Generalizing about Public Health Interventions: A Mixed-Methods Approach to External Validity

Laura C. Leviton

The Robert Wood Johnson Foundation, Princeton, New Jersey 08543-2316; email: llevito@rwjf.org

Annu. Rev. Public Health 2017. 38:371–91

First published online as a Review in Advance on January 6, 2017

The *Annual Review of Public Health* is online at publhealth.annualreviews.org

https://doi.org/10.1146/annurev-publhealth-031816-044509

Keywords

external validity, generalization, applicability, sampling, context, complexity, dissemination, adoption, evidence-based interventions, core components, adaptation

Abstract

Public health researchers and practitioners are calling for greater focus on external validity, the ability to generalize findings of evidence-based interventions (EBIs) beyond the limited number of studies testing effectiveness. For public health, the goal is applicability: to translate, disseminate, and implement EBIs for an impact on population health. This article is a review of methods and how they might be combined to better assess external validity. The methods include (*a*) better description of EBIs and their contexts; (*b*) combining of statistical tools and logic to draw inferences about study samples; (*c*) sharper definition of the theory behind the intervention and core intervention components; and (*d*) more systematic consultation of practitioners. For population impact, studies should focus on context features that are likely to be both important (based on program theory) and frequently encountered by practitioners. Mixed-method programs of research will allow public health to expand causal generalizations.

INTRODUCTION

Programs, policies, and practices that have been tested for effectiveness are termed evidence-based interventions (EBIs). Public health professionals urgently seek to address the challenge of external validity: the extent to which researchers can generalize about the effects of EBIs across variation in populations, settings, treatments, measurements, and historical periods (11, 23, 76). The purposes of this article are to explore reasons that external validity receives so little attention and to describe methods to address the challenges. Two strong and widely used public health interventions are used throughout to illustrate these issues and some ways forward: HIV prevention and home visiting for vulnerable pregnant women and young families (2, 14, 57).

The Problem

The sense of urgency stems from practitioners' need to apply EBIs in real-world situations in order to improve the health of populations. Urgency also stems from policy makers' recognition that dissemination by itself is not sufficient for widespread and high-quality implementation of EBIs (40–42, 85) and from researchers who want to replicate a study or conduct systematic reviews of EBIs (37). Yet, little research attention is paid to external validity of EBIs.

Federal, state, and privately supported registries encourage practitioners to use EBIs, whether through endorsement, technical assistance, or funding (12, 14, 15, 20, 25, 73, 84, 87). To better achieve their aims, the registries need to match their focus on tested interventions with more focus on external validity. In principle, registries aim to give guidance so that practitioners do not waste their time and to ensure that public resources are spent wisely. Yet, a simplistic assumption underlying the registries is that an EBI found to be effective in a limited number of studies will probably be effective in other contexts. Practitioners are supposed to select EBIs that seem to fit their situation. However, the information to support these EBIs is not sufficient to assist practitioners. The populations, settings, and implementation issues presented in the real world of public health practice and policy are more various than the contexts in which the EBIs were originally studied (50); hence application of the EBIs in new settings has unknown external validity, and guidance is lacking about how to address these variations. For applying all but the very simplest EBIs, this situation poses a serious barrier.

Calls for Greater Attention to External Validity

Across many policy sectors, practitioners are asking for more information about external validity. In medicine, the details of EBIs and their contexts are often so poor that clinicians cannot possibly judge whether the results apply to particular kinds of patients (37, 72). Public health practitioners have a similar difficulty in generalizing EBIs to a variety of populations and communities (78). The systematic reviews that guide the designation of EBIs often do not attend to external validity (3, 66). In youth development, most reports of EBI evidence provide at least some description of study participants, but relatively few reports describe the study settings, how the EBIs were implemented, or the EBIs' cost (47). A systematic review of leading economics journals indicates that external validity considerations are simply missing from reports of effectiveness that might strongly influence policy (68). Year after year, the Community Preventive Services Task Force addresses external validity under the rubric of "Applicability" (20) and, year after year, reports the lack of information to Congress:

> Even when enough evidence exists for the Task Force to recommend an intervention, information may be missing that could help users decide if the intervention will work in their specific setting or meet

their unique needs. Among the 13 interventions the Task Force recommended in FY 2015, a number of evidence gaps repeatedly surfaced:

- Were the interventions effective in rural settings?
- Did effectiveness differ by race, ethnicity, age, disability, education, income, or insurance status?
- Which components of the interventions are essential for the interventions to be effective? What other components or combinations of components can increase effectiveness?
- Does effectiveness differ according to the way the intervention is delivered (e.g., the length of the program, face-to-face versus telephone, instructor-to-learner ratio)? (21, p. 2)

Overview of the Article

Barriers to knowledge about external validity include confusion over its formal definition, thorny problems of sampling, and an ongoing need to attend to the logic of causal generalization and inductive inferences. They also include poor descriptions of EBIs themselves and a lack of information about the underlying principles, program theory, and core components.

Bodies of evidence in public health can help mitigate these barriers: the living, growing knowledge bases about interventions and classes of interventions. In the twenty-first century, it is no longer necessary to rely on single studies to assess either internal or external validity. In fact, one should not do so, given the likely flaws in any single study. Bodies of evidence lend themselves to systematic reviews and identify important gaps in knowledge. Bodies of evidence give indications of how robust an effect may be when EBIs are applied to different populations, settings, systems, and time periods. Most important for external validity and public health practice, bodies of evidence include a wide range of study types—randomized experiments or quasi-experiments, but also implementation studies, surveys, and qualitative information—that provide essential information. For this reason, bodies of evidence are critically important for construct validity as well. They assist with better classification of context features, provide explanations for the roles of context features in affecting outcomes, and focus attention on those features that matter to effectiveness.

With focus, it is possible to better understand features of context that matter to external validity. To improve population-level impacts from EBIs in public health, the focus should be on context features that are likely to be

- important, based on prior information about the theory behind the intervention (program theory) as well as practice experiences that further specify core components, and
- frequently encountered in practice, such as barriers, opportunities, and adaptations in response to these features.

These are the areas in which expanding on generalizations will have the largest impact on public health (36, 45, 54). Setting priority on these features will require studies of effectiveness but also a range of other types of studies. It will incorporate practitioner experiences systematically in those studies, adding practice-based evidence to evidence-based practice (41, 42).

WHY IS ASSESSING EXTERNAL VALIDITY SUCH A CHALLENGE?

The Logic of Causal Inference and Causal Generalization

Establishing credible evidence of effectiveness depends on internal validity (11, 23, 76). Internal validity is defined as the extent to which a causal inference is warranted about an intervention's effects. Research designs including randomized trials and strong quasi-experiments systematically eliminate biases that pose alternative explanations for the results, the well-known threats to internal

validity (11, 23, 76). Although these research designs have been challenged many times, no credible standards for causal inference have replaced them (77). Research on effectiveness gives priority to internal validity in order to be certain about effectiveness in the first place. Yet, in individual studies, a focus on internal validity often limits external validity (38, 52, 56). As seen below, tests of effectiveness require limiting the variation in populations, settings, treatments, and measures, whereas assessment of external validity requires expanding such variation.

In contrast with the deductive logic of internal validity, external validity depends on inductive logic. Inductive logic is based on probabilities, not certainty (24).[1] One can be fairly certain about the necessary conditions for an EBI to produce an effect, but certainty about the sufficient conditions is not possible because the EBI cannot be tested under all possible conditions. Applied social research has difficulty contending with this problem, common responses being a shrug and "this is important, but there are few good answers."

The Need for Better Description of Interventions and Context

Journals rarely give enough space to describe the details of implementation and context, although some acknowledge the problem and are starting to provide space online for the details (42). Reporting guidelines increasingly ask for these details (10, 46; **http://www.squire-statement.org/**). With such details, individual practitioners can better exercise judgment about whether their own context is sufficiently similar to those in which the EBI was tested. Unfortunately, there is little agreement beyond a need for better description. A recent literature review identified 25 frameworks to assess external validity, applicability, or transferability in health research. None of them covered all the reviewers' criteria for assessment, and none of them demonstrated perceived utility for the reader (10).

