Behavior change intervention research in community settings: how generalizable are the results?

DAVID A. DZEWALTOWSKI1, PAUL A. ESTABROOKS1,2, LISA M. KLESGES3, SHEANA BULL4 and RUSSELL E. GLASGOW2
1Community Health Institute, Kansas State University, Manhattan, KS, USA, 2Kaiser Permanente Colorado, Denver, CO, USA, 3University of Tennessee Health Science Center, Memphis, TN, USA and 4University of Colorado Health Sciences Center, Denver, CO, USA

SUMMARY
This review examines the extent to which recent behavioral intervention studies conducted in community settings reported on elements of internal and external validity, with an emphasis on whether research has been conducted in representative settings with representative populations. A targeted review was conducted on community-based intervention studies that promoted good nutrition, physical activity or smoking cessation/prevention, and were published in 11 leading health behavior journals between 1996 and 2000. The RE-AIM framework (reach, efficacy, adoption, implementation and maintenance) was used to evaluate the extent to which each paper reported on elements of reach, efficacy/effectiveness, adoption, implementation and maintenance. A total of 27 publications were reviewed. Although most studies (88%) reported participation rates among eligible members of the target audience (‘reach’), only 11% of studies reported the participation rate (‘adoption’) among eligible community-based organizations or settings. Few studies reported if participating individuals or settings were representative of those found in the broader population. Although a majority of studies (59%) reported whether the intervention was delivered (‘implementation’), few reported whether individuals maintained behavior change (30%) or whether organizations maintained or institutionalized interventions (0%). To increase the potential to translate community research findings to practice, studies should place a greater emphasis on obtaining and reporting external validity information, such as representativeness. The lack of external validity information limits researchers’ and practitioners’ ability to judge the generalizability of effects and the comparative utility of interventions. Improved reporting will facilitate implementation of proven and broadly applicable intervention strategies in communities. To make significant progress, all parties, including researchers, reviewers, editors and funders, need to take responsibility for increased emphasis on external validity information and ask what role they can best play to facilitate this process.

Key words: behavioral interventions, external validity, RE-AIM

INTRODUCTION
There is a well recognized gap between research findings and the implementation of evidence-based prevention strategies in community settings (McGinnis and Foege, 2000). Research should inform community leaders or facilitate using proven intervention strategies in community environments. However, community leaders and health promotion experts suggest that a barrier to the adoption of research-based, efficacious interventions is that these strategies may not meet community needs (Green and Mercer, 2001). Interventions may be too complex, difficult or costly to integrate with existing activities.

Part of the problem may be researchers’ attempts to find the most efficacious program rather than a program that could be implemented and delivered with limited resources to many people. A lack of reporting on critical external validity issues might contribute to the problem. Recent reviews of behavioral interventions
delivered through schools (Estabrooks et al., 2002), worksites (Bull et al., 2003) and health care settings (Glasgow et al., 2002) demonstrated a lack of reporting about external validity issues. We will address the research–practice gap by examining how often internal and external validity issues are addressed in the community-based health behavior research literature.

In this article, we evaluate the extent to which recent studies have included descriptions of internal and external validity, with an emphasis on whether research has reported the representativeness of community-based settings and populations in community interventions targeting dietary change, physical activity or smoking. The 2010 Healthy People report defines community as a specific group of people, often living in a defined geographical area, who share a common culture, values and norms, and who are arranged in a social structure according to relationships the community has developed over a period of time (World Health Organization, 1998; US Department of Health and Human Services, 2000). This definition suggests that community-based strategies and initiatives are implemented in the major social structures that provide channels to reach and influence defined populations (Mullen et al., 1995). Community settings include public facilities, local government agencies, social service, faith and civic organizations (US Department of Health and Human Services, 2001). We also included interventions aimed at the entire population by using community settings and/or media to target behavior change in homes. We excluded interventions that were delivered through schools, worksites or health care settings, as this work has been reported elsewhere (Bull et al., 2003; Estabrooks et al., 2002; Glasgow et al., 2002).

