designed to be useful not only to researchers but also to policymakers, by helping identify the types of knowledge that might be generated from any possible evaluation given the strength of evidence available in response to each of five questions; and also to funders and commissioners of research, by helping support a more systematic consideration of resource allocation decisions depending on the types of knowledge required. Ideally, the questions should be used to stimulate and structure debate by all parties as part of the "iterative, bidirectional circuitry of scientific discovery" involving knowledge exchange and research translation (Ginexi and Hilton 2006; Ogilvie et al. 2009; Simmons et al. 2009).

References

Bond, L., P. Craig, M. Egan, K. Skivington, and H. Thomson. 2010. Health Improvement Programmes: Really Too Complex to Evaluate? *BMJ* 340:c1332.

Bradford Hill, A. 1965. The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine* 58:295–300.

Breckland Council. 2009. Report of the Director of Services and the Deputy Chief Executive to the General Purposes Committee—7th January 2009. Thetford—Healthy Town. Available at http://democracy.breckland.gov.uk (accessed July 14, 2010).

Craig, P., P. Dieppe, S. Macintyre, S. Michie, I. Nazareth, and M. Petticrew. 2008. *Developing and Evaluating Complex Interventions: New Guidance*. London: Medical Research Council.

Department of Health. 2008a. Healthy Community Challenge Fund: Invitation to Local Areas. Available at http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/DH_085328 (accessed July 13, 2010).

Department of Health. 2008b. *Healthy Weight, Healthy Lives: A Cross Government Strategy for England*. London: Department of Health.

Dudley Healthy Towns. 2010. Dudley Healthy Towns. Available at http://www.dudleyhealthytowns.co.uk (accessed July 14, 2010).

Exworthy, M. 2008. Policy to Tackle the Social Determinants of Health: Using Conceptual Models to Understand the Policy Process. *Health Policy & Planning* 23:318–27.

Freedman, B. 1987. Equipoise and the Ethics of Clinical Research. *New England Journal of Medicine* 317:141–45.

Ginexi, E., and T. Hilton. 2006. What's Next for Translation Research? *Evaluation & the Health Professions* 29:334–47.

Hawe, P., A. Shiell, and T. Riley. 2009. Theorizing Interventions as Events in Systems. *American Journal of Community Psychology* 43:267–76.

House of Commons Health Committee. 2009. *Health Inequalities. Third Report of Session 2008–09*. London: Stationery Office.

Last, J., ed. 2001. *A Dictionary of Epidemiology*. Oxford: Oxford University Press.

Leviton, L., L. Khan, D. Rog, N. Dawkins, and D. Cotton. 2010. Evaluability Assessment to Improve Public Health Policies, Programs and Practices. *Annual Review of Public Health* 31:213–33.

Lorenc, T., G. Brunton, S. Oliver, K. Oliver, and A. Oakley. 2008. Attitudes to Walking and Cycling among Children, Young People and Parents: A Systematic Review. *Journal of Epidemiology & Community Health* 62:852–57.

Mackenzie, M., C. O'Donnell, E. Halliday, S. Sridharan, and S. Platt. 2010. Do Health Improvement Programmes Fit with MRC Guidance on Evaluating Complex Interventions? *BMJ* 340:c185.

Medical Research Council. 2000. *A Framework for Development and Evaluation of RCTs for Complex Interventions to Improve Health*. London: Medical Research Council.

Moving Thetford Forward. 2010. Moving Thetford Forward. Available at http://www.movingthetfordforward.com (accessed July 14, 2010).

National Institute for Health and Clinical Excellence. 2008. *Promoting and Creating Built or Natural Environments That Encourage and Support Physical Activity*. London: National Institute for Health and Clinical Excellence.

National Obesity Observatory. 2009. *Standard Evaluation Framework for Weight Management Interventions*. Oxford.

Nutbeam, D. 1998. Evaluating Health Promotion—Progress, Problems and Solutions. *Health Promotion International* 13:27–44.