Underdeveloped Program Theory and Construct Validity

Better descriptive frameworks to assess generalizability would be helpful, but they are not enough. A general failing of descriptive frameworks is their focus on surface similarity: attending to features that are most readily available on the surface. These may not be the features that make an EBI effective; indeed, some notable failures to replicate based on simple descriptions support this contention (28). Program theory specifies the core components of an EBI (also known as essential program elements as distinct from peripheral or incidental program elements) (44). Core components are necessary to achieve intermediate program objectives. Core components may be prosaic (e.g., making families aware of their eligibility to participate in home-visiting programs), but they can also bring about the moderators of intervention effects (e.g., staff development and monitoring to assure high-quality implementation) or the mediators of an effect (e.g., home visiting addresses the individual needs of those served, which is hypothesized to improve parenting and child outcomes) (57). Unfortunately, program theory appears to be badly underdeveloped in many outcome studies of prevention interventions for physical and mental health, as seen in two systematic reviews (27, 30), as well as a scan of EBIs for violence prevention (67). Criteria are lacking to define the core components (8), and even when core components are defined, the actual strength and integrity of the intervention-as-implemented are usually not reported.

The underlying challenge involves construct validity: the ability to make inferences from specific activities back to the theoretical concepts on which they were based (11, 23, 26, 76). Although

[1] Of course, statistical estimation is always a matter of probabilities.

construct validity is often discussed as a measurement issue, in fact it applies just as much to treatments: assessing whether theory has been made operational. As in measurement, the challenge is to infer from specifics back to underlying concepts. Knowledge about construct validity is essential to expanding knowledge about external validity (11, 23, 26, 76). Assessing surface similarity is important, but attention to the underlying theoretical constructs helps to extend causal generalizations with more confidence (see below).[2]

Because program theory is underdeveloped for many EBIs, practitioners often state that they cannot do the EBI by the book. Manuals of operations often do not have the flexibility to offer guidance for common barriers. When practitioners then have to adapt EBIs in order to respond to the situation, it is unclear whether the adaptation is still consistent with the core components or whether the adapted EBI has enough strength to bring about the desired effect. Consistency and adequate strength of core components are termed fidelity to the EBI. Assessing fidelity is a key area of concern not only for external validity, but also for the construct validity of the EBIs themselves.

The Fallacy of Generalizing from Limited Samples

Guidance from EBI registries assumes that EBIs that were tested in a small number of contexts would be effective in a wider variety of contexts. This assumption is simply wrong on empirical grounds—for example, Hawe (44) describes failures to replicate several EBIs, not to mention the inconsistent or even contradictory effect sizes that appear in systematic reviews. The assumption is also tenuous on logical grounds because most studies of EBIs use convenience or purposive samples, not representative ones (80). Such samples create interactions of selection and treatment at several levels, and these problems present threats to external validity.[3] Participants' propensity to benefit from an EBI is unknown, so it is not possible to estimate what the effect would be on an unselected population. Also, the organizations and practitioners in the original studies often have greater capacity to implement the EBIs than do unselected organizations and practitioners, an interaction of setting and treatment that limits generalization. In all fairness, many effectiveness studies in public health recruit very hard-to-engage or needy populations, and they implement an EBI in lower-capacity organizations. Yet, practitioners still do not know how study participants differ from the general populations that might receive the EBI or the contexts in which it is implemented (39). Public health can improve the registries' usefulness by expanding a focus on external validity. This article now turns toward methods to do so.

PERSPECTIVES ON EXTERNAL VALIDITY FROM SOCIAL SCIENCE AND STATISTICS

External Validity as Uncertainty Reduction

Deciding to endorse, pay for, or use an EBI are examples of decision-making under conditions of uncertainty and risk (55). The risk for policy makers and local practitioners alike is that scarce

[2]Pawson & Tilley's (65) realistic evaluation presents some concepts that overlap with what follows; however, Campbell & Stanley (11), Cook & Campbell (23), and Cronbach & Shapiro (26) predate their exposition and expand it in useful directions. Pawson & Tilley, along with Chen (17), Weiss (88), and many others (77) have contributed by consulting stakeholders to build program theory, a major advantage for external validity (see below).

[3]Formal definitions of well-known threats to external validity are not repeated here but include interactions of the treatment with selection, treatment with setting, treatment with instrumentation, other reactive features of experimental arrangements, additive effects of multiple treatments, interaction of the causal relationship with treatment variations, and context-dependent mediation of the causal relationship (11, 23, 76).

public health resources might be wasted. Cronbach & Shapiro (26) reframed external validity as information to reduce uncertainty. Doing so frees external validity from some restrictions of inductive logic because it involves assessing the likelihood of effectiveness using a range of information, not just accumulation of definitive tests in a limited number of instances. For example, both Greenhalgh et al. (43) and Hawe (44) review features of EBIs and their context that predictably affect the uptake and implementation of an EBI.[4]

Two Different Purposes

As Cook (22) and Cronbach & Shapiro (26) point out, there are two purposes to assessing external validity: (*a*) representation of a class of interventions and contexts ("Is this EBI effective for population X in context Y?") and (*b*) extrapolation from existing studies of the EBI to new, unstudied contexts ("Will this EBI work in new populations and contexts? Will this EBI work in my particular context with my population?"). Some social scientists simply have a scholarly interest in representation: Does a causal relationship hold across a class of interventions, populations, settings and so on? For public health, in contrast, the information is applied to health planning, policy and assurance, so both representation and extrapolation are important.

Unpacking the Undifferentiated Bundles of Context and Complex Interventions

Sometimes researchers caution against overgeneralization by saying that context is unknown or that interventions are too complex and therefore unpredictable. However, these cautions are not helpful when they obscure opportunities to assess external validity. As Patton (64) notes, stepping back to see many examples allows one to detect the patterns in complex or complicated situations, allowing for theory building and the ability to classify, compare, and contrast. External validity requires assessment of five sets of characteristics that vary across individual studies. Within a single study, they are inextricably confounded (22, 26, 76).

The populations studied. These are program participants or other units (e.g., families, communities) affected by policies, programs, and practices. For example, home-visiting programs aim to provide support for vulnerable pregnant women and young families. The demographic and background characteristics of participants vary substantially across studies of home visiting EBIs, making it difficult to disentangle the effects as being a function of participant characteristics, EBI characteristics, or both (57). Would an EBI found to be effective for adolescent mothers also be effective for older women who have had several children already? Beyond surface variables such as demographics or location, what other features of these populations matter, for example, their levels of trust in health care? These are less often measured but are sometimes more powerful to mediate or moderate behavior changes.

The settings studied. Settings are a set of nested context variables, including practitioners' knowledge, skills and abilities, training and supervision, the capacity of the organizations in which they are embedded, the communities where services are provided, and the systems that support or impede implementation. At the level of individual studies, these variables interact. Explaining results in one case may be quite different from those in another case. For example, the effects of

[4]The focus on reduced uncertainty justifies statements of probability in this article, such as "likely" and "might" instead of "is" and "does."

the Nurse Family Partnership EBI were not replicated in Britain, in spite of reasonable fidelity to the model (69). The failure to replicate may have been due to Britain's systems of support for new parents, which are much more comprehensive and readily accessible than they are in the United States. These services may have produced a ceiling effect, such that no marginal benefit of home visiting could be observed.

The particular treatments implemented. EBIs implemented in open systems are rarely exact replications of a study intervention—they cannot be (26). Also, as Green (39) points out, EBIs in public health are often more akin to processes than products: Think about the differences between vaccines, which must be standardized to ensure the public's health, and efforts to immunize all children, which are complex processes (or at least complicated ones). Batalden et al. (5) cite a long tradition of research in business and public administration to support their distinction between health care as a uniform product (e.g., correctly taking and reading an X-ray) versus services such as primary care, which are coproduced with the patient. Services and processes require that practitioners modify their actions in response to the context. In the EBI literature, such modification is termed adaptation.