Our review is structured around the RE-AIM framework [reach, efficacy/effectiveness (depending on research goal), adoption, implementation and maintenance; http://www.re-aim.org (Glasgow et al., 1999; Glasgow et al., 2001)]. This framework was designed to organize and integrate internal and external validity dimensions that are important in the translation of research to practice for health promotion interventions (Rogers, 1995; Abrams et al., 1996; Oldenburg et al., 1999; Oldenburg et al., 2000).

Reach is defined as the percent of potentially eligible individuals who participate in the intervention study, and how representative they are of the target population from which they are drawn. Efficacy/effectiveness is the intended positive impact of the intervention and its possible unintended consequences on quality of life and related factors. Reach and efficacy/effectiveness operate at the individual level.

Adoption is the percent of potential settings and intervention agents that participate in a study and how representative they are of targeted settings/agents. Implementation refers to the quantity and quality of delivery of the intervention’s various components. Adoption and implementation are setting-level dimensions.

Finally, the maintenance dimension includes individual- and setting-level indices. At the individual level, maintenance is defined as the longer-term efficacy/effectiveness of an intervention. Outcomes at ≥6 months post-intervention contact reflect longer-term individual maintenance. The setting level definition of maintenance refers to the institutionalization of a program and is assessed according to the percent of settings that continue the intervention program, in part or in whole, beyond the study duration (Glasgow et al., 1999; Glasgow et al., 2001).

**METHODS**

This project was carried out in association with the Behavior Change Consortium [Behavior Change Consortium: National Institutes of Health (NIH), 2001 (http://www1.od.nih.gov/behaviorchange)], a collaboration among 15 NIH-funded research projects studying theory-based health behavior change around diet, physical activity and smoking cessation. Hence, our review is focused on interventions that target these behaviors.

The review examined the extent to which recent community-based studies reported in leading health promotion journals addressed RE-AIM dimensions. We reviewed studies meeting these criteria published from 1996 to 2000 in the nine journals listed in Table 1. Journals were selected based on feedback from a panel of leading health promotion researchers and based on the following criteria: circulation, frequency of reporting intervention studies, and coverage of general health promotion. Specialty journals and those restricted to a single discipline or target behavior were excluded. We included experimental or quasi-experimental studies that used some type of comparison or control condition and which were designed to influence the behavioral outcomes of
diet (e.g. caloric intake, fruit and vegetable consumption, percent of calories from fat), physical activity, and smoking cessation. Studies reporting pre-/post-test results of single interventions were not included.

**Coding criteria**

Articles were coded using the RE-AIM criteria. Under ‘reach’, we coded: (i) whether a study reported the percent of eligible participants who were enrolled in the study; (ii) whether analyses were conducted that compared the eligible participants who did not participate with those who participated or to the greater community population (representativeness); and (iii) whether the study reported the percent of potentially eligible participants excluded from participation based on researcher criteria. We also categorized studies by the method of subject recruitment and coded participation rates. Due to the selection criteria, all studies reported efficacy/effectiveness information. The inclusion of quality of life or other measures that could reveal potentially negative consequences of the intervention was also coded.

Adoption was coded by recording if a study reported: (i) the number or percent of potential study sites that investigators excluded from consideration; (ii) the percent of eligible sites that participated; and (iii) the characteristics of sites that participated compared with sites that declined or with the greater population of sites (representativeness). We also coded if the percent of potential intervention agents (e.g. leaders within sites) who agreed to deliver the intervention was reported, and if analyses were conducted of the characteristics of potential intervention agents who took part in program delivery versus those who did not.

