Research on Social Work Practice
Volume 19 Number 5
September 2009 560-568
© 2009 The Author(s)
10.1177/1049731509335497
http://rswp.sagepub.com

Critical Measurement Issues in Translational Research

Russell E. Glasgow
Kaiser Permanente

This article summarizes critical evaluation needs, challenges, and lessons learned in translational research. Evaluation can play a key role in enhancing successful application of research-based programs and tools as well as informing program refinement and future research. Discussion centers on what is unique about evaluating programs and policies for implementation impact (or potential for dissemination). Central issues reviewed include the importance of context and local issues, robustness and external validity issues, multiple levels of evaluation, implementation fidelity versus customization, choosing evaluation designs to fit questions, and who participates and characteristics of success at each stage of program recruitment, delivery, and outcome. The use of mixed quantitative and qualitative methods is especially important, and the primary redirection that is needed is to focus on questions of decision makers and potential adoptees rather than the research colleagues.

Keywords: *translation; evaluation; dissemination; implementation; external validity*

It is widely recognized that there is a serious problem concerning the slow and incomplete transfer of research findings into practice (Institute of Medicine & Committee on Quality Health Care in America, 2003). This seems to be true across diverse content areas, countries, and areas of specialization (McGlynn et al., 2003). There are multiple and interacting reasons for the present situation, including resources, training, reimbursement, and other policies, priorities, vested interests, and political will (Kingdon, 1995). This article focuses on how measurement can facilitate successful transfer of research to real-world practice and policy. It provides a general overview and background for more specific evaluation issues discussed in later papers. The article begins by discussing the unique characteristics of implementation and dissemination research and the related evaluation implications. It illustrates application of a research translation model (RE-AIM) to focus attention on key measurement issues and concludes with a list of specific translation challenges and measurement recommendations.

Unique Features of Translational Research

Key features of implementation research are that it is concerned with either (a) evaluating typical citizens in typical settings receiving interventions delivered by typical staff (Glasgow & Emmons, 2007a; Green & Ottosen, 2004) or (b) determining whether a program works in a specific type of real-world setting and what types of patients, staff, and delivery conditions are associated with success (Pawson, Greenhalgh, Harvey, & Walshe, 2005). This is in contrast to efficacy research in which the evaluation focuses on a subset of persons most likely to benefit and without confounding factors, in which the research is often conducted in leading universities.

The primary purpose of translational research is to address practical questions that key decision and policy makers are likely to have (e.g., can this program work here, how much will it cost, who can successfully deliver the program?). Theory is important in implementation and dissemination research, but questions revolve around how the theory is operationalized and implemented rather than around more basic theoretical questions.

Almost all who have conducted this type of research have commented on the importance of context. To capture context adequately often requires relatively comprehensive measurement, as described in later sections. The programs and policies evaluated are usually complex and often multilevel, which present added evaluation challenges. A related feature of translational

Author's Note: Correspondence concerning this article should be addressed to Russell E. Glasgow, PhD, Kaiser Permanente Colorado, P. O. Box 378066, Denver, CO 80237-8066; e-mail: russg@re-aim.net.

research is that interventions or policies often evolve over time (Rotheram-Borus, Flannery, & Duan, 2004). Sometimes this is intentional as in rapid cycle quality improvement programs (Berwick, 1996), and other times it is unintentional and due to changes or drift in staff, available resources, or priorities.

The characteristics of translational research that are intended to inform policy and practice also have important implications for how evaluations are conducted. These implications are discussed below.

Context

When evaluating implementation, it is helpful to ask the who, what, where, why, and how questions that a journalist might ask (Table 1). A key question that is challenging to answer, but which has enormous public health implications, is ‘‘who participates—and who does not?’’ Typically, this question is answered only in terms of the numerator of the number of participants (e.g., citizens, employees, students) who take part, the number of staff who delivered the program, or the number of settings included. By itself, such information is only moderately helpful. Much more informative is to also collect data on the denominators of the numbers of participants, staff, and settings invited to participate and on the similarities and differences between those who take part and those who do not at each of these levels. In particular, greater attention is needed to the range and representativeness of settings and staff involved in a project as reviews have found that this information is reported less often than is representativeness data on patients (Glasgow, Klesges, Dzewaltowski, Bull, & Estabrooks, 2004).