Adaptation is a normal feature of the adoption of innovations (70), and it occurs across many policy sectors that use EBIs (54). Adaptation in good faith is not the same as well-known implementation failures caused by carelessness or dishonesty (54). Adaptation is a cause for concern when it is framed as a departure from fidelity to the EBI, because some adaptations have been known to reduce program effect sizes (32, 48). Yet in other cases adaptations actually improve outcomes (6, 30, 79). Surely, these mixed findings require more systematic study (7, 71). Adaptation is not necessarily a departure from fidelity; in fact, fidelity and adaptation can exist side by side in some EBIs (4, 31). A central question becomes, what is consistent with the EBI and what is not?

The specification of measurement variables and study design. These features in the sample of studies concern the particular ways that effects are measured, for example, the length of follow-up to assess sustained effects. Investigators often make different decisions about methods and measure things differently from one study to the next. In the case of home visiting, not all EBIs target or measure all 8 outcomes; for example, 14 of the EBIs improve positive parenting practices, whereas 11 improve child readiness for school (2).

The historical period in which the studies were conducted. Studies are conditioned by the times in which they take place. Home visiting in 2016 is markedly different from the same practice in 1986, when the first randomized experiment on nurse home visiting was published (63). Resources for maternal and child health have diminished and increased again, systems and agency responsibilities have changed beyond all recognition, and implementation has been refined over time (2). The historical period matters to HIV interventions for different reasons (see below).

External Validity as a Sampling Challenge

External validity is often defined as a sampling problem: Ideally, one could include random samples in studies, capture the most relevant variation, and arrive at better conclusions about an average population effect (81). But this ideal research method is difficult and seldom achieved.[5]

[5]Big data may eventually carry public health beyond the sampling conundrum because the analytics themselves may provide detailed information about external validity. However, public health may be a long way from understanding the variation in systems, social, environmental, and behavioral factors that would permit such detailed knowledge.

There are four related reasons why improved sampling is necessary but not sufficient to assess external validity. First, an average effect size across samples of studies, populations, or settings does not much help individual practitioners who want to know whether the results of an EBI can generalize to their specific, local context (22, 26). For that matter, an average effect is not helpful enough to policy makers who need to know who will benefit from an EBI under which circumstances.

The second challenge is the confounding of populations, settings, treatments, measurements, and historical periods at the level of the study. Cook (22) points out that it is a fallacy to attribute study outcomes only to the treatment, when in fact it is the interaction among these five characteristics that causes the outcomes. It is not possible to test all the interactions of these variables (76). Also, given the resources even of the largest studies of effectiveness, subgroup analysis is severely constrained by sample sizes and the degrees of freedom required to assess covariates.

The third challenge is that, in many cases, the most important sources of variation in these domains are simply not known, making it difficult to sample on anything beyond demographics, location, or general types of practitioners and settings (81). In mature bodies of evidence about an EBI, there may be a solution because theory together with practice go beyond these surface characteristics (see HIV example and discussion of the Pareto principle, below).

The fourth challenge is the prevalence of purposive or convenience samples for even the best randomized experiments on EBIs. When a study randomly samples units and includes them all in an intent-to-treat framework, then a conclusion about effectiveness can be generalized to the inference population, that is, those who would be covered by a health policy or program. Yet, it is estimated that only 3% of social experiments use the twofold strategy of random sampling then random assignment (80). Sometimes researchers try to generalize retrospectively from purposive samples by statistically adjusting for participants' propensity to benefit; however, Tipton (80) demonstrates the biases that such adjustments can introduce. Tipton et al. (82) developed a method of prospectively selecting participants for randomized experiments using measured propensity to benefit from intervention, so the study sample is similar to the population of inference. This method potentially offers advantages, but its requirements are not easy to meet (83). Also, individual units are nested within settings, systems, and other context features, and small samples of these larger units often prevent causal generalization from going much beyond the individual participant level (81). Propensity to benefit needs to be better understood in many cases in order to select participants that reflect the population—it may not be sufficient to rely on surface similarity (e.g., demographics and location). And finally, while this method may assist with external validity to represent a set of contexts, it is not helpful to extrapolate to new contexts (22).

The Importance of Maximizing Heterogeneity Across the Five Study Characteristics

Without systematic sampling on the most important sources of variation, external validity requires that investigators maximize the heterogeneity of the populations, settings, treatments, measurements, and historical periods studied. That way, researchers can determine how robust the outcomes are across studies and also assess whether differences across studies are relevant to the outcomes. However, effectiveness studies tend to reduce heterogeneity in settings, types of practitioners, populations, and the very definition of the topic at hand (22). Moreover, meta-analysis often restricts its focus to randomized experiments only, thus eliminating studies that employ quasi-experiments or simple observation, studies which often provide more real-world contexts than do randomized experiments (22).

TOWARD SOLUTIONS: UNCERTAINTY REDUCTION

Logic to Reduce Uncertainty About Context

Shadish et al. (76) provide five principles to extend causal generalizations where variations in populations, settings, and treatments have not been studied directly. Leviton & Trujillo (54) demonstrate how more systematic interaction with practitioners can help to apply these principles—a structured form of participatory research.

Principle 1: Assessing surface similarity between what is studied and the target of generalization. Surface similarity might include what is specified in a manual of operations or described in studies of the EBI: for example, a target number of visits, qualifications of practitioners, and characteristics of a target population. If practitioners replicate these features, there is less uncertainty about effectiveness. Yet, some surface features are just trivial: It is probably irrelevant that the home visitor wore blue pants instead of gray pants. Other surface features might, in some cases, be false positives for what is needed to bring about the effect; or features omitted from reports may be false negatives. Moreover, so much is invisible with surface similarity alone—the underlying program theory may provide much more insight (see below). In addition, public health would do well to extract knowledge more systematically from practitioners about their experience with EBIs. Doing so would help to "surface" the similarity that practitioners see, but researchers do *not* see (54).

Principle 2: Ruling out irrelevancies, context attributes that do not change a generalization. If an EBI is effective across variations in populations, settings, treatments, measurements, and historical periods, then those variations are likely irrelevant for causal generalization. For example, home-visiting interventions are apparently effective for families in the various American tribal communities, so in this case culture and setting may be irrelevant: They do not limit generalization. However, the EBIs may require some cultural adaptation (29); it is an empirical question how much adaptation is necessary.

Principle 3: Identifying context attributes that limit generalization. Effectiveness does not generalize to EBI implementation failures: no program, no effect (54). Also, characteristics of culture, systems, or other setting variables may limit generalization, and practitioners can assist in identifying some of these on the basis of their experiences. Some context attributes may not impair effectiveness as such but may work to moderate the strength of implementation and the obtained effects. For example, challenges for home visiting on tribal reservations include the need to travel long distances in remote areas, missed appointments due to competing family priorities, and substantial participant attrition (29).

Principle 4: Interpolating to unsampled values within a sample range and extrapolating beyond the sample range. To get beyond the problems of selection-treatment interaction, one can investigate whether an EBI is effective for participant samples at the extremes of some important characteristic, as well as modal instances of that characteristic. Thus, samples of participants may be easy or difficult to engage, have greater versus lesser disadvantage, or (if it is known) have a greater, versus lesser, propensity to benefit from intervention. If the EBI is effective for participants at different points in such sampling ranges, it is likely effective within the range and perhaps beyond it. In the same way, implementing and testing the EBI in settings with both high and low capacity lends some confidence to generalize about settings with intermediate capacity. Extrapolation must often be more tentative than interpolation; for example, it may still be essential to determine

the minimum capacity required for a setting to implement an EBI, or a point at which participants experience diminishing benefit from additional intervention (86).

Principle 5: Causal explanation, in which scientists develop and test explanatory theories about the target of generalization. Well-developed program theory improves causal generalizations (22, 26, 28) as well as overall effectiveness of public health interventions (35, 42). One reason is that it provides a basis for understanding variation in implementation: what is consistent with the EBI, what weakens it, and what constitutes an appropriate practitioner adaptation to population and setting characteristics. Program theory can also identify mediators that describe the mechanisms linking an EBI to effects, as well as context features that moderate the effects (see Reference 54 for elaborated examples).

