Research design issues for evaluating complex multicomponent interventions in neighborhoods and communities

Kelli A. Komro,[1] Brian R. Flay,[2] Anthony Biglan,[3] Alexander C. Wagenaar[4]

[1]Department of Behavioral Sciences and Health Education, Rollins School of Public Health, Emory University, Atlanta, GA, USA

[2]Department of Public Health, Oregon State University, Corvallis, OR, USA

[3]Oregon Research Institute, Eugene, OR, USA

[4]Department of Health Outcomes & Policy, College of Medicine, University of Florida, Gainesville, FL, USA

Correspondence to: K Komro
kkomro@emory.edu

Abstract

Major advances in population health will not occur unless we translate existing knowledge into effective multicomponent interventions, implement and maintain these in communities, and develop rigorous translational research and evaluation methods to ensure continual improvement and sustainability. We discuss challenges and offer approaches to evaluation that are key for translational research stages 3 to 5 to advance optimized adoption, implementation, and maintenance of effective and replicable multicomponent strategies. The major challenges we discuss concern (a) multiple contexts of evaluation/research, (b) complexity of packages of interventions, and (c) phases of evaluation/research questions. We suggest multiple alternative research designs that maintain rigor but accommodate these challenges and highlight the need for measurement systems. Longitudinal data collection and a standardized continuous measurement system are fundamental to the evaluation and refinement of complex multicomponent interventions. To be useful to T3–T5 translational research efforts in neighborhoods and communities, such a system would include assessments of the reach, implementation, effects on immediate outcomes, and effects of the comprehensive intervention package on more distal health outcomes.

Keywords

Health, Well-being, Evaluation design, Complex intervention, Translational research

Recent research suggests that significant improvements in the life trajectories of at-risk young people are now possible [1]. Unfortunately, actual well-being lags far behind, especially in areas of concentrated poverty, where the risk of young people developing multiple behavioral and health problems is much higher than other neighborhoods [2]. Major advances in population health will not occur unless we translate existing knowledge into effective multicomponent strategies, implement them in at-risk and underserved communities, and develop rigorous evaluation methods to ensure continual improvement. The success of translational research methods to bring about change in entire communities demands the evolution of methods that are most appropriate for the task at hand. In this paper, we present approaches to translational research stages T3–T5 [3] that advance the adoption, implementation, and maintenance of effective and replicable multicomponent interventions. We highlight experimental methods that can sharpen our understanding of the impact of complex interventions as well as each component program, policy, or practice that is implemented to support community development. These experimental methods will enable continuous quality improvement to ensure that neighborhood and community change efforts become increasingly effective and sustained over time. The methods are appropriate and applicable for research stages T3–T5 nationally and globally.

Most health outcomes have "multiple interacting influences crossing socio-ecological levels" and require complex, multi-level strategies to foster health and well-being [4] To illustrate the need for multi-level strategies for child health promotion, we created a causal model to guide community transformation efforts [5]. We first identified the major cognitive, behavioral, social, and health outcomes at each phase of child development and specified the major proximal and distal influences on each of these outcomes. We then identified evidence-based policies, programs, and practices that, implemented together, would best help achieve the desired social, health, and behavioral outcomes [6]. Disseminating, implementing, and evaluating complex comprehensive interventions in entire

<div style="border:1px solid">

Implications

Practice: Develop strong partnerships with researchers to implement and rigorously evaluate innovations for the promotion of health and well-being.

Policy: Fund national longitudinal and standardized measurement systems implemented and usable at the local level to advance the evaluation and refinement of multicomponent interventions to promote health.

Research: Partner with local and state agencies to implement and evaluate complex multicomponent intervention trials using multiple design elements to optimize causal inference.

</div>

neighborhoods and communities pose new challenges for the application of translational prevention science to community development. Hence, we present here major methodological tools for evaluating dissemination and implementation efforts that effectively and efficiently promote positive health outcomes.

CHALLENGES FOR OPTIMAL EVALUATION RESEARCH

We focus on three major challenges to optimal evaluation of health promotion and disease prevention efforts: (a) multiple contexts of evaluation/research, (b) complexity of packages of interventions, and (c) phases of evaluation/research questions. We then discuss alternative rigorous research designs able to accommodate these challenges.