Most projects can collect numerator and denominator information by simply keeping careful records and summarizing this information as well as the number of and reasons for exclusions at each of the setting, staff, and patient levels in Point 1 of the table. Due to confidentiality or logistical issues, it can be more challenging to collect information on characteristics of those who decline to participate. In such cases, a useful fall-back strategy is to rely on existing data sources, such as local census data, reports from health departments, or organizational records, to compare the characteristics of the target population in that area (e.g., all employees in the workforce, all patients in a health plan—see www.re-aim.org) to those who participate.

Table 1
Journalist Questions for Assessing Implementation and Potential for Dissemination

1. Who comes? (and who does not)—at following levels:
   a. Setting: Which organizations (e.g., worksites, medical plans, schools) were approached—How many participated?
   b. Staff: Which staff members participated?
   c. Individual patients, consumers, end users: How many and what types of people participated?
2. What outcomes are produced? (intended and unintended)
   a. How much change is observed on key dependent variables?
   b. What is the impact on quality of life?
   c. Were any negative impacts produced?
3. Where and for whom will this program work?
   a. What types of settings and staff members are most successful?
   b. What patient/user characteristics are associated with success?
4. Why were these results found?
   a. How did change come about (what were the mediators?)
   b. What contextual factors were important?
5. How consistently was the program/policy delivered?
   a. Across different program components?
   b. Across staff?
   c. Over time? (Did program change?)
6. How long-lasting are the effects?
   a. What was the attrition—at setting, staff, and individual levels—and how did this impact results?
   b. To what extent were changes maintained over time?
   c. Was the program or policy institutionalized, modified (and how), or discontinued?

What outcomes? The second key question in Table 1 concerns the magnitude and breadth of improvements produced by a program or policy. Often research reports are limited to a narrow assessment of impact on a preidentified key outcome. Magnitude of change on this primary dependent variable is certainly one important aspect of outcome measurement. Equally important, however, are answers to the related question of impact on quality of life—considered by many to be the ultimate outcome of social and public health interventions (Kaplan, 2003)—and to know if any negative or unanticipated results occurred. Quality-of-life measures provide a common metric that is helpful in making resource decisions across different content areas. Often program and policy developers have difficulty identifying potential adverse events that might occur as a result of a new program. One of the main ways that programs can have an unintended negative impact is that by focusing efforts on a given area (e.g., diabetes or obesity), busy and underresourced settings may do less in other areas that are also part of their mission, such as mental health or cancer screening.

Conditional outcomes. The third set of questions in Table 1 assesses the breadth of conditions under which a program is successful, which is sometimes referred to as robustness of effects. At the setting level, this refers to organizational characteristics related to success. For example, are only well-resourced settings that employ a multidisciplinary team approach able to achieve success? At the individual or consumer level, a key robustness issue is whether results are uniform or differential across recipient characteristics, such as race, ethnicity, income, education, gender, age, health literacy, and risk levels related to health inequities.

Understanding why. The fourth key issue is to provide information on how and why the pattern of outcomes observed was found. Many disciplines call this process or mechanism evaluation (Linnan & Steckler, 2002), and the goal is to understand how the program or policy achieves its effects (or why it did not succeed, or was only effective for a subset of participants). Qualitative approaches are often helpful to elucidate such understandings, which can inform both program refinement and the underlying theory. For quantitative data, Baranowski, Lin, Wetter, Resnicow, and Davis (1997) and Reynolds, Buller, Yaroch, Maloy, and Cutter (2006) discussed specific analysis steps to determine whether hypothesized theoretical variables are causally related to (mediate) outcomes.

Implementation consistency. The fifth question in Table 1 concerns how consistently programs are delivered across different intervention components, staff, recipients, and time. Consistency of delivery by typical staff is especially key in evaluations conducted in real-world settings because failure to adequately implement a program is one of the most frequent reasons for failure in dissemination research (Basch, Sliepcevich, & Gold, 1985; Bond, 2007). It is important to understand both the extent to which different components of a program are delivered as intended and whether there are staff characteristics (e.g., education, profession, experience, similarity to recipients) associated with successful program implementation.

It is also important to track program/policy implementation over time. Intervention delivery patterns can drift over time, both intentionally and unintentionally. The issue of program fidelity versus customization is currently an active area of investigation and controversy, and is discussed in more detail below as well as in several of the articles in this issue.