The Strong Potential of Bayesian Analysis to Assess External Validity

Bayesian analysis uses prior information to reduce uncertainty about statistical estimates (51). It permits statements such as "there is an X percent likelihood that the EBI increases the health of children of Y kind, in Z context." This method is respected among statisticians (e.g., see **http://community.amstat.org/sbss/home**), and effectiveness researchers show a growing interest—for example, a search of the Cochrane Library reveals 50 methods articles on Bayesian analysis (see **http://onlinelibrary.wiley.com/o/cochrane/cochrane_clcmr_articles_fs.html**). It can be combined with inferential statistics in advantageous ways (9). Yet, Bayesian analysis seems uncommon in published reports of EBIs in public health.

Bayesian analysis has several advantages. One often hears that external validity can be informed by examining the effects of an EBI for subsamples within effectiveness studies. To illustrate, a study of home visiting might include families in tribal communities, and the question is whether the effects generalize to other such families. Yet, in many cases these subsamples are too small for analysis using inferential statistics, offering only tantalizing hints (26, 76). Bayesian analysis addresses this problem by "borrowing strength" across sites and other studies and using other information to reduce variance and increase power (9). Bayesian analysis can use the overall effect in a trial to increase the precision of the estimate for the subsample. This approach will be feasible if the estimate for the subsample is reasonably similar to the overall effect. Thus, if tribal families have child health outcomes that are close to those for the average of home visiting recipients in a given study or are close to the average for other similar tribal families across studies, then the Bayesian approach can increase the precision of the estimate. If outcomes for tribal families were to be markedly dissimilar to those for the overall average of participating families, then that would also be valuable to know, precisely because the results appeared not to generalize to such families. Such a pattern might be the basis for additional study of tribal families, with high information value for policy and programs.

Bayesian analysis can also use prior information about barriers to implementation, whether reported by practitioners, reported in literature reviews (43), or quantified through correlation and regression in studies of the EBI. A short list will likely include lack of time, resources, and training, cultural differences, and unsupportive organizations (54). It makes sense to utilize practitioners' knowledge about contexts that may facilitate or impair implementation and thus effectiveness. It is simply efficient to do so and then test effectiveness in those particular contexts. Researchers frequently bemoan the small number of sites involved in most trials of EBI effectiveness, which makes it difficult to generate insights about context features that affect implementation and outcomes as well as generalization (81). Detailed information about site characteristics and capacities

might inform estimates of the prior probability of implementation and other context features that limit generalization.

The potential of Bayesian analysis continues to unfold. Recent clinical trials and simulations using Bayesian adaptive design for certain types of short-term outcomes achieved higher levels of certainty about effects with substantially smaller samples than did conventional trials (33). The implications for external validity should be obvious: If trials can be conducted with many fewer cases, and thus fewer resources, then more trials may be conducted or more variations within a study could be assessed.

Combining Bayesian Analysis with Other Inductive Strategies

In spite of these advantages, Bayesian analysis is still prone to the same logical errors and threats to validity that need attention when using inferential statistics. Biased studies, no matter how large and detailed, still provide biased estimates. Practitioners have been known to share an unfounded belief in the efficacy of an intervention, or they may share an unfounded belief about modifications that will increase the strength and integrity of interventions, or about the settings and personnel that have the capacity to implement an intervention. Diverse sources of information that do not share the same biases are still crucial for both causal inference and causal generalization.

Keeping this caution in mind, Bayesian analysis is particularly advantageous when there is a body of diverse information about the EBI. For example, this approach can utilize information from other studies of the EBI to increase precision about subsamples. Thus, if tribal families participated in several previous studies of home visiting, then their outcomes can improve the precision of the estimate in a current study or a systematic review. There might still be a selection bias or other threat to external validity. Nevertheless, thanks to the five principles of causal generalization, more can be done to rule out such alternative explanations and to reduce uncertainty (76):

- To check for selection biases, one can determine how families were recruited for the studies. If they were not all recruited in the same way, but effect sizes are consistently positive across studies, then the method of recruitment is likely irrelevant, increasing the probability that home visiting benefits tribal families in general.
- The methods of selection may represent a range or continuum (based on need or the distance between the family and a home-visiting agency); if effect sizes are consistent and positive, then interpolation and perhaps extrapolation become feasible.
- The possibility that selection interacted with the intervention requires an explanatory theory: What other mechanisms might interact with the EBI, causing families who participate in the trial to have healthier children than unselected tribal families who also receive home visiting? Ruling out such interactions may be difficult but makes it increasingly plausible that results generalize to other tribal families and children. For example, what if the families participating in the trials of home visiting are more trusting of health care than nonparticipants in the trials? Does trust affect their ability to benefit from home visiting? This empirical question can be answered via survey, from literature reviews, key informants, or other sources.

TOWARD SOLUTIONS: UNPACKING TREATMENT VARIATION TO FURTHER SPECIFY THE CORE COMPONENTS

Construct Validity: Many Roads to Rome

A critically important point is that well-developed program theory often permits a variety of potential activities, or "roads to Rome," to implement the EBI's core components and achieve its

intermediate objectives (44, 54). So long as the intermediate objectives are achieved, the specific activities can vary. This approach deliberately invites a degree of adaptation, always evaluating the implementation for strength and consistency with the underlying core components. For example, practitioners in HIV prevention programs discovered early that adaptation was necessary, given the range of situations that pose a risk of infection (53). Today, adaptation packages assist the practitioner to modify EBIs thoughtfully (13), an approach that is now being replicated for violence prevention interventions (34, 59).

Figure 1 illustrates the use of underlying program theory, tailored to local context. It describes a class of behavioral interventions for HIV/AIDS prevention, borrowed from Leviton & Trujillo (54). It is deliberately presented not as a detailed logic model, but as an abstract sketch or flow chart. Thus, prevention planning requires local intelligence to assure maximum reach and relevance. Engagement depends on increasing people's motivation to protect themselves and then developing and reinforcing their skills to do so. Where there are frequent barriers to self-protection, such as partner abuse, addictions, and HIV-positive serostatus, these populations need tailored interventions to assist with coping. In some programs, follow-up support may be required to ensure that the intervention has sufficient strength and that behavior changes are sustained.

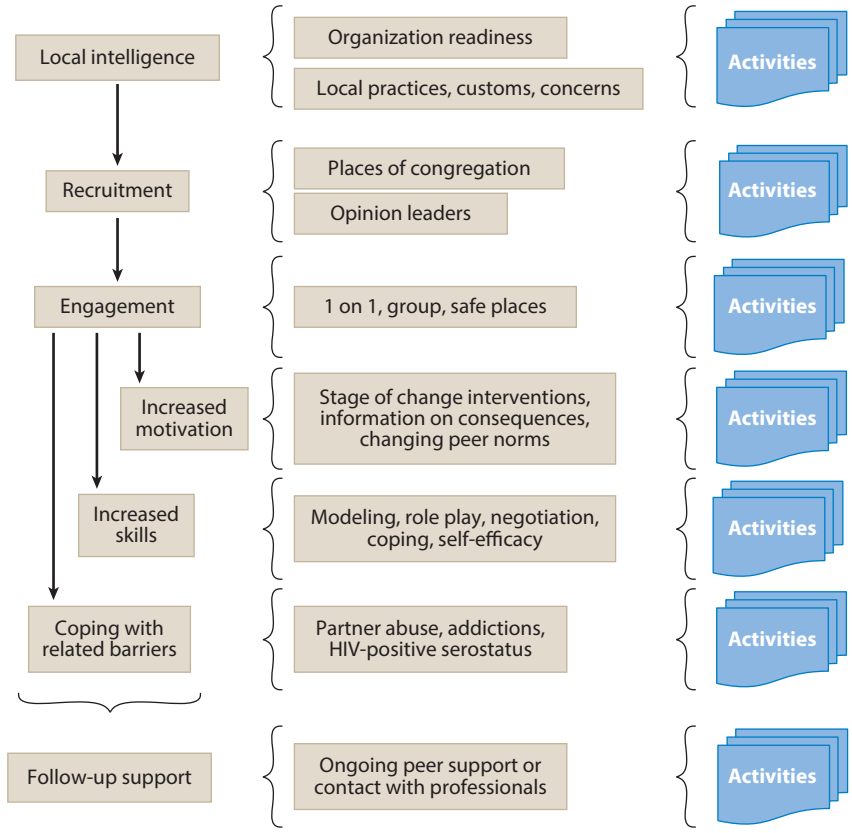

Figure 1

Greatly oversimplified flow chart of program components for HIV prevention. Source: Leviton & Trujillo (54). Theoretical constructs appear on the left, commonly used core components in the middle, and a variety of activities to implement the core components on the right.