Multiple contexts

The evaluation of complex, community-based, multicomponent interventions will need to take place within and across communities. Each community must collect evaluation data to monitor the adoption and implementation of each intervention component, track effects of the interventions, and evaluate the impact of the overall package. Also, across communities, evaluation research will be necessary to answer questions about effects of the initiatives en bloc.

To achieve sizable results, communities must provide a package of interventions across many settings and age groups. No single intervention will solve all problems in a community. For several reasons, even the best evidence-based interventions produce relatively small effects when taken to scale [7], meaning they affect only a portion of the population, or the effects are small for any one person or institution but important and large when extrapolated to the population as a whole. Different interventions also target particular outcomes or sub-populations (e.g., developmental stages, age groups, risk status). Therefore, multiple evidence-based interventions need to take place concurrently to produce larger effects at the community or population level. Because effects of multiple interventions could be due to additive effects across interventions or to synergistic mutually reinforcing effects, a viable approach to evaluation must consider (a) the effects of specific interventions separately and (b) the total effects of the package of interventions (usually not a simple sum of the component effects).

Complex multicomponent interventions

Patton sums up the difference between evaluating stand-alone programs and complex multicomponent interventions [8]. Simple programs, like recipes, are relatively easy to replicate and evaluate (with expected similar effects across replications). Complicated programs are more difficult to reproduce with fidelity and to replicate effects and, hence, more difficult to evaluate. Complex, multicomponent interventions, like those needed to socialize a population of children effectively, are even more difficult to replicate.

An evidence-based practice or strategy is most like a recipe. An individual program combines practices ranging from simple to complicated. A policy or structural change, including the process of getting a policy passed, implemented, and enforced, is more complicated. A package of multiple interventions is even more complex, with potential synergies and interactions; in some ways, the parts are indivisible from the whole. A successful package of interventions to help a community achieve large results is likely to contain multiple practices, programs, and policies. We use the term "component interventions" to refer to any one of these practices, programs, or policies and the term "multicomponent interventions" to refer to a set of multiple interventions implemented simultaneously.

The complexity of multicomponent intervention packages requires evaluation designs that are equally complex, that acknowledge differences between communities, and that assess intervention implementation, and how the strategies are adapted over time as communities improve their practice and learn what does and does not work. Traditionally, development and evaluation of individual programs have been "top-down"; that is, researchers or developers determine what is needed and how to achieve it, develop a program, evaluate its efficacy, then its effectiveness, and then offer it to the world [9]. In contrast, development and evaluation of a complex package of interventions is more likely to be "bottom-up" [10] neighborhoods determine what they need, adopt/adapt particular combinations of interventions, and evaluate for effectiveness (both individually and as a package) in a specific real-world setting [i.e., pragmatic rather than explanatory evaluation–11]. They are likely to focus as much on implementation as on outcomes–because they cannot achieve outcomes if they fail to deliver the interventions well. This approach is similar to Patton's [8] concept of "developmental evaluation," which involves changing the intervention, adapting it to changed circumstances, and altering tactics based on emergent conditions. This can be particularly useful for programs that evolve over time as they address emerging issues in changing environments.

Critical evaluation/research questions

Evaluating the effectiveness of interventions in real-world settings requires focus on assessing the multiple phases of program effects along specified causal pathways. First, multiple evidence-based programs must be adopted and staff trained to deliver them. Then, if an evidence-based intervention does not produce the expected long-term effects, most likely it was not implemented with integrity, did not reach and engage the target audience, or did not produce immediate expected proximal effects. In most cases, it should be possible to specify the expected immediate effects (on causal mediators) of these interventions that would instill confidence that they will result in longer-term

improvement in outcomes. For example, a parenting intervention should achieve some immediate improvement in parenting practices that mediates longer-term outcomes for youth (e.g., academic performance, behavioral skills, or health status).

Understanding the mechanisms of effects leads to an assessment of whether a package of interventions is performing in the way it is intended along the full range of its implementation, rather than simply an evaluation of its ultimate impact. Evaluating the expected mechanism of effect helps us understand which elements of the package are functioning well and which might be improved to achieve larger effects. For communities that receive support for complex, multicomponent interventions, a wide range of factors will affect implementation and outcomes, including cultural, societal, geographical, and political factors, as well as the presence of existing investments and activities addressing some or all of the same outcomes. Given the multiplicity of factors that influence outcomes, one of the main goals of evaluation is to assess a multicomponent intervention's *contribution* to changes in outcomes. Research designs incorporating many repeated measures, both before and after intervention implementation, are necessary to follow the trends in implementation, reach, and impact.