Sustainability. The final question concerns the longevity of programs and their longer-term effects at both the setting and individual levels. If an organization or government agency is going to make an investment in a new program or policy and devote the time and resources involved in training, supervision, infrastructure, and so on, it wants to have a reasonable expectation that both the program (policy) and its effects will stand up over time.

At the individual level, there are two key evaluation issues related to sustainability. The first is attrition. It is often challenging to track participants over time in today's mobile society, but attrition can produce misleading conclusions if it is not taken into account using appropriate imputation methods. This is especially the case if attrition rates are high, are related to participant characteristics (especially to success), or are differential across program conditions. The other well-known issue is that of maintenance of change over time. Many problem behaviors and societal issues can be modified over a short period of time, but long-term maintenance is a much greater challenge (Orleans, 2000).

At the setting level, the key question concerns whether the policy or program is continued intact, discontinued entirely, or modified following an initial evaluation period. There are few data as to the extent to which organizations adapt, modify, or discontinue programs over time (Glasgow et al., 2004), but it is rare that a program is continued in exactly the same way it was initially introduced. Rotheram-Borus et al. (2004) have discussed the need to study evolution of programs over time to enhance understanding of translation issues.

Comprehensiveness

One of the central ways in which measures for programs intended for wide-scale implementation are different than those used in other types of research is that they need to be more comprehensive. This need arises from the complexity of programs that are ready for translation, the multilevel, contextual issues discussed above, and the importance of addressing concerns of multiple stakeholders and decision makers.

This point can be illustrated with a story. Imagine that a specific genetic basis for depression (or cancer or obesity) is discovered next week and that a major pharmacogenetic company rapidly develops a pharmacogenetic intervention in record time and documents its efficacy. Imagine further that the governmental drug approval agency after reviewing the key double-blind randomized controlled trial (RCT) efficacy study that demonstrated a large effect size—a 50% reduction in depression compared to a double-blind placebo control—decides to rush this new drug to market because of the public health need.

This exciting breakthrough would then need to be put into practice to actually impact public health. Here is where the story gets interesting and where the enormous impact of other behavioral, social, economic, and policy factors come into play. Further assume that the government and the pharmaceutical company combine resources in an unprecedented manner to rush the new

Table 2
The Reality of Translating an Evidence-Based Depression Intervention Into Practice

Translation Step	RE-AIM Element[a]	Success Rate	Population-Wide Impact
Persons having genetic risk factor	Population prevalence	40%-60%	40%-60%
Health care settings that participate	Adoption—setting Level	40%-60%	16%-36%
Physicians who prescribe	Adoption—clinician Level	40%-60%	6%-22%
Patients who accept	Reach	40%-60%	2%-13%
Delivery/medication adherence	Implementation(follow regimen correctly)	40%-60%	0.8%-8%
RCT efficacy results	Effectiveness (percentage success in RCT)	40%-60%	0.3%-5%
Continued longer-term effects	Maintenance (individual level)	40%-60%	0.1%-3%

Note: RCT = randomized controlled trial.
a. www.re-aim.org.

drug into widespread use. Table 2 describes what are likely realistic to optimistic estimates of the actual impact of a nationwide dissemination effort to promote use of this breakthrough treatment. The right-hand column of Table 2 shows the bottom-line public health impact or percentage of all depressed persons who would benefit from such an effort.

The left-hand column summarizes the series of steps involved in translating any basic science breakthrough into real-world practice. The second column labels the step according to its categorization in the RE-AIM framework of reach, effectiveness, adoption, implementation, and maintenance (Glasgow & Emmons, 2007a; Glasgow, Lichtenstein, & Marcus, 2003). The third column displays the success rate for that step and includes estimates that vary from 40% to 60% for each stage to bracket the likely overall impact. For most steps, a 40% to 60% success rate would be considered very good for results from a nationwide campaign over a 1- to 2-year period and especially if the 40% to 60% impacted were representative and included those most at risk (which unfortunately is often not the case).

Let’s begin with the assumption that 40% to 60% of the depressed population has the genetic profile that puts them at risk. This would be several times higher than the vast majority of genetic disorders to date, but let’s be optimistic for purposes of illustration. If 40% to 60% of all mental health and primary health care clinics were to adopt this new treatment approach, that would be a phenomenal success. To accomplish this, a convincing case would need to be made to diverse organizations that would include both mental and physical health care settings; government, military and private agencies; outpatient and hospital settings; community health centers; and so on—most of which have their own lengthy approval processes and many of which are underresourced.