For implementation, we coded: (i) whether any information was reported that could describe the extent to which the intervention was delivered to participants; (ii) whether a delivery rate was presented based on the extent to which the intervention was implemented as intended; and (iii) if the requirements of the intervention were described in terms of time or cost. For maintenance, we coded whether a study included ≥6 months post-intervention follow-up data on participants. If follow-up information was available, we coded if: (i) the characteristics of participants versus dropouts were reported; (ii) either imputation or intent-to-treat analyses were used to evaluate the impact of attrition on study outcomes; and (iii) a report was given of whether the intervention was continued at the setting after the research study was completed.

**Reliability of coding**

We randomly selected 10 papers so that we could examine coding reliability. We report the agreement indices on whether or not a study included information on the relevant RE-AIM component. Percent agreement scores were uniformly high across the various RE-AIM dimensions. Agreement for reach ranged from 80% for whether the participation rate was reported, to 100% for whether both the exclusion and representativeness elements were reported. Agreement on adoption factors ranged from 90% for whether the number of sites that investigators excluded from consideration was reported, to 100% for other adoption components. Implementation dimensions ranged from 90% for whether there was information on delivery and
delivery rates, to 100% of agreement if the paper reported the time or cost of implementation. Finally, for maintenance there was 80% agreement of whether imputation or intent-to-treat analyses were used in evaluating the impact of attrition, and 100% agreement on the other maintenance components. The Kappa statistic (Cohen, 1968), which adjusts for chance agreement across all coding categories, was in the excellent range at 0.87.

RESULTS

Description of the studies

We located 27 studies that met our inclusion criteria (Table 1). These studies were distributed fairly evenly across the various health behaviors, and only three studies combined more than two of these health behaviors. Six studies targeted the whole community, nine studies targeted community sites (such as fitness centers or churches) and 12 studies targeted homes. The home-based studies all had a media component as a primary intervention strategy. Some of the home studies added limited face-to-face contact to the media strategy. The Appendix includes citations for articles meeting our inclusion criteria.

Seventy-four percent (n = 20) of the 27 studies were randomized controlled trials (RCTs). We report results by both intervention site and overall because all community-wide studies used quasi-experimental designs (n = 6). Eighty-nine percent of the site-based studies and 100% of the home-based studies used RCT designs. Only four of the 27 studies used cluster randomization, where settings such as supermarkets were randomized to conditions. Only one study reported both positive and negative outcomes, such as a quality of life index or other indices of possible adverse outcomes.

Reach

Most studies (88%) specified the number of eligible individuals who participated in the study (Table 2). Figure 1 shows that studies that recruited by advertising for volunteers had higher participation rates compared with studies that recruited from a targeted list of individuals and studies that recruited participants primarily through random digit dial telephone lists. These differences may be due to selection bias from a study’s definition of eligible participants rather than the recruitment method. That is, random digit dial recruitment or site-based studies could begin with a list of eligible individuals that is representative of a targeted population, whereas mass media advertising might begin with a list of volunteers that was already a self-selected sample. Only 11% of the studies reported the percent of potential participants that investigators excluded and information on representativeness. Therefore, we cannot determine if the method of recruitment actually influenced how representative the samples were on key sociodemographic characteristics. However, the low level of individual participation rates in studies that recruited from a representative targeted population raises questions about generalizability.

Hafstad and associates (Hafstad et al., 1997) provide an excellent example of how Reach can be reported in a study of an anti-smoking mass media campaign. Although the campaign was delivered in the community at a county level, they recruited from all students aged 14 and 15 years in lower grades of a secondary school. Because recruitment occurred through an organizational setting, even though the intervention was delivered to the community at large, the research team was able to define a representative sample (i.e. all students in the county) and obtain parental consent for 95% of the eligible students.

Adoption

Few studies reported information on the number and types of settings that agreed to participate compared to those that were recruited (Table 2). Only two home-based studies and one site-based study reported the potential number of sites that might participate compared with those that actually did participate. No studies reported the potential number of intervention agents within sites who agreed to participate. Two site-based studies reported the number of excluded organizations and if these settings were representative. Voorhees et al. (Voorhees et al., 1996), in their church-based smoking cessation study, defined a catchment area for recruitment of settings (21 census tracts with 71 291 persons). Within the catchment area, 130 churches were identified and 23 were invited to participate. The study Steering Committee recruited a convenience sample of churches that were believed to be active in social and health issues, and for which the committee...
members knew the pastors. Of the 23 churches invited, all but one participated.