Thus, basic questions to determine which efficient research designs will help evaluate complex multicomponent interventions include:

1. Implementation: Is each intervention implemented with fidelity, so that it is likely that the effects can be replicated elsewhere? Evaluating effectiveness of evidence-based interventions in real-world settings requires much more focus on assessing adoption patterns, adaptations, and implementation level/intensity and integrity [12]. Key questions include the following: How many implementers or settings adopt the intervention? How often and how well is it delivered?

2. Reach: Is each intervention reaching the members of its intended target audience? An intervention cannot have expected effects on the target audience if it does not reach them [12]. Methods for testing strategies to expand reach to a high proportion of the target population must be included. We need to know the following: What is the level of engagement in and satisfaction with the intervention? Is the reach sufficient to achieve a potentially critical tipping point required for group-level normative or behavior change, and thereby change population-level outcomes?

3. Immediate effects: Does an intervention have the expected immediate effects on the mediating or proximal behaviors or processes? Without reliable immediate effects on mediators, chances of longer-term change on ultimate outcomes remain slim [13]. Here, the key questions include the following:

Does delivery of the intervention produce expected immediate effects, especially in terms of changes in behaviors or organizational practices? Are results the same for everyone? How well are these changes maintained over time?

4. Outcomes: Are there changes in longer-term outcomes? If so, is it possible to discern which intervention components have contributed to each effect? Key questions include: Does continuous delivery of the multicomponent intervention lead to the expected outcomes? How long do they take to occur? How well are they maintained? Do outcomes differ for different subgroups? Can we attribute effects on these outcomes to specific component interventions?

These questions not only are important for research but also are vital for effective policymaking and service delivery, maintenance of service delivery quality, continuous improvement in interventions, and public support needed to maintain interventions. Indeed, we see a human service and prevention system evolving, where distinctions between research and practice diminish as careful measurement and experimental evaluation of intervention processes and their effects become integral to program operation and service provision.

The evaluation designs described here will improve understanding of large-scale programmatic strategies for translational research to address the complex needs of vulnerable and underserved communities. Such research will provide rigorous, non-partisan, multidisciplinary, and independent assessments of complex multicomponent interventions to inform policymakers, the scientific community, program implementers, and stakeholders in education, social work, public health, and related fields.

EVALUATION RESEARCH DESIGNS

Designs for implementation and reach questions

We can answer these questions by tracking and continuous monitoring of who (or what settings/places) and how many adopt a particular intervention, and how often and how well they deliver that intervention. Collecting these data helps decision-makers and implementers ensure that they adopt each of numerous component interventions planned and that they are combined in a sequence or interacting package of interventions as planned.

Management information system records, observations, standardized reporting by implementers, or survey measures of implementation quality—together with comparison against established norms or benchmarks—help determine the quality of implementation of an intervention. This information helps providers improve the intensity level and quality of implementation. Understanding these aspects of implementation is critical to the interpretation of evaluation results and the enhancement of future or ongoing adoption or adaptation.

Designs for evaluating intervention effects

The main research designs are the same whether assessing immediate or long-term intervention effectiveness. In this section, we discuss designs in order of scientific rigor for establishing a causal relationship. After explication of stand-alone designs, we describe how a hierarchy of nested designs is useful.

Randomized controlled trials—Most of the substantial progress in the prevention and education sciences in recent decades is due to the increasing use of randomized controlled trials (RCTs); there is increasing advocacy for the use of RCTs to guide and evaluate government initiatives [14]. We cannot assume that existing evidence, often derived from implementations under optimal conditions, will guarantee the success of these interventions when implemented in new and more challenging settings.

Within communities, randomized trials provide the strongest evidence of effectiveness of component interventions, when circumstances permit randomization. For example, within a community, a lottery can select children for charter schools, as they did in the Harlem Children's Zone [15]. Where multiple interventions are delivered in one kind of setting (e.g., families, schools, clinics), and there are insufficient resources to offer them to all instances of the setting, then settings might be randomly assigned to receive or not receive the package of interventions. This approach is likely to be politically unpopular because community members likely prefer that everyone benefits, but a lottery may be the fairest method of distribution when resources prevent universal implementation or in situations of equipoise—whether the intervention helps or not is truly unknown. Randomization, with enough units for adequate statistical power, generates the strongest evidence of effectiveness in real-world settings.