The third row in Table 2 illustrates the impact of physician reaction to a newly approved medication and again optimistically assumes that 40% to 60% of physicians would test patients and prescribe this medication to all of their eligible patients. The remaining rows of Table 2 illustrate the impact of later steps in this sequential story of the national rollout of this new depression wonder drug. Only in the fourth and following rows of Table 2 do we even begin to include the impact of patient reactions to such a medication.

Three points should be made in summary: (a) The 40% to 60% estimates for the percentage of patients who would accept and could pay for what would likely be an expensive medication, who would take the medication as prescribed over a sufficient period of time, assuming no major side effects, and who would continue to maintain benefits long term are likely overestimates. (b) Only in the next to last row do the results of the groundbreaking initial study come into play—The issues in all the other rows are typically ignored in the types of efficacy-style RCTs often designed to answer only the narrow question of whether a treatment will work under optimal conditions. (c) Finally, the bottom line impact after 1 to 2 years is that approximately 0.1% to 3% of the depressed population would benefit in a lasting way from this revolutionary breakthrough in pharmacogenetics.

Lessons learned. The purpose of this exercise is not to disparage pharmacogenetic approaches—The same issues apply to real-world applications of behavioral, socioenvironmental, or policy interventions. The point is that evidence needs to expand beyond the narrow domain of studying only the impact on a single primary dependent variable. There is also a more subtle but optimistic message embedded in Table 2.

This message is that there are numerous and multiple opportunities—represented by each row in Table 2—to enhance the ultimate success rate in the bottom right of the table. Improving any of the steps of adoption, reach, implementation, or maintenance could also substantially increase the resulting public health benefit. These various steps also make clear the opportunities for transdisciplinary collaboration to address translation issues—The potential contributions of diverse fields, such as social marketing, health communication, behavioral approaches to adherence, patient–provider communication, risk and decision analysis, health economics, and health policy, are apparent.

A lesson learned, especially when conducting evaluations with limited budgets, is that it is often costly and overly burdensome to collect quantitative measures on all of the issues in Table 1. In addition, validated scales or practical instruments frequently do not exist for the specific research questions in a particular project. In such cases, using a multimethod approach that includes qualitative assessment (Crabtree & Miller, 1999) can help provide a more complete evaluation. Use of qualitative, semistructured interviews are particularly helpful in elucidating reasons for results and understanding factors related to trouble spots in a program (e.g., why certain subgroups choose not to participate, why certain program components are not delivered consistently?).

An excellent example of using quantitative and qualitative assessments together comes from the WISEWOMAN project to reduce cardiovascular risk among low-income women in the United States (Besculides, Zaveri, Farris, & Will, 2006). These investigators first used quantitative measures from the RE-AIM model to evaluate program reach, effectiveness, adoption, implementation, and maintenance (www.re-aim.org). Using these measures, they identified grantee sites that were especially high or low on RE-AIM dimensions and conducted observations, qualitative interviews, and focus groups within these sites to better understand factors associated with success.

Measuring Implementation

Cost. Data on program costs and cost-effectiveness are one of the least frequently reported types of data in research reports (Glasgow et al., 2004). This is unfortunate since program cost is one of the first questions that decision and policy makers ask, and is often a major barrier to dissemination. Part of the reason there have not been more cost analyses is that program developers and researchers have felt overwhelmed by the complexity, magnitude of the task, and the time and costs involved in performing economic analyses. Fortunately, recent, focused approaches are now available that do not attempt to answer every economic issue but restrict focus to the costs of a program as delivered, of replication under different conditions, or the cost per unit change in key outcomes. Such models are practical for most translational purposes (Ritzwoller, Toobert, Sukhanova, & Glasgow, 2006), answer the questions that decision makers usually have, and do not require a great deal of economist time (unlike more complicated issues such as determining cost-benefit or impact on health care utilization).