**Implementation**

Many studies (59%; \( n = 16 \)) provided information to evaluate the extent to which the intervention was delivered to participants. Considerable variation existed across studies in how program implementation was defined. Eleven of the 16 studies (70%) reporting implementation provided estimated percentages of consistent delivery of the intervention according to a protocol. The mean implementation rate was

<table>
<thead>
<tr>
<th>Component</th>
<th>Community intervention site</th>
<th>Mass community</th>
<th>Home-based</th>
<th>Site-based</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reach</td>
<td>50</td>
<td>100</td>
<td>100</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>Participation rate of individuals</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Representativeness of individuals</td>
<td>0</td>
<td>17</td>
<td>11</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Investigator exclusions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effectiveness</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Behavioral outcome measure</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Quality of life or negative outcomes</td>
<td>0</td>
<td>0</td>
<td>22</td>
<td>7.4</td>
<td></td>
</tr>
<tr>
<td>Adoption</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participation rate of settings</td>
<td>0</td>
<td>16</td>
<td>11</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Representativeness of settings</td>
<td>0</td>
<td>0</td>
<td>22</td>
<td>7.4</td>
<td></td>
</tr>
<tr>
<td>Investigator exclusions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implementation</td>
<td>50</td>
<td>58</td>
<td>67</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>Delivery of intervention information</td>
<td>83</td>
<td>67</td>
<td>44</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>Delivery rates</td>
<td>0</td>
<td>58</td>
<td>44</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>Maintenance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Individual level: 6 month post-intervention</td>
<td>33</td>
<td>33</td>
<td>25</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>If so, attrition numbers (or analysis of completers and non-completers)</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>If so, use of imputation or intent to treat analyses</td>
<td>0</td>
<td>50</td>
<td>100</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>Site level: institutionalization</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 1:** Mean participation rate results of reach dimension by recruitment strategy.
85% (range: 19–100%). However, some measures of implementation may have confused intervention delivery (ability to deliver intervention as intended) with intervention receipt (degree to which the target audience actually received the intervention) (Burgio et al., 2001). Because implementation elements are not uniformly reported, the average implementation rate should be interpreted with caution. It is unclear whether those not including implementation information had poor, average or excellent intervention delivery, so their effect on the average rates is unknown. Other implementation elements, such as time or cost related to intervention delivery, were also inconsistently reported. Mass media interventions generally reported some information regarding the time of intervention (83% of mass community studies), but across all of the studies, only 63% provided time or cost information.

More work is needed to define valid and reliable approaches to measure program implementation consistently across studies (McGraw et al., 2000). For example, a community development study targeting smoking cessation measured if residents had ‘heard of the organization that was delivering the initiative’ (Fisher et al., 1998), which could be considered a measure of intervention receipt. However, this does not describe the extent to which the program was delivered. Implementation could be operationalized in terms of both the extent to which an intervention was delivered as intended and the degree to which the target audience actually received the intervention (Burgio et al., 2001). In studies with a media component, implementation delivery is often assumed to be 100%. For example, an intervention is deemed delivered as intended when the mail is sent for a mass mailing campaign. However, documenting delivery does not provide information about whether or not the targeted participants received and understood the intervention. Therefore, some investigators have attempted to measure implementation as the participants’ receipt of the intervention. Brug and colleagues (Brug et al., 1998) surveyed participants and found that 99% of individuals receiving a tailored letter indicated that they had received and read the letter. Furthermore, Resnicow and colleagues (Resnicow et al., 1997) reported the number of booster telephone calls that were received, which is a shared implementation process of organizational delivery and target audience behavior (answering the telephone).