Together, the increased motivation and skills for self-protection mediate behavior that avoids infection. Moderators include the characteristics of outreach workers and educators, the relevance of engagement tactics, and the strength of the intervention during engagement and follow-up. Some might object, arguing that this program theory sketch is too linear and ignores reciprocal influences between the intervention and context. Yet, the theoretical concepts and core components do have a temporal sequence: One does, in fact, lead to the next, and achieving the first intermediate objective is necessary for the rest to be achieved. Where reciprocal influence does occur is in the specific activities, which adapt in response to changes in the local context and the needs of populations.

Theory and Umbrella Concepts Offer a Guide to Adaptation

Equating core components with the specific activities of a manual of operations treats public health interventions as products to be standardized, not as services to be employed in context. In contrast, thinking of core components as umbrella concepts permits reasonable adaptation to context and generates a variety of activities to achieve intermediate objectives. For example, a health education curriculum might need to be translated into Spanish. Yet, this is an adaptation. How will one know that it is consistent with the original EBI—because the developer says so? More than surface similarity is involved here. Word-for-word translation is no guarantee that the cultural meaning of a curriculum will be conveyed. That requires going below surface similarity, and the core component is transformed in the process. Translation, of course, achieves better reach, an important aim of EBIs. Would one avoid adaptation under such circumstances? Beyond translation, whether cultural adaptations increase effectiveness or are largely irrelevant for a given EBI is an empirical question and a matter of current, active inquiry (54).

Adaptation Is an Opportunity to Further Specify Program Theory and Core Components

Evidence increasingly suggests that a thoughtful adaptation of activities can work to better define program theory and core components. For example, Collins (19) developed the multiphase optimization strategy (MOST), which identifies what the EBI program theory actually requires to achieve intermediate objectives, and then conducts successive tests of whether specific activities or components are necessary to achieve outcomes. This process informs cost-effectiveness, reach, and sustainability. Another effective approach in certain mental health interventions is to train practitioners on the underlying principles until mastery is observed and then to supervise, ensuring that adaptation in response to clients is still consistent with EBIs (54). In line with participatory evaluation, some strategies have been proposed for more systematic interaction between researcher/developers and practitioners to improve implementation and adaptation of EBIs (16). Green & Glasgow (40) suggest a process of consultation with practitioners, termed "mapping, matching, pooling and patching," (p. 20) that reflects derivation from theory to practice and back again. Expert practitioners have a larger repertoire of responses to practice contexts than do novices, so they can fit theory to context in various ways and thus further specify the program theory and core components (54).

In general, providing implementation supports for practitioners can cast light on appropriate adaptation (8). Such implementation supports can improve outcomes for EBIs: For example, a randomized experiment using the Getting to Outcomes© process (85) has been found to improve outcomes of an EBI for teen pregnancy prevention (18). The field of quality improvement in medicine encourages practitioners to use rapid cycle improvements to achieve an aim; this process

is inherently one of adaptation, and a systematic review indicates positive but limited evidence of effectiveness (75).

None of these approaches suggests that EBIs can be adapted willy-nilly. A careful process is needed to determine what is consistent with the core components, what is inconsistent, and what offers sufficient strength to achieve outcomes. Adaptations can be helpful, harmful, or irrelevant, and if irrelevant, they can waste time and resources that are better used by adhering to the EBI. The point is that for most EBIs, not enough is known about such adaptations, either for construct validity or external validity.

TOWARD SOLUTIONS: FOCUSING THE STUDY OF HETEROGENEOUS CONTEXTS

The Robert Wood Johnson Foundation (RWJF) is exploring a strategy to better inform external validity, in partnership with selected grantees and the Centers for Disease Control and Prevention (CDC) Division of Violence Prevention (34, 54). The specific focus is on the assessment of adaptations, but because adaptations are a response to context, the process extends to other context features. The idea is first to collect a high volume of practice experiences and then to focus attention on important and frequent variation in those experiences. It is not necessary to independently test the effectiveness of every adaptation in every context. Instead, researchers and practitioners together can capture the features of context that are likely to be frequently encountered and important (on the basis of program theory). These are the features that have high information value for further tests of effectiveness. While this strategy cannot provide a complete answer to external validity, it reduces uncertainty and expands generalizations in the areas of greatest consequence to population health (45, 54).

Phase 1: Crowd-Sourcing Practitioner Experiences

One failing of the EBI registries is that there is no formal, systematic way for practitioners to provide feedback about how an EBI worked in their contexts. This is a major limitation for external validity, which requires understanding many heterogeneous contexts. Instead, feedback on EBIs can provide a high volume of practice experiences to obtain the necessary heterogeneity. Individually, practitioner experiences with an EBI are limited. But taken together, practitioners have seen a broad and heterogeneous array of contexts—more so than the researchers or developers of a given EBI (54). Individual participatory studies of implementation cannot provide this variation because any single study is limited in scope. Instead, practitioner experiences can be systematically sampled. For example, under a grant from the RWJF, Naylor and her colleagues (60) have conducted a snowball survey to describe and quantify practitioner adaptations of the transitional care model (M.D. Naylor, personal communication, November 2015); likewise, in previous studies, they identified the most frequent barriers to implementing the model. Glasgow and colleagues (R.E. Glasgow, T.L. Hall, J.S. Holtrop, & L.M. Dickinson, personal communication, October 2016) used RWJF support to describe and quantify the frequency of 49 adaptations of the patient-centered medical home model.

Phase 2a: Winnowing Context Features for Further Study

Out of the heterogeneous, high volume of cases, it will be important to select the context features that should be studied further. Following the lead of the Community Preventive Services Task Force, investigators can use collective judgment, literature review, and quantitative studies to identify context features that are likely to be important on the basis of program theory (21). With

RWJF support, the CDC Division of Violence Prevention moved quickly to this phase because three EBIs had important context features that practitioners had already encountered frequently (34).

It will be helpful to winnow out the less important, less frequent context features. The process can start by eliminating the infrequent features on the basis of practitioner and client surveys or consultation with seasoned technical assistance providers. Other context features can be ruled out for further investigation because they are irrelevant to the EBI based on program theory, collective judgment, systematic reviews of existing studies, or subgroup analysis within a single study. Some adaptations do not generalize at all; they are important to identify precisely because they limit generalization. Such adaptations have insufficient strength, are inconsistent with the core components, or are actively harmful. Some features of adaptation are irrelevant additions. They will appear benign, but they can still introduce cost or take up time.

Phase 2b: Setting Priorities for Study

The Pareto principle justifies a focus on the important context features that also occur most frequently. It states that the large majority of effects come from a relative minority of causes. It fits an impressive array of economic, social, and business applications (61). **Figure 2** is an example from public health quality-improvement training (1). Examples such as these often come from individual settings, but why not apply them across populations and settings to assist causal generalizations? For instance, one could assess the frequency with which practitioners list barriers to implementation of an EBI. Their frequency could be empirically verified, along with generic adaptations to overcome the barriers. A lack of time and resources to implement EBIs might account for a large majority of cases, followed by an unsupportive organization and, in some cases, cultural barriers (see also Reference 43 for a longer list). These may be the most frequent and important context features that limit generalization. It is not possible to study all the context features in complex interventions. Yet, if the Pareto principle applies, then the important, frequent context features are hardly mysterious or particularly complex. Moreover, addressing the important, frequent features works to reduce variation and improve services. The field can then turn its attention to the next most frequent set of important features.