Customization versus fidelity. One of the current areas of active research and debate is how to resolve the inherent tension when translating research into practice between customizing programs to local situations, to make policies/programs culturally relevant (Castro, Barrera, & Martinez, 2004) using principles of participatory research (Viswanathan et al., 2004), and the need to maintain fidelity to an evidence-based program (Bellg et al., 2004; Bond, 2007). There is agreement that the extremes on either end of this continuum are not good. For example, having users make wholesale modifications to evidence-based interventions without sufficient justification or because they are not experienced with a certain component (e.g., omitting role playing from skills training) would not be recommended. Neither would we expect an underresourced rural mental health clinic that serves a low-income, low-literacy Hispanic population to conduct a program exactly as it was at the Mayo Clinic and to use precisely the same materials.

The two most promising measurement approaches to balancing customization and fidelity seem to be a theoretical principles approach and a more pragmatic essential components assessment. The theoretical (vs. procedural) fidelity approach (Rovniak, Hovell, Wojcik, Winett, & Martinez-Donate, 2005) evaluates program implementation based upon what theoretical principles are addressed by a given component. If a modification retains a similar emphasis on specified, theoretically important principles, then the adaptation is said to have theoretical fidelity.

The essential ingredients approach involves having experienced program developers or a panel of experts with practical experience in the content area and knowledge of the research-based program designate a priori some intervention components as essential or necessary to be consistent with the original program and other components to be modifiable (Ory, Mier, Sharkey, & Anderson, 2007).

Measurement Issues in Practical Trials

There is ongoing controversy about the types of research designs that are most appropriate for translational research. This overall issue is beyond the scope of this article, but a key point is that ''if we want more evidence-based practice, then we need more practice-based evidence'' (Green & Ottosen, 2004, p. 17). Fortunately, there are feasible design alternatives that are often acceptable to stakeholders and can retain internal validity while substantially enhancing external validity. These strategies have been referred to as practical trials (Glasgow, Magid, Beck, Ritzwoller, & Estabrooks, 2005; Tunis, Stryer, & Clancey, 2003). Such designs can be randomized trials or they can use other experimental designs such as interrupted time series or multiple baseline across settings designs that control for threats to internal validity. This section discusses measurement issues related to practical trials.

The distinguishing characteristics of practical trials are that they address four key concerns of decision makers relevant to generalizability. The first issue is inclusion of heterogeneous patients—instead of selecting the most motivated, least complex patients who have the fewest confounding factors and who are maximally homogeneous; samples are purposefully selected (Shadish, Cook, & Campbell, 2002) to represent the range of patients encountered in the real-world settings to which one wants to generalize. The specific measures used to assess representativeness should reflect what is known in that particular research area but will frequently include factors such as age, gender, race, ethnicity, and health literacy that have been associated with inequities.

A second characteristic is that the interventions are conducted in multiple settings. The emphasis is on including a range of settings that reflect those in typical practice—in contrast to only the settings that have the greatest expertise, the most resources, and the highest chances of successfully delivering an intervention. Measurement issues concerning settings include comparing key characteristics of participating settings related to intervention capacity (e.g., number and expertise of staff, level of resources) to either (a) settings invited to participate that decline or to (b) all organizations of that type in the region(s) the study is conducted.

The third factor is one of the most significant ways in which practical clinical (Tunis et al., 2003) and behavioral trials (Glasgow, Davidson, Dobkin, Ockene, & Spring, 2006a) differ from research as usual. This criterion is that comparison conditions represent current standards of care or alternative treatments—rather than no treatment or placebo controls. The rationale for this is that to justify changes in practice, the additional education and quality control modifications necessary, and the frequently higher costs of a new treatment, the innovation should be significantly better than current, familiar, and less-expensive interventions.

The final characteristic of practical trials reflects back to our story and is that multiple outcomes, and especially outcomes relevant to clinicians, decision makers, and the community, should be included. These concerns address factors such as staff and implementation requirements, costs, range of applicability, and impact on quality of life or benefit relative to alternative uses of resources. In summary, measurement for practical trials provides important information on the influence of contextual factors and generalizability that is often missing from traditional efficacy studies.

Challenges and Conclusions

Table 3 summarizes several key challenges and related assessment strategies associated with measurement in translational research. It uses the RE-AIM model (Glasgow, 2008; Glasgow, McKay, Piette, & Reynolds, 2001; www.re-aim.org) to consider both common challenges and possible solutions. RE-AIM is a framework for translational research that focuses attention on key issues and related measures to assist at each of several essential steps involved in integrating research into practice.