Maintenance

Community-based studies do not routinely examine the long-term effects of interventions. We did not analyze four of the 27 studies on the individual maintenance level because their research designs did not include individual-level longitudinal data. These studies were primarily multiple year studies that reported group-level data. Of the 23 remaining studies, only 30% (n = 7) reported follow-up data after the intervention concluded. The mean rate of attrition was 20%, and ranged from 4% to 35%. All studies that reported follow-up data provided information on attrition, such as a comparison of baseline data of dropouts to participants. However, only 57% of those reporting attrition used this information in their analytic procedures to adjust their analyses using imputation or intent-to-treat methods. No study reported the extent to which the intervention was continued beyond the formal study period.

DISCUSSION

This review examined the extent to which behavioral intervention studies reported on elements of internal and external validity, with an emphasis on whether research has been conducted in representative community-based settings with representative populations. Overall, studies have not reported sufficient information to judge the representativeness of community-based settings and populations. In contrast, most studies have documented individual level participation rates and efficacy/effectiveness outcomes.

The RE-AIM reporting criteria were developed to place a balanced emphasis on internal and external validity by addressing five components of interventions important for translation of research findings (Glasgow et al., 1999; Glasgow et al., 2001). Internal validity is central to establishing confidence that a variation in an effect is due to the intervention and not some other causal force(s) or ‘nuisance’ factor(s) (Cook and Campbell, 1979; Brewer, 2000). All selected studies reported internal validity elements of the RE-AIM criteria, including effectiveness/efficacy results based on controlled or randomized controlled design testing an intervention’s influence on a behavioral outcome measure. A majority of the studies reported at
least some information on the level of the intervention’s implementation. In contrast, few studies reported external validity elements of the RE-AIM criteria. Although most studies reported on the number of individuals who participated compared with those who were eligible, the participants’ representation of a more general population was not described. Higher participation rates were found in studies that selected volunteers compared with studies that invited participants from a list of individuals that represented a targeted population. Because no journals require a standardized reporting of Reach information, we do not know if participants in research studies are representative of community populations targeted for health behavior change interventions or to what extent selection bias occurs.

Multi-level analyses and interventions have recently been called for in health research (Anderson, 1998; Berkman and Kawachi, 2000; McKinlay and Marceau, 2000; US Department of Health and Human Services, 2001) to improve our understanding of contextual factors related to behavior and behavior change. Despite these appeals, our literature review shows a lack of emphasis on these elements. Only 11% of studies reported organizational-level information compared with 88% that reported individual participant-level information. It is unclear whether researchers do not collect such organizational-level information or whether disincentives exist to preclude its being reported. Understanding organizational characteristics related to the adoption of programs is critical to our understanding the dissemination and effectiveness of community-based behavioral interventions as called for by policy makers (US Department of Health Services, 2000).

Lack of reported maintenance information at both the individual and site level in these community-based research studies is problematic because it limits the estimation of the long-term impact of such interventions. This lack may indicate the difficulty in conducting research in community settings, which in turn influences reporting at both the organizational and individual level. It may also reflect the unavailability of funding to conduct post-intervention assessments, and methods to increase the support of these activities require attention. Otherwise, researchers conducting efficacy studies may view studying maintenance or dissemination as beyond their purview.

Many studies in this review addressed dimensions of internal validity as indicated by reports of short-term implementation. However, implementation may also be important for assessing external validity. That is, methods to increase short-term implementation goals, such as the use of research staff to deliver the intervention, may be less effective than methods designed for longer-term institutionalization of programs. For example, participatory research methods (Green and Mercer, 2001) may lack high implementation rates initially and may not emphasize high internal validity. However, participatory methods may ensure an increased emphasis on external validity and the institutionalization of programs over time. Studies would need to report both implementation information and the long-term institutionalization of programs to assess this proposition.