One might object that context has different forces working at different levels of the social ecological model, and the Pareto principle might identify frequent and important forces at each

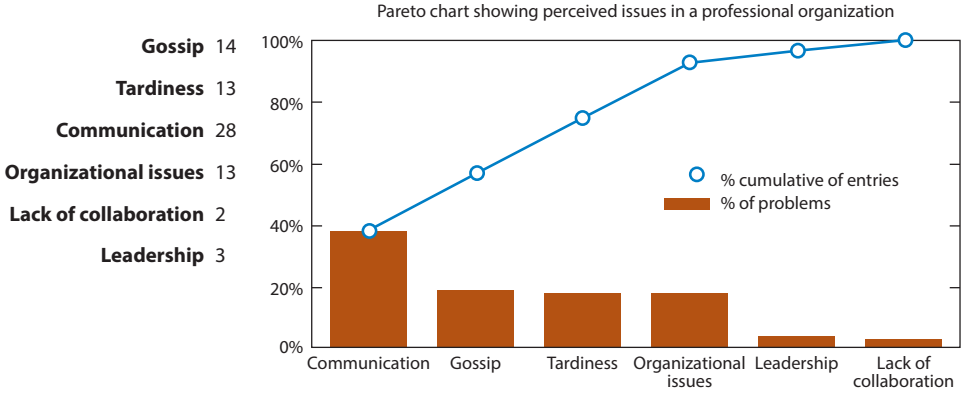

Figure 2

Distribution of causes and effects. Source: Arizona Public Health Training Center (1).

level. Thus 20% of implementation forces may be the most frequent and important, but nested within 20% of systems or policy barriers that are most frequent and important, and so on for cultural barriers and other population forces. Nevertheless, the number of choices is finite, and the most frequent combinations can still receive priority for investigation.

Related to the Pareto principle is the operations management concept of runners, repeaters, and strangers, which categorizes products and services according to their volume and variety (62, 89). Runners appear frequently, repeaters appear regularly but at longer time intervals, and strangers appear at irregular and sometimes unpredictable intervals. Different processes are often needed for these three groups. Over time, strangers can become repeaters or even runners through routinization and recognition of patterns. In the case of an EBI, runners would be the most frequently encountered context features, repeaters would be somewhat less common, and strangers would be relatively uncommon. This framework may be helpful to assess external validity because, using inductive logic and the tools described in this article, public health may be able to generalize about runners at both individual and setting levels. Repeaters may require more of a conceptual approach, as well as more adaptation of the EBI. Strangers may require the most thought, study, and discussion about whether an EBI can generalize and, if so, what kinds of adaptation are consistent with the underlying theory and core components. Strangers are valuable to contribute to both external validity and construct validity, but it will be premature to focus too much on them until the EBI is sufficiently mature and the theory behind it is sharp enough to accommodate first the runners and then the repeaters.

HIV prevention illustrates how a maturing body of evidence can employ this framework (53). For many years, a reasonably capable organization could intervene with at-risk gay men—not everyone would avoid infection, of course, but intervention prevented many cases. Epidemiology began to identify repeaters: people with a variety of special situations for which tailored interventions were needed, such as runaway youth. Because these tailored interventions are now EBIs in the CDC Compendium (14), the repeaters have become runners. However, an important stranger then emerged because the availability of antiretroviral therapy meant that AIDS was no longer a death sentence. Some younger gay men now believed they no longer had to protect themselves. It was, and is, essential to come to grips with this development in HIV prevention.

Phase 3: Empirical Studies to Reduce Uncertainty

After winnowing unimportant and infrequent context features, some features will be plausible enough, or controversial enough, to proceed to further empirical study. In the interest of speed, these can focus on whether context features affect the moderators and mediators of outcomes. For example, a cultural adaptation may recruit more members of an ethnic group or better mediate the desired behavior changes (49). Note that Phase 3 does not require a randomized experiment, although Collins (19) has shown that one is beneficial. This phase can also further specify the core components, if necessary. Qualitative comparative analysis (QCA) (74) can use the underlying program theory to further explore the necessary and sufficient conditions to place diverse activities under the program theory umbrella. The variety of quantitative methods to assess construct validity can also further specify program theory—for example, the multitrait-multimethod matrix might be used to identify adaptations that are consistent with the construct or are irrelevant (22, 76).

Phase 4: New Tests of Effectiveness

The first three phases are likely to identify important gaps in knowledge about EBIs, so new studies of outcomes are desirable as a check on the inductive process. In particular, it may be desirable to

test the effectiveness of frequent and important adaptations. Also, some context features may not be addressed satisfactorily by the first three phases and may require their own independent tests of effectiveness. These are likely to be cases where program theory needs further development, where not enough is known concerning combinations of populations and settings, or where gray areas of practice are simply not addressed by the available literature.

A process such as this could meet the criterion of usefulness that current frameworks for external validity cannot satisfy (10). The reason is that utility is built in through ongoing consultation with the users. Thus insights about appropriate adaptation might be garnered, stored, and shared. At the same time, as external validity is expanded, policy makers and researchers will gain a more nuanced understanding about what works for whom and in which circumstances.

CONCLUSIONS: TOWARD A PROGRAM OF RESEARCH ON EXTERNAL VALIDITY

Public health should develop a program of research on external validity that focuses on frequent and important context features, because these are the features that have the most consequences for population health. Such a program should draw on the entire body of evidence, including descriptive and qualitative information, practitioner experience, correlational and regression studies, and tests of effectiveness. Doing so has several advantages. (*a*) A systematic approach explicates the undifferentiated bundle otherwise lumped as "context." (*b*) One can go beyond surface similarity to more powerful generalizations on the basis of program theory. (*c*) One can utilize practitioner knowledge in meaningful ways. (*d*) With the accrual of information, one can be increasingly specific about situations where a causal inference is likely to hold. (*e*) One can also be more selective about where to focus additional attention and resources, whether the focus is on causal inference or causal generalization. Peer-reviewed journals may also be more welcoming of the papers that inquire into causal generalization and that provide richer description of interventions, if they were to have a better sense of the direction that such a program of inquiry was taking.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Tracy Costigan, Randall Brown, Frank Davidoff, and Sharon Wilson gave valuable suggestions and comments; Thomas Cook, John Ovretveit, Gareth Parry, and Matt Trujillo contributed to related papers on external validity; and Robert Wood Johnson Foundation grantees are implementing research on local adaptation that contributed to insights: Kimberly Freire, Leah Perkinson, Meredith Stocking, Russell Glasgow, Mary Naylor, and Felipe Gonzalez-Castro.

LITERATURE CITED

1. Ariz. Public Health Train. Cent. 2013. *Pareto Chart*. YouTube video, 2:55, posted by West. Reg. Public Health Train. Cent., Nov. 22, **https://youtu.be/pLWBG_CZ4ZY**
2. Avellar S, Paulsell D, Sama-Miller E, Del Grosso P, Akers L, Kleinman R. 2015. *Home Visiting Evidence of Effectiveness Review: Executive Summary*. Washington, DC: Off. Plan., Res. Eval., Adm. Children Fam., US Dep. Health Hum. Serv. **http://homvee.acf.hhs.gov/HomVEE_Executive_Summary_2015.pdf**