The chief challenge to assessing reach is that too often evaluations include only participants who are easy to access, most likely to benefit, or especially motivated, and thus recruitment expectations for translation are unrealistically high. Another danger is of casting too narrow of a net in evaluating results (effectiveness), focusing only on restricted outcomes, and omitting measures of possible negative effects, mediating variables and process measures that can help to understand why and how program/policy effects (or lack of effects) occur. Table 3 presents ways to broaden this perspective. Decision makers are concerned about impact on participants like those in their setting. Moderator analyses, or evaluations of whether a program is differentially effective across subgroups that differ on important psychosocial and demographic factors, can help to clarify applicability (Glasgow et al., 2006b).

More research should be conducted in representative or low-resource settings. Equal priority should be given to the recruitment and representativeness of intervention settings (adoption)—for example, health care clinics, worksites—as is given to the representativeness of individual participants. Two crucial implementation issues often present challenges to interpretation and research translation. The first is the failure to identify the characteristics of settings and staff who are able to successfully implement programs or policies. Staff characteristics

Table 3
Common Challenges in Evaluating Translational Research and Possible Remedies

Challenge	Remedy
Reach	
Not including a relevant, high risk, or representative sample or being able to evaluate representativeness	Use population-based recruitment or overrecruit high-risk subgroups Report on participation rate, exclusions, and representativeness Avoid too many exclusion criteria
Effectiveness	
Not thoroughly understanding outcomes or how they come about	Assess broad set of outcomes, including possible negative ones
No knowledge of mediators	Include measures of hypothesized mediators
No assessment of moderator variables	Conduct subgroup analyses to identify moderator effects
Conflicting or ambiguous results	
Inadequate control conditions to rule out alternative hypotheses	Design control condition to fit your question
Adoption	
Program only studied in high functioning, optimal settings	Involve potential adoptee using CBPR principles beginning with initial design phase
Program not ever adopted or endorsed—or only used in academic settings	Approach a representative or broad group of settings early on when revision is still possible and report on setting exclusions, participation, and representativeness
Implementation	
Protocols not delivered as intended (Type III error)	Assess if treatment is too complicated, too intensive, or not compatible with other duties to be delivered consistently
Not able to answer key questions about costs, time, or staff requirements	Systematically vary staff characteristics and evaluate staff impact as well as costs
Deciding if a program adaptation or customization is good or bad	Specify a priori the critical theoretical components Identify essential elements that cannot be changed and those that can be adapted
Maintenance	
Program or effects not maintained over time	Include maintenance phase in both protocol and in evaluation plan
Substantial attrition of settings, delivery staff, and/or participants over time	Plan for institutionalization, sustainability, and dissemination and their evaluation Take steps to minimize attrition, address attrition using appropriate methods, evaluate, and report impact of attrition

Note: CBPR = community-based participatory research.

that may moderate implementation include expertise, education, training, age, race/ethnicity, gender, experience, and similarity to the target audience. The second issue is that estimates of the program costs are often not available. It is now feasible for most programs to collect valid estimates of program implementation costs (Ritzwoller et al., 2006) so that cost-effectiveness analyses and return on investments can be calculated and made available to decision makers.

Resolving the tension between fidelity (delivering a program exactly as in a research protocol) and customization or adaptation to local settings, culture, and history are among the most important measurement challenges. As discussed above, recommended approaches include specifying key or critical components of a program and evaluating delivery of the theoretical principles or mechanisms that are hypothesized to lead to desired outcomes. Logic models (Glasgow & Linnan, 2007b) are useful for depicting predicted relationships and in guiding measurement decisions.

There is a dearth of information on maintenance or sustainability of programs at the setting level. We need much greater understanding of the extent to which settings continue implementation, make adaptations, or discontinue interventions over time (Goodman, McLeroy, Steckler, & Hoyle, 1993). At the individual level, participant attrition is a common challenge. Recommendations for addressing the impact of attrition include analyzing the characteristics of those present (vs. those who drop out) at follow-up assessments and then deciding which imputation methods are most appropriate for that particular missing data situation.

The key to successfully overcoming the challenges summarized in Table 3 is to plan for and anticipate trouble spots (Green & Kreuter, 2005; Klesges, Estabrooks, Glasgow, & Dzewaltowski, 2005). These issues can be addressed using RE-AIM, PRECEDE-PROCEED, or other planning and evaluation frameworks. The world is complex and program effects are often context dependent. Our evaluations should reflect this complexity, and reports should transparently describe program challenges, adaptations, and contextual issues so that both internal and external validity concerns are addressed.