Ogilvie, D., P. Craig, S. Griffin, S. Macintyre, and N. Wareham. 2009. A Translational Framework for Public Health Research. *BMC Public Health* 9:116.

Pawson, R., and N. Tilley. 1997. *Realistic Evaluation*. London: Sage.

Petticrew, M., M. Whitehead, S. Macintyre, H. Graham, and M. Egan. 2004. Evidence for Public Health Policy on Inequalities: 1: The Reality According to Policymakers. *Journal of Epidemiology & Community Health* 58:811–16.

Sallis, J., M. Story, and D. Lou. 2009. Study Designs and Analytic Strategies for Environmental and Policy Research on Obesity, Physical Activity, and Diet: Recommendations from a Meeting of Experts. *American Journal of Preventive Medicine* 36(2 suppl.):S72–S77.

Shepperd, S., S. Lewin, S. Straus, M. Clarke, M. Eccles, R. Fitzpatrick, G. Wong, and A. Sheikh. 2009. Can We Systematically Review Studies That Evaluate Complex Interventions? *PLoS Medicine* 6: e1000086.

Simmons, R., D. Ogilvie, S. Griffin, and L. Sargeant. 2009. Applied Public Health Research—Falling through the Cracks? *BMC Public Health* 9:362.

Stamatakis, E., U. Ekelund, and N. Wareham. 2007. Temporal Trends in Physical Activity in England: The Health Survey for England 1991 to 2004. *Preventive Medicine* 45:416–23.

Thetford Healthy Town. 2010. Thetford Healthy Town. Available at http://www.thetfordhealthytown.co.uk (accessed July 14, 2010).

Thomson, H., R. Hoskins, M. Petticrew, D. Ogilvie, N. Craig, T. Quinn, and G. Lindsay. 2004. Evaluating the Health Effects of Social Interventions. *BMJ* 328:282–85.

Trevisan, M., and Y. Huang. 2003. Evaluability Assessment: A Primer. *Practical Assessment, Research & Evaluation* 8:20.

Wanless, D. 2004. Public Health Evidence. In *Securing Good Health for the Whole Population: Final Report*. London: HM Treasury.

Westley, H. 2007. Childhood Obesity: Thin Living. *BMJ* 335:1236–37.

White, M., J. Adams, and P. Heywood. 2009. How and Why Do Interventions That Increase Health Overall Widen Inequalities within Populations? In *Social Inequality and Public Health*, ed. S. Babones, 65–82. Bristol: Policy Press.

Wholey, J. 1979. *Evaluation: Promise and Performance*. Washington, DC: Urban Institute.

Wilson, P., and M. Petticrew, on behalf of the Medical Research Council's Population Health Sciences Research Network knowledge transfer project team. 2008. Why Promote the Findings of Single Research Studies? *BMJ* 336:722.

Windsor, R., N. Clark, N. Boyd, and R. Goodman. 2004. *Evaluation of Health Promotion, Health Education and Disease Prevention Programs*. Boston: McGraw-Hill.

Yin, R. 1979. *Changing Urban Bureaucracies: How New Practices Become Routinized*. Lexington, MA: DC Heath.

Acknowledgments: This is an independent report commissioned and funded by the Policy Research Programme in the Department of Health. The views expressed are not necessarily those of the department. The funder had no involvement in the study design, collection, analysis, and interpretation of data, writing of the article, or decision to submit the article for publication.

David Ogilvie and Andy Jones also were supported by the Centre for Diet and Activity Research (CEDAR), and Martin White was supported by Fuse, the Centre for Translational Research in Public Health. Both CEDAR and Fuse are UKCRC Public Health Research Centres of Excellence, with funding from the British Heart Foundation, Cancer Research UK, Economic and Social Research Council, Medical Research Council, National Institute for Health Research (NIHR), and Wellcome Trust, under the auspices of the UK Clinical Research Collaboration, which is gratefully acknowledged (see www.esrc.ac.uk/publichealthresearchcentres). Steven Cummins also was supported by an NIHR Senior Fellowship.