3. Avellar SA, Thomas J, Kleinman R, Sama-Miller E, Woodruff SE, et al. 2016. External validity: the next step for systematic reviews? *Eval. Rev.* **https://doi.org/10.1177/0193841X16665199**. In press
4. Backer TE. 2002. *Finding the Balance: Program Fidelity and Adaptation in Substance Abuse Prevention. A State of the Art Review.* **http://www.enap.ca/cerberus/files/nouvelles/documents/CREVAJ/Baker_2002.pdf**
5. Batalden M, Batalden P, Margolis P, Seid M, Armstrong G, et al. 2015. Coproduction of healthcare service. *BMJ Qual. Saf.* 25:509–17
6. Bishop DC, Pankratz MM, Hansen WB, Albritton J, Albritton L, Strack J. 2014. Measuring fidelity and adaptation: reliability of an instrument for school-based prevention programs. *Eval. Health Prof.* 37:231–57
7. Blakely CH, Mayer JP, Gottschalk RG, Schmitt N, Davidson WS, et al. 1987. The fidelity adaptation debate: implications for the implementation of public sector social programs. *Am. J. Community Psychol.* 15:253–68
8. Blase K, Fixsen D. 2013. *Core Intervention Components: Identifying and Operationalizing What Makes Programs Work*. Washington, DC: Off. Assist. Secr. Plan. Eval., Dep. Health Hum. Serv. **https://aspe.hhs.gov/basic-report/core-intervention-components-identifying-and-operationalizing-what-makes-programs-work**
9. Brown R. 2016. *Producing better evidence for building a culture of health*. Presented at the Robert Wood Johnson Found. Conf. Shar. Knowl. Build Cult. Health, 1st, March 10, Baltimore, Md.
10. Burchett H, Umoquit M, Dubrow M. 2011. How do we know when research from one setting can be useful in another? A review of external validity, applicability and transferability frameworks. *J. Health Serv. Res. Policy* 16:238–44
11. Campbell DT, Stanley JC. 1966. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally
12. CDC (Cent. Dis. Control Prev.). 2015. *CDC Community Health Improvement Navigator*. CDC, Atlanta. **http://www.cdc.gov/chinav/**
13. CDC (Cent. Dis. Control Prev.). 2015. *High impact HIV/AIDS Prevention Project (HIP) is CDC's approach to reducing HIV infections in the United States.* CDC, Atlanta. **https://effectiveinterventions.cdc.gov/en/HighImpactPrevention/Interventions.aspx**
14. CDC (Cent. Dis. Control Prev.). 2016. *Compendium of evidence-based interventions and best practices for HIV prevention.* Updated Sept. 1, CDC, Atlanta. **http://www.cdc.gov/hiv/prevention/research/compendium/index.html**
15. Cent. Study Prev. Violence. 2016. *Blueprints for Healthy Youth Development.* Inst. Behav. Sci., Univ. Colo., Boulder. **http://www.colorado.edu/cspv/blueprints/**
16. Chambers DA, Glasgow RE, Stange KC. 2013. The dynamic sustainability framework: addressing the paradox of sustainment amid ongoing change. *Implement. Sci.* 8:117
17. Chen HT. 2010. The bottom-up approach to integrative validity: a new perspective for program evaluation. *Eval. Prog. Plan.* 33:205–14
18. Chinman M, Acosta J, Ebener P, Malone PS, Slaughter M. 2016. Can implementation support help community-based settings deliver evidence-based sexual health promotion programs? A randomized trial of Getting To Outcomes®. *Implement. Sci.* 11:78
19. Collins LM. 2016. *Multiphase optimization strategy (MOST).* Methodol. Cent., Pa. State Univ., University Park. **https://methodology.psu.edu/ra/most/**
20. Community Prev. Serv. Task Force. 2016. *The Guide to Community Preventive Services.* Updated Sept. 26. Atlanta: Community Guide. **http://www.thecommunityguide.org/index.html**
21. Community Prev. Serv. Task Force. 2016. *Using Evidence to Improve Health Outcomes: Annual Report to Congress, Federal Agencies, and Prevention Stakeholders.* Atlanta: Community Guide. **http://www.thecommunityguide.org/annualreport/**
22. Cook TD. 2014. Generalizing causal knowledge in the policy sciences: external validity as a task of both multiattribute representation and multiattribute extrapolation. *J. Policy Anal. Manag.* 33:527–36
23. Cook TD, Campbell DT. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. New York: Rand McNally
24. Copi IM, Cohen C, Flage DE. 2007. *Essentials of Logic*. Upper Saddle River, NJ: Pearson Educ. 2nd ed.

25. County Health Rankings and Roadmaps. 2016. *What Works for Health*. Univ. Wis. Popul. Health Inst., Madison. **http://www.countyhealthrankings.org/roadmaps/what-works-for-health**
26. Cronbach LJ, Shapiro K. 1982. *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass
27. Dane AV, Schneider BH. 1998. Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clin. Psychol. Rev.* 18:23–45
28. Davidoff F, Dixon-Woods M, Leviton L, Michie S. 2015. Demystifying theory and its use in improvement. *BMJ Qual. Saf.* 24:228–38
29. Del Grosso P, Kleinman R, Mraz Esposito A, Sama-Miller E, Paulsell D. 2014. *Assessing the Evidence of Effectiveness of Home Visiting Program Models Implemented in Tribal Communities*. Washington, DC: Off. Plan., Res. Eval., Adm. Child Fam., US Dep. Health Hum. Serv. **https://www.mathematica-mpr.com/our-publications-and-findings/publications/assessing-the-evidence-of-effectiveness-of-home-visiting-program-models-implemented-in-tribal-communities**
30. Durlak JA, DuPre EP. 2008. Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. *Am. J. Community Psychol.* 41:327–50
31. Dusenbury L, Brannigan R, Hansen WB, Walsh J, Falco M. 2005. Quality of implementation: developing measures crucial to understanding the diffusion of preventive interventions. *Health Educ. Res.* 20:308–13
32. Elliott DS, Mihalic S. 2004. Issues in disseminating and replicating effective prevention programs. *Prev. Sci.* 5:47–53
33. Finucane MM, Martinez I, Cody S. 2015. *What Works for Whom? A Bayesian Approach to Channeling Big Data Streams for Policy Analysis.* Work. Pap. No. 40. Princeton, NJ: Mathematica Policy Res. **https://www.mathematica-mpr.com/-/media/publications/pdfs/health/bayesian_approach_channeling_wp.pdf**
34. Freire KE, Perkinson L, Morrel-Samuels S, Zimmerman MA. 2015. Three Cs of translating evidence-based programs for youth and families to practice settings. *N. Dir. Child Adolesc. Dev.* 2015:25–39
35. Glanz K, Bishop DB. 2010. The role of behavioral science theory in development and implementation of public health interventions. *Annu. Rev. Public Health* 31:399–418
36. Glasgow RE, Vogt TM, Boles SM. 1999. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *Am. J. Public Health* 89:1323–27
37. Glasziou P, Chalmers I, Altman DG, Bastian H, Boutron I, et al. 2010. Taking healthcare interventions from trial to practice. *BMJ* 341:c3852
38. Godwin M, Ruhland L, Casson I, MacDonald S, Delva D, et al. 2003. Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. *BMC Med. Res. Methodol.* 3:28
39. Green LW. 2001. From research to "best practices" in other settings and populations. *Am. J. Health Behav.* 25:165–78
40. Green LW, Glasgow RE. 2006. Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. *Eval. Health Prof.* 29:126–53
41. Green LW, Glasgow RE, Atkins D, Stange K. 2009. Making evidence from research more relevant, useful, and actionable in policy, program planning, and practice: slips "twixt cup and lip." *Am. J. Prev. Med.* 37(6S1):S187–91
42. Green LW, Nasser M. 2012. Furthering dissemination and implementation research: the need for more attention to external validity. In *Dissemination and Implementation Research in Health: Translating Science to Practice*, ed. RC Brownson, GA Colditz, EK Proctor, pp. 305–26. New York: Oxford Univ. Press
43. Greenhalgh T, Robert G, MacFarlane F, Bate P, Kyriakidou O. 2004. Diffusion of innovations in service organizations: systematic review and recommendations. *Milbank Q.* 82:581–629
44. Hawe P. 2015. Lessons from complex interventions to improve health. *Annu. Rev. Public Health* 36:307–23
45. Hill LG, Maucione K, Hood BK. 2007. A focused approach to assessing program fidelity. *Prev. Sci.* 8:25–34
46. Hoffmann T, Glasziou P, Boutron I, Milne R, Perera R, et al. 2014. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ* 348:g1687
47. Horne CS. 2016. Assessing and strengthening evidence-based program registries' usefulness for social service program replication and adaptation. *Eval. Rev.* **https://doi.org/10.1177/0193841X15625014**. In press