This article has summarized key measurement approaches, challenges, and lessons learned in assessing programs and policies intended for broader translation across a variety of content areas. Assessment costs, participant burden, and other trade-offs must be considered when planning formative, outcome, impact, and process evaluation efforts in a complex world (Glasgow & Linnan, 2007b; Linnan & Steckler, 2002). Some of the interventions and policies we assess will prove effective and should be considered for sustainability and translation; others will not. Data collected to answer the questions above will reveal program effects, limitations, processes, and pathways of change, as well as insights about how to improve the theory guiding a policy or program. Such broad-based measurement approaches should help lead to generalizable improvements in programs, policy, and theory.

References

Baranowski, T., Lin, L. S., Wetter, D. W., Resnicow, K., & Davis, H. M. (1997). Theory as mediating variables: Why aren't community interventions working as desired? *Annals of Epidemiology*, *7*, S89-S95.

Basch, C. E., Sliepcevich, E. M., & Gold, R. S. (1985). Avoiding type III errors in health education program evaluations. *Health Education Quarterly*, *12*, 315-331.

Bellg, A. J., Borrelli, B., Resnick, B., Ogedegbe, G., Hecht, J., Ernst, D., et al. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the Behavior Change Consortium. *Health Psychology*, *23*, 443-451.

Berwick, D. M. (1996). A primer on leading the improvement of systems. *British Medical Journal*, *312*, 619-622.

Besculides, M., Zaveri, H., Farris, R., & Will, J. (2006). Identifying best practices for WISEWOMAN programs using a mixed-methods evaluation. *Preventing Chronic Disease*, *3*, 1-9.

Bond, G. R. (2007). Modest implementation efforts, modest fidelity, and modest outcomes. *Psychiatric Services*, *58*, 334.

Castro, F. G., Barrera, M., Jr., & Martinez, C. R., Jr. (2004). The cultural adaptation of prevention interventions: Resolving tensions between fidelity and fit. *Prevention Science*, *5*, 41-45.

Crabtree, B. F., & Miller, W. L. (1999). Depth interviewing. In B. F. Crabtree & W. L. Miller (Eds.), *Doing qualitative research* (pp. 89-107). Thousand Oaks, CA: Sage.

Glasgow, R. E. (2008). What types of evidence are most needed to advance behavioral medicine? *Annals of Behavioral Medicine*, *35*, 19-25.

Glasgow, R. E., Davidson, K. W., Dobkin, P. L., Ockene, J., & Spring, B. (2006a). Practical behavioral trials to advance evidence-based behavioral medicine. *Annals of Behavioral Medicine*, *31*, 5-13.

Glasgow, R. E., & Emmons, K. M. (2007a). How can we increase translation of research into practice? *Annual Review of Public Health*, *28*, 413-433.

Glasgow, R. E., Klesges, L. M., Dzewaltowski, D. A., Bull, S. S., & Estabrooks, P. (2004). The future of health behavior change research: What is needed to improve translation of research into health promotion practice? *Annals of Behavioral Medicine*, *27*, 3-12.

Glasgow, R. E., Lichtenstein, E., & Marcus, A. C. (2003). Why don't we see more translation of health promotion research to practice? Rethinking the efficacy to effectiveness transition. *American Journal of Public Health*, *93*, 1261-1267.

Glasgow, R. E., & Linnan, L. (2007b). Evaluation of theory-based interventions. In K. Glanz (Ed.), *Health education: Theory, research and practice* (4th ed. pp. 487-506). Hoboken, NJ: Jossey-Bass.

Glasgow, R. E., Magid, D. J., Beck, A., Ritzwoller, D., & Estabrooks, P. A. (2005). Practical clinical trials for translating research to practice: Design and measurement recommendations. *Medical Care*, *43*, 551-557.

Glasgow, R. E., McKay, H. G., Piette, J. D., & Reynolds, K. D. (2001). The RE-AIM framework for evaluating interventions: What can it tell us about approaches to chronic illness management? *Patient Education and Counseling*, *44*, 119-127.