Research involving multiple communities would also benefit from randomized trials if a sufficient number of communities adopt a similar set of interventions. Indeed, funders could rank community proposals in order of reviewers' scores and then randomly assign half of the best proposals to get funding for the intervention and the other half to be controls (perhaps committing funding to the controls at a future date). This approach is difficult to achieve, and there might still be substantial variation between communities in choice of interventions and how well they are implemented.

Regression-discontinuity designs—The regression-discontinuity design (RDD) is one of the most powerful "quasi-experimental" designs that, when implemented properly, can produce conclusions as clear as those obtained from randomized trials [16]. In RDDs, participants are assigned to conditions based on a continuous variable, often a measure of need, merit, or risk (e.g., children assigned to receive school lunch programs if their household income falls below a specified threshold). The functional relationship between the known assignment variable (e.g., household income) and the outcome variable (e.g., health, school achievement), estimated separately for the treated group that falls below the threshold and the control group that falls above the threshold, provides the basis for causal inference. Because treatment assignment is determined fully by the assignment variable, inference of a treatment effect of the program is warranted if there is a discontinuity at the threshold where the treatment is introduced. Strong causal inferences are possible when the design's rigorous implementation standards are met.

RDDs are statistically less efficient than randomized trials and need about three times as many cases to reach the same statistical power as a randomized trial [17]. As a result, they are rarely appropriate when the assignment unit is a larger aggregate (e.g., city or state), but highly effective when individuals are assigned to treatments based on a cutoff variable. For example, instead of using a lottery to assign children to a new form of school, one might use scores on a prior test or an admissions exam. This approach would be politically acceptable because it would allow communities to provide the new services to those most in need of them.

Limitations of RCTs and RDDs—By themselves, RCTs and RDDs are insufficient for advancing translational prevention research in communities. As the size of the unit receiving an intervention increases (e.g., from individuals to whole communities or even states) and the complexity of the intervention increases (e.g., from one clearly defined program to multiple policies, programs, and practices), it becomes more difficult to conduct RCTs or RDDs. The most obvious reason for this is that it becomes difficult to include a sufficient number of units to estimate intervention effects reliably. Second, RCTs and RDDs require standardization of the intervention across all units receiving a "treatment." This is more difficult to achieve in neighborhood or community interventions.

Finally, the most significant problem with employing RCTs or RDDs to evaluate complex multicomponent community interventions at this stage of our knowledge is that they preclude further improvements in the intervention. A method is required for systematically evaluating the functional effects over time of intervention components in a continuous quality improvement mode of operations. To accumulate enough knowledge to justify the next round of large-scale trials, methods must reliably replicate relationships between intervention components and measured processes in individual communities.

Matched control group designs—Other quasi-experimental designs are alternatives to RCTs or RDDs [17]. Most quasi-experimental designs involve identifying appropriate comparison groups (counterfactuals). Recent work suggests that the use of matched controls can provide the same estimates of effects as RCTs or RDDs, but *only* when the controls are local and closely matched on the outcome variable (or some close approximation of it), which must be measured with high reliability [18].

Recent work also suggests that careful application of modern analytic methods can reduce biases in non-matched and self-selected control groups [19]. Such methods include (a) ordinary linear regression, predicting outcome from condition and a rich set of observed covariates, and (b) propensity score stratification following careful weighting and covariance adjustment criteria to maximize balance [20]. Even with the use of propensity score methods, great care must be taken to avoid "hidden bias" [19]. For mediation models, adding propensity scores has been shown to reduce bias [21].

Designs involving repeated measures–Repeated measures strengthen randomized trials [17], but intensive longitudinal designs with many repeated observations are particularly useful in non-randomized designs. The fundamental feature of these designs is observable change in a series of data points after introduction of an intervention or independent variable. At the simplest level, a change over time in either the slope or intercept of the repeatedly measured process or outcome is evidence of the effect of the intervention [17]. Of importance in community-based work, these designs allow for adaptations or adjustments (in light of most recent results) in the program or its implementation to improve effectiveness, with the effects of the adjustment observed in further repeated data waves. The fundamental issue in these designs is whether we can have confidence that any observed change in a repeatedly measured process is, in fact, due to the intervention, and not other simultaneous events. The two most important design elements include a long time series, along with well-designed sets of comparisons.