Recommendations

It is clear that recent community-based studies in these 11 leading journals infrequently address RE-AIM dimensions related to external validity. Therefore, our recommendations center on the need to increase the frequency of reported elements of external validity, specifically representativeness of samples and settings, and information on long-term maintenance of intervention programs. Appealing to researchers to adopt these recommendations is insufficient. Funders, journal editors and reviewers must also consider and support the inclusion of external validity dimensions; therefore we recommend ways to facilitate reporting on the translation of research to practice. First, we recommend that a comparison of the study sample with either the broader target population or with those that decline, with respect to basic demographic data, should be completed (Reach). This comparison can often be made using available datasets (e.g. census data). Where such a dataset is unavailable, researchers should attempt to gather basic demographic data on all participants contacted for recruitment and subsequently compare those that agreed to participate with those that declined. Secondly, we recommend that a comparison of the setting and/or delivery agent (e.g. community leader) be made with the broader target population of settings and/or delivery agents on the resources available or agent education-level (adoption). For example,
community-based interventions should compare organizational structure, available tax base and facilities for physical activity of intervention sites with the target population of settings. In terms of intervention delivery agents, the required level of education and expertise should be compared with that of individuals who deliver the intervention in effectiveness trials. Thirdly, we
Community settings

recommend that researchers record the level of fidelity with which the intervention is delivered (implementation). This evaluation should include how much of the intervention protocol was followed as intended, the timeliness of protocol implementation, and any adaptations of the intervention protocol. We also recommend that researchers do not solely rely on self-report data to document implementation (Davis et al., 2000).

To facilitate our recommendations it is useful to identify a standardized way to report on these issues. A reporting template such as that in Figure 2 might be one efficient reporting method. This template adds external validity information to what many health journals currently require concerning recruitment and retention (Moher et al., 2001). By making explicit participation and implementation results at the organizational (as well as individual) level, the public health impact becomes more clear. Such a template could also be helpful in designing intervention programs with great potential for being adopted widely.

Caveats
This review did not attempt an exhaustive search of the extant literature, but sought to characterize recently published research in selected journals. Because our sampling may not represent current research practices, the estimates of reported RE-AIM dimensions may not reflect the larger community intervention literature. However, we speculate that our reporting rates exceed the actual rate of reporting external validity in earlier issues and in less demanding journals. Likewise, the lack of reporting does not mean that the information was not collected. The need to preserve journal space may lead to external validity information being deleted, which does not preclude researchers’ awareness of the issue.

Researchers may contend that their studies were intended to demonstrate the efficacy of the intervention, not its effectiveness or generalizability. Flay (Flay, 1986) defines an efficacy trial as a well controlled test of an effect under ideal conditions, which is compared with an effectiveness trial that studies the strength of an intervention effect under real-world conditions. Therefore, the reporting of external validity information may be of greater importance when conducting an effectiveness trial. However, it is more likely that the generalizability of an effect will be established by replicating studies using systematically sampled settings and participants. To assimilate research findings across settings, both efficacy and effectiveness studies must provide representativeness information for settings and participants.

CONCLUSIONS
To address the gap between research findings and implementation of prevention strategies in community settings (McGinnis and Foege, 2000), improvements in reporting on elements of external validity are required. Among those most needed are methods to improve how studies are conducted, how representativeness of samples and settings are reported, and whether programs are maintained in the long term. The lack of such reporting limits researchers’ and practitioners’ ability to judge the generalizability of effects and the comparative utility of interventions. Improved reporting will facilitate implementation of proven and broadly applicable intervention strategies in communities.

ACKNOWLEDGEMENTS
This research was supported by a grant from the Robert Wood Johnson Foundation. The first two authors’ participation in the Behavior Change Consortium was supported by the National Institutes of Health (NICHD/NINR/ODP/NIAID/ODS-RO1 HD37367).

Address for correspondence:
David A. Dzewaltowski
Community Health Institute
Umburger 101
Kansas State University
Manhattan
KS 66506
USA
E-mail: dadx@ksu.edu

REFERENCES


APPENDIX