Glasgow, R. E., Strycker, L. A., King, D., Toobert, D., Kulchak Rahm, A., Jex, M., et al. (2006b). Robustness of a computer-assisted diabetes self-management intervention across patient characteristics, healthcare settings, and intervention staff. *American Journal of Managed Care*, *12*, 137-145.

Goodman, R. M., McLeroy, K. R., Steckler, A., & Hoyle, R. (1993). Development of level of institutionalization scales for health promotion programs. *Health Education Quarterly*, *20*, 161-178.

Green, L. W., & Kreuter, M. W. (2005). *Health promotion planning: An educational and ecological approach* (4th ed.). Mountain View, CA: Mayfield.

Green, L. W., & Ottosen, J. M. (2004, January 12-13). *From efficacy to effectiveness to community and back: Evidence-based practice vs. practice-based evidence*. Proceedings from conference From Clinical Trials to Community: The Science of Translating Diabetes and Obesity Research, National Institutes of Diabetes, Digestive and Kidney Diseases, Bethesda, MD.

Institute of Medicine & Committee on Quality Health Care in America. (2003). *Crossing the quality chasm: A new health system for the 21st century.* Washington, DC: National Academies Press.

Kaplan, R. M. (2003). The significance of quality of life in health care. *Quality of Life Research*, *12*, 3-16.

Kingdon, J. (1995). *Agendas, alternatives, and public policy* (2nd ed.). New York: Harper Collins.

Klesges, L. M., Estabrooks, P. A., Glasgow, R. E., & Dzewaltowski, D. (2005). Beginning with the application in mind: Designing and planning health behavior change interventions to enhance dissemination. *Annals of Behavioral Medicine*, *29*, 66S-75S.

Linnan, L., & Steckler, A. (2002). Process evaluation and public health interventions: An overview. In A. Steckler & L. Linnan (Eds.), *Process evaluation in public health interventions and research* (pp. 1-23). San Francisco: Jossey-Bass.

McGlynn, E. A., Asch, S. M., Adams, J., Keesey, J., Hicks, J., DeCristofaro, A., et al. (2003). The quality of health care delivered to adults in the United States. *New England Journal of Medicine*, *348*, 2635-2645.

Orleans, C. T. (2000). Promoting the maintenance of health behavior change: Recommendations for the next generation of research and practice. *Health Psychology*, *19*, 76-83.

Ory, M. G., Mier, N., Sharkey, J. R., & Anderson, L. A. (2007). Translating science into public health practice: Lessons from physical activity interventions. *Alzheimer's and Dementia*, *3*, S57.

Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. (2005). Realist review: A new method of systematic review designed for complex policy interventions. *Journal of Health Services Research & Policy*, *10*, S21-S39.

Reynolds, K. D., Buller, D. B., Yaroch, A. L., Maloy, J. A., & Cutter, G. R. (2006). Mediation of a middle school skin cancer prevention program. *Health Psychology*, *25*, 616-625.

Ritzwoller, D. P., Toobert, D., Sukhanova, A., & Glasgow, R. E. (2006). Economic analysis of the Mediterranean Lifestyle Program for postmenopausal women with diabetes. *Diabetes Educator*, *32*, 761-769.

Rotheram-Borus, M. J., Flannery, D., & Duan, N. (2004). Interventions that are CURRES: Cost-effective, useful, realistic, robust, evolving, and sustainable. In H. Rehmschmidt, M. L. Belfer & I. Goodyer (Eds.), *Facilitating pathways: Care, treatment, and prevention in child and adolescent health* (pp. 235-244). New York: Springer.

Rovniak, L. S., Hovell, M. F., Wojcik, J. R., Winett, R. A., & Martinez-Donate, A. P. (2005). Enhancing theoretical fidelity: An e-mail-based walking program demonstration. *American Journal of Health Promotion*, *20*, 85-95.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference.* Boston: Houghton Mifflin.

Tunis, S. R., Stryer, D. B., & Clancey, C. M. (2003). Practical clinical trials. Increasing the value of clinical research for decision making in clinical and health policy. *Journal of the American Medical Association*, *290*, 1624-1632.

Viswanathan, M., Ammerman, A., Eng, E., Gartlehner, G., Lohr, K. N., Griffith, D., et al. (2004). *Community-based participatory research: Assessing the evidence* (Evidence Report/Technology Assessment No. 99; Rep. No. AHRQ Pub No. 04-E022-2). Rockville, MD: Agency for Healthcare Research and Quality.