Interrupted time-series designs–Rigorous interrupted time-series designs require many data points, typically a minimum of 30 and often a hundred or more, both before and after an intervention [22]. Where sufficient pre-intervention data exist on a regular basis (e.g., daily, weekly, monthly, or annual assessments), repeating the same assessments post-intervention creates an interrupted time series. Time-series designs can be used for either individual interventions or packages of interventions and for a single community or multiple communities. Having a sufficiently large number of repeated data points provides greater confidence in the reliability of the claimed relationship between the independent variable and the dependent variable. However, replicating that relationship across cases (e.g., persons, neighborhoods, schools, or cities) is also very important. Confidence in the evidence from those replications improves as a function of (a) whether the data come from multiple cases at the same time (as opposed to a series of replications over time), (b) whether the timing of implementation is staggered across cases, (c) whether receipt of the intervention across cases is determined at random, (d) how large the effect is on intercept or slope, (e) how reliable these effects are (as a function of the number of data points and size of the effects), and (f) whether the relationship between independent and dependent variables is

shown to reverse when or if the intervention is subsequently withdrawn.

Multiple-baseline designs–The label multiple-baseline intuitively refers to multiple pre-intervention measures or pretests. However, in the behavior analysis literature since the 1960s, the term has been used to refer to multiple settings, units/cases, or outcomes. We use the term multiple-baseline to refer to both multiple pretests and multiple cases, where there is staggered introduction of the intervention across cases and where cases can be groups or places (e.g., communities or subsets of communities). In some respects, this type of design might be the workhorse of community intervention evaluation and also of dissemination and implementation translational research [23]. As interventionists start work with one case or one group of cases (e.g., a preschool, a family, a classroom, several blocks of a community, a community), they can adapt and improve the intervention before working with the next case(s).

Strengthening research designs—a hierarchical approach

Adding more waves of measurement, both before and after interventions start, improves all designs, including randomized trials. Including multiple settings also helps; indeed, including multiple communities can improve all of the above designs. At the community level, any design will (and should) include multiple dependent variables, and it is expected that the package of interventions will produce effects on all of the targeted outcomes. The inclusion of comparison-dependent variables that are *not* expected to change due to the intervention is an additional way to improve the design. Careful use of sophisticated analytical approaches can also improve interpretation of results, including casual inference, from any design, randomized or not.

It is unlikely that the first attempt to implement a package of interventions will be totally successful; adjustments will be necessary. Therefore, a continuous improvement approach is desirable. A time-series or multiple-baseline design with monthly assessments of implementation and intermediate outcomes could simultaneously address the needs for a rigorous evaluation and for a practical way of implementing and refining the intervention. Careful monitoring of implementation fidelity, uptake, and effects of each of the components of the intervention can provide continuous feedback to guide refinement of the intervention. Thus, intervention implementation procedures and their effects improve over time. This is an important learning process that is natural in the refinement of any intervention. In this way, continuous evaluation and refinement of the intervention (including implementation strategies and intervention components) can occur within one neighborhood or community and the lessons learned applied to implementation in subsequent communities.

THE FUNDAMENTAL IMPORTANCE OF A MEASUREMENT SYSTEM

Traditionally, organizations have delivered educational and human service interventions without accompanying efforts to monitor the quality or effects of the intervention. It is increasingly clear, however, that effective education and human services need ongoing systems for monitoring dissemination and the fidelity of implementation of evidence-based interventions, the proportion of the target population reached, their immediate effects on targeted mediators or processes, and their ultimate effects on longer-term outcomes.

A standardized measurement system is fundamental to the evaluation of complex multicomponent interventions. Such a system, to be usable in local neighborhoods and communities across the country and between countries, would include assessments of (a) implementation of each intervention component, (b) reach of each component, (c) effects of each component on immediate effects, and (d) effects of the comprehensive intervention package on outcomes. Such a measurement system would include the following nine dimensions:

1. Be standardized and usable for evaluations within and between communities.
2. Include a comprehensive set of reliable and valid measures.
3. Be easy to use by a variety of individuals and organizations within and across communities.
4. Allow local evaluators to select the measures most relevant to their community.
5. Include key measures of process, intermediate and primary outcomes.
6. Be readily accessible to communities within a country and across countries (e.g., economically, technologically, and culturally).
7. Rapidly provide results to relevant individuals and organizations to inform continuous quality improvement.
8. Be secure and protect confidentiality.
9. Support the use of unique identifiers in order to aggregate data from multiple contexts and at multiple levels (e.g., individuals within families, within schools/organizations, within neighborhoods/communities, etc.).