48. Hulleman C, Cordray DS. 2009. Moving from the lab to the field: the role of fidelity and achieved relative intervention strength. *J. Res. Educ. Eff.* 2:88–110
49. Jagers RJ, Sydnor K, Mouttapa M, Flay BR. 2007. Protective factors associated with preadolescent violence: preliminary work on a cultural model. *Am. J. Community Psychol.* 40:138–45
50. Kessler R, Glasgow RE. 2011. A proposal to speed translation of healthcare research into practice. *Am. J. Prev. Med.* 40:637–44
51. Kruschke JK. 2011. *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Burlington, MA: Academic
52. Leviton LC. 2015. External validity. In *International Encyclopedia of the Social and Behavioral Sciences*, ed. JD Wright, pp. 617–22. Oxford, UK: Elsevier. 2nd ed.
53. Leviton LC, Guinan ME. 2003. HIV prevention and the evaluation of public health programs. In *Dawning Answers: How the HIV/AIDS Epidemic Has Helped to Strengthen Public Health*, ed. RO Valdiserri, pp. 155–76. Oxford, UK: Oxford Univ. Press
54. Leviton LC, Trujillo MD. 2016. Interaction of theory and practice to assess external validity. *Eval. Rev.* **https://doi.org/10.1177/0193841X15625289**. In press
55. March JG. 1994. *A Primer on Decision Making: How Decisions Happen*. New York: Free Press
56. Mercer SM, DeVinney BJ, Fine LJ, Green LW, Dougherty D. 2007. Study designs for effectiveness and translation research: identifying trade-offs. *Am. J. Prev. Med.* 33(2):139–54
57. Minkovitz CS, O'Neill KMG, Duggan AK. 2016. Home visiting: a service strategy to reduce poverty and mitigate its consequences. *Acad. Pediatrics* 16:S105–11
58. Deleted in proof
59. Morrel-Samuels S, Hutchison P, Perkinson L, Bostic B, Zimmerman M. 2014. *Selecting, Implementing and Adapting Youth Empowerment Solutions*. Ann Arbor: Univ. Mich. Sch. Public Health
60. Naylor MD, Feldman PH, Keating S, Koren MJ, Kurtzman ET, et al. 2009. Translating research into practice: transitional care for older adults. *J. Eval. Clin. Pract.* 15:11–70
61. Newman MEJ. 2005. Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* 46:323–51
62. NHS Scotl. 2008. *The Glenday Sieve.* NHS Scotl., Edinburgh. **www.qihub.scot.nhs.uk/media/215450/glenday_sieve_presentation.ppt**
63. Olds DL, Henderson CR Jr., Tatelbaum R, Chamberlin R. 1986. Improving the delivery of prenatal care and outcomes of pregnancy: a randomized trial of nurse home visitation. *Pediatrics* 77:16–28
64. Patton MQ. 2010. *Developmental Evaluation*: *Applying Complexity Concepts to Enhance Innovation and Use*. New York: Guilford
65. Pawson R, Tilley N. 1997. *Realistic Evaluation*. Thousand Oaks, CA: Sage
66. Pearson M, Coomber R. 2010. The challenge of external validity in policy-relevant systematic reviews: a case study from the field of substance misuse. *Addiction* 105:136–45
67. Perkinson L. 2012. *Environmental scan of adaptation guidance*. Unpubl. Rep., CDC Found., Atlanta
68. Peters J, Langbein J, Roberts G. 2015. *Policy Evaluation, Randomized Controlled Trials, and External Validity: A Systematic Review*. Ruhr Econ. Pap. 589. Bochum, Ger.: Ruhr Econ. Pap. **http://en.rwi-essen.de/media/content/pages/publikationen/ruhr-economic-papers/rep_15_589.pdf**
69. Robling M, Bekkers M-J, Bell K, Butler CC, Cannings-John R, et al. 2016. Effectiveness of a nurse-led intensive home-visitation programme for first-time teenage mothers (Building Blocks): a pragmatic randomised controlled trial. *Lancet* 387:146–55
70. Rogers E. 2003. *Diffusion of Innovations*. New York: Free Press. 5th ed.
71. Rohrbach LA, Grana R, Sussman S, Valente TW. 2006. Type II translation: transporting prevention interventions from research to real-world settings. *Eval. Health Prof.* 29:302–33
72. Rothwell PM. 2005. External validity of randomised controlled trials: "To whom do the results of this trial apply?" *Lancet* 365:82–93
73. SAMHSA (Subst. Abuse Ment. Health Serv. Adm.). 2014. *Evidence based programs/NREPP*. SAMHSA, Rockville, Md. **http://www.samhsa.gov/data/evidence-based-programs-nrepp**
74. Schatz F, Welle K. 2016. *Qualitative Comparative Analysis: A Valuable Approach to Add to the Evaluator's 'Toolbox'? Lessons from Recent Applications*. CDI Pract. Pap. 13. Brighton, UK: Inst. Dev. Stud. **https://www.ids.ac.uk/F3B60FE0-C360-11E5-86F9005056AA4991**
75. Schouten LM, Hulscher ME, van Everdingen JJ, Huijsman R, Grol RP. 2008. Evidence for the impact of quality improvement collaboratives: systematic review. *BMJ* 336:1491–94

76. Shadish WR, Cook TD, Campbell DT. 2002. *Experimental and Quasiexperimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin
77. Shadish WR, Cook TD, Leviton LC. 1991. *Foundations of Program Evaluation: Theorists and Their Theories*. Newbury Park, CA: Sage
78. Steckler A, McLeroy KR. 2008. The importance of external validity. *Am. J. Public Health* 98:9–10
79. Sundell K, Beelmann A, Hasson H, von Thiele Schwarz U. 2015. Novel programs, international adoptions, or contextual adaptations? Meta-analytical results from German and Swedish intervention research. *J. Clin. Child Adolescent Psychol.* 45:784–96
80. Tipton E. 2013. Improving generalizations from experiments using propensity score subclassification: assumptions, properties, and contexts. *J. Educ. Behav. Stat.* 38:239–66
81. Tipton E, Hallberg K, Hedges LV, Chan W. 2016. Implications of small samples for generalization: adjustments and rules of thumb. *Eval. Rev.* **https://doi.org/10.1177/0193841X16655665**. In press
82. Tipton E, Hedges LV, Vaden-Kiernan M, Borman GD, Sullivan K, Caverly S. 2014. Sample selection in randomized experiments: a new method using propensity score stratified sampling. *J. Res. Educ. Eff.* 7:114–35
83. Tipton E, Peck LR. 2016. A design-based approach to improve external validity in welfare policy evaluations. *Eval. Rev.* **https://doi.org/10.1177/0193841X16655656**. In press
84. US Prev. Serv. Task Force. 2014. *The Guide to Clinical Preventive Services 2014*. Rockville, MD: Agency Healthc. Res. Qual. (AHRQ). **http://www.uspreventiveservicestaskforce.org/Home/GetFileByID/989**
85. Wandersman A, Alia K, Cook BS, Hsu LS, Ramaswamy R. 2016. Evidence-based interventions are necessary but not sufficient for achieving outcomes in each setting in a complex world: Empowerment Evaluation, Getting To Outcomes, and demonstrating accountability. *Am. J. Eval.* **https://doi.org/10.1177/1098214016660613**. In press
86. Wang VL, Ephross PH, Green LW. 1975. The point of diminishing returns in nutrition education through home visits by aides: an evaluation of EFNEP. *Health Educ. Monogr.* 3:70–88
87. Wash. State Inst. Public Policy. 2015. *Interventions to promote health and increase health care efficiency: December 2015 update*. Wash. State Inst. Public Policy, Olympia. **http://www.wsipp.wa.gov/ReportFile/1622/Wsipp_Interventions-to-Promote-Health-and-Increase-Health-Care-Efficiency-December-2015-Update_Report.pdf**
88. Weiss CH. 1997. *Evaluation: Methods for Studying Programs and Policies*. New York: Prentice Hall. 2nd ed.
89. Williams SJ, Aitken J, Radnor Z, Esain A. 2016. *Patient-centric and process-centric healthcare supply chains: the role of the broker*. Presented at Int. Organ. Behav. Healthc. Conf., 10th, April 5–6, Cardiff, UK