Measuring process, implementation, and reach—Systematic methods to monitor community change processes have been developed as part of complex community-wide intervention research. Others have designed standardized web-based systems to track complex dimensions of community participation and action, and systems to monitor school- and family-based implementation fidelity. Such systems could grow into a standardized, yet highly flexible, web-based system that becomes a routine function within community-based organizations. A user-friendly measurement system to document and monitor adoption, implementation, and reach of each intervention component is not only vital to an overall evaluation effort; it is part of a continuous quality improvement system to improve the effectiveness of interventions over time. Other side benefits include improved management of local efforts, since the data from the measurement system also serve as a management information system for local agencies and organizations.

Monitoring outcomes—NIH and CDC support several initiatives to monitor health systematically at the state and national levels. However, these systems do not allow for monitoring of outcomes at the local level, where many preventive initiatives exist. Typically, these systems have a limited focus on a single health issue. Therefore, we propose the development of a standardized and comprehensive measurement system that local organizations can implement feasibly and routinely at the local level. It should include measures of distal and proximal influences (measureable intermediate outcomes of individual community strategies) and comprehensive primary outcomes of child health, including cognitive, social and emotional, behavioral, and physical health domains. The availability of individual-level time series data within and across communities will eventually be invaluable for translational research and the evaluation of community-level interventions as long as analysts carefully apply research design elements improving causal attribution and correctly apply modern statistical methods.

SUMMARY RECOMMENDATIONS

Thanks to the accumulation of numerous evidence-based preventive interventions, the growing availability of comprehensive measures of child and adolescent well-being, and advances in research design and statistical analysis, the methodological practices of translational prevention research are evolving. As neighborhoods and communities attempt to disseminate and implement existing evidence-based practices to achieve population-wide benefits, they will need tools that support continuous quality improvement of component interventions and complex packages of interventions, as well as strategies for ensuring reach and fidelity of implementation. In this paper, we have described major options available for these tasks.

Careful attention to rigorous evaluation strategies for each individual intervention component can help design an overall evaluation of the effectiveness of complex multicomponent initiatives. Attention to designing rigorous evaluations and continuous quality improvement efforts for each intervention component will provide the building blocks for a more successful overall effort. The elements and principles necessary to strengthen causal inference should be applied to the evaluation of each intervention component, as well as to the overall effort. Causal models should be prepared for each intervention component, specifying the intervention component's inputs and adoption,

intervention delivery, and immediate and ultimate outcomes—see our child health and well-being model as one example [5, 6]. Critical for a rigorous evaluation of each component, as well as the overall evaluation, is implementation of a standardized, efficient system to measure adoption, delivery and reach, and both immediate and ultimate outcomes for each intervention component. To guide decisions for optimum use of multiple design elements for constructing experiments and quasi-experiments, we suggest a hierarchical decision-making approach as follows:

1. Randomization by units or by time whenever feasible and acceptable at community, neighborhood, family, organization, and/or individual levels.
2. When randomization is not feasible, use matched comparisons (sites and/or outcomes), in combination with (a) time-series or multi-level designs with (at least) monthly assessments of implementation, intermediate and ultimate outcomes, in addition to (b) elements of adaptive intervention designs, feedback loops, and continuous quality improvement and (c) careful use of modern statistical methods.

A monitoring and measurement system is a necessary component of the kind of evaluation/research we envision for complex multicomponent interventions. Bringing about permanent improvements in the prevalence of successfully developing children requires such a system, in the same way that having good measures of economic performance is vital to management of our economy.

In summary, the interventions that neighborhoods and communities will implement should lead to permanent changes in practices—as long as the data show their value—leading to the final outcome of translational research, which is universal change in attitudes, policies, and social systems [3]. Thus, we envision a system of continuous quality improvement and translational research methods that begin with the implementation of practices that previous research stages have found valuable. Then, the evidence of their effects in specific neighborhoods and communities further shapes changes to social systems. The maintenance and institutionalization of effective interventions will require both ongoing evidence of their continued value and ongoing quality implementation.

Compliance with ethical standards
Conflict of interest: The authors declare that they have no competing interests.

Adherence to ethical standards: No human subject research was involved.

References

1. National Research Council, Institute of Medicine. *Preventing mental, emotional, and behavioral disorders among young people: progress and possibilities*. Washington, DC: The National Academies Press, 2009.
2. Yoshikawa H, Aber JL, Beardslee WR. The effects of poverty on the mental, emotional, and behavioral health of children and youth: implications for prevention. *Am Psychol*. 2012; 67(4): 272-84. doi:10.1037/a0028015.
3. Fishbein DH, Stahl M, Ridenour TA, Sussman S. Translating emerging research into prevention of high-risk behaviors: state-of-the-science, barriers, and innovations. Translational behavioral medicine 2015
4. Holmes BJ, Finegood DT, Riley BL, Best A. Systems thinking in dissemination and implementation research. In: Brownson RC, Colditz GA, Proctor EK, eds. *Dissemination and implementation research in health*. New York: Oxford University Press; 2012: 175-91.
5. Komro KA, Flay BR, Biglan A, the Promise Neighborhoods Research Consortium. Creating nurturing environments: a science-based framework for promoting child health and development within high-poverty neighborhoods. *Clin Child Fam Psychol Rev*. 2011; 14(2): 111.
6. Komro KA, Tobler AL, Delisle AL, O'Mara RJ, Wagenaar AC. Beyond the clinic: improving child health through evidence-based community development. *BMC Pediatr*. 2013; 13(172): 1471-2431.
7. Glasgow RE, Emmons KM. How can we increase translation of research into practice? Types of evidence needed. *Annu Rev Public Health*. 2007; 28: 413-33.
8. Patton MQ. *Utilization-focused evaluation*. 4th ed: Guilford Press, 2008
9. Flay BR, Biglan A, Boruch RF, et al. Standards of evidence: criteria for efficacy, effectiveness and dissemination. Prevention Science 2005;6(3):151–75 doi: DOI 10.1007/s11121-005-5553-y
10. Chen HT. The bottom-up approach to integrative validity: a new perspective for program evaluation. *Eval Program Plann*. 2010; 33(3): 205-14.
11. Thorpe KE, Zwarenstein M, Oxman AD, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *Can Med Assoc J*. 2009; 180(10), E47.
12. Brownson RC, Colditz GA, Proctor EK. *Design and analysis in dissemination and implementation research*: Oxford University Press, 2012
13. MacKinnon DP, Fairchild AJ. Current directions in mediation analysis. *Current Directions in Psychological Science*. 2009; 18(1): 16-20.
14. Haskins R, Baron J. *Building the connection between policy and evidence: The Obama evidence-based initiatives*. Nesta: London, England; 2011.
15. Dobbie W, Fryer RG Jr. *Are high quality schools enough to close the achievement gap?* NBER: Evidence from a social experiment in Harlem; 2009.
16. Cook TD, Shadish WR, Wong VC. Three conditions under which experiments and observational studies produce comparable causal estimates: new findings from within-study comparisons. *J Policy Anal Manage*. 2008; 27(4): 724-50.
17. Shadish WR, Cook TD, Campbell DT. *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin Company; 2002.
18. Cook TD, Steiner PM. Case matching and the reduction of selection bias in quasi-experiments: the relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis. *Psychol Methods*. 2010; 15(1): 56-68.
19. Shadish WR, Clark MH, Steiner PM. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *J Am Stat Assoc*. 2008; 103(484): 1334-44.
20. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*. 2001; 2(3): 169-88.
21. Coffman DL. Estimating causal effects in mediation analysis using propensity scores. *Structural Equation Modeling: A Multidisciplinary Journal*. 2011; 18(3): 357-69.
22. Wagenaar AC, Komro K. Natural experiments: research design elements for optimal causal inference without randomization. In: Wagenaar AC, Burris S, eds. *Public health law research: theory and methods*. San Francisco: Jossey-Bass; 2013.
23. Landsverk J, Brown CH, Chamberlain P, et al. Design and analysis in dissemination and implementation research. In: Brownson RC, Colditz GA, Proctor EK, eds. *Dissemination and implementation research in health*. New York: Oxford University Press; 2012.