POLICY
Appropriate Research Designs for Evaluating Community-level Alcohol Interventions: What Next?

Conor Gilligan1,* , Rob Sanson-Fisher1 and Anthony Shakeshaft2

1Priority Research Centre for Health Behaviour, School of Medicine and Public health, University of Newcastle, Newcastle, Australia and 2The Sax Institute, Sydney and the National Drug and Alcohol Research Centre, University of NSW, Sydney, Australia

*Corresponding author: Discipline of Health Behaviour Sciences, University of Newcastle, David Maddison Building, Cnr King & Watt Sts, Newcastle, Australia. Tel: +61-2-4913-8635; Fax: +61-2-4913-8148; E-mail: Conor.Gilligan@newcastle.edu.au

(Received 3 March 2010; accepted 10 June 2010)

Abstract — Aim: The aim of this study was to explore the logistical and methodological strengths and weaknesses of some of the more common research designs which can be used to evaluate the impact of system- or population-level approaches for reducing alcohol-related harms. Method: This paper identifies studies that have evaluated system or population approaches to reduce alcohol-related harms. It highlights the tension caused by a desire for the most rigorous research designs, such as randomized controlled trials (RCTs), the most potentially efficacious interventions and the practical problems in applying the RCT to population-level research. Alternative research designs, which possess methodological rigour and are more feasible, are identified and described. The design with the strongest methodological characteristics and feasibility in allowing the evaluation of population interventions is considered to be the multiple baseline. Results: The multiple baseline design addresses potential problems of sample sizes, selection bias, the suitability and baseline stability of outcome measures, statistical analyses and the practicalities of conducting rigorous research in system- or population-level settings. Conclusion: The multiple baseline design has the capacity to allow methodologically and statistically stringent evaluations with relatively small sample sizes, low cost and fewer of the complications imposed by RCTs. Like all research designs it has limitations, but arguably represents the most practical and methodologically rigorous approach to the evaluation of system- or population-level strategies.

GLOBAL BURDEN OF ILLNESS
Alcohol is a major avoidable risk factor for disease and injury, responsible for an estimated 3.8% of deaths and 4.6% of disability adjusted life years (DALY) globally (Rehm et al., 2009). The short- and long-term harms of alcohol consumption (including factors such as crime and violence, treatment costs, loss of productivity and premature death) were estimated to cost Australia $15.3 billion in 2004–2005 (NHMRC, 2009; Mathers et al., 2001). Risky or high-risk alcohol consumption was responsible for 31,000 deaths between 1991 and 2001 and 577,269 hospital episodes between 1993 and 2001. Alcohol consumption accounted for 3.3% of the total burden of disease in Australia in 2003. Alcohol dependence, harmful use of alcohol and road traffic accidents are the leading causes of disease burden for young Australians aged 15 to 25 (Mathers et al., 2001).

The prevalence of alcohol misuse in Europe and the UK, particularly among young people, is a major cause of concern for the British Medical Association. Young British women aged between 16 and 24, for example, have the highest alcohol consumption of their age group in Europe (Gill, 2002; BMA, 2008). Alcohol intoxication was responsible for 6% of road casualties and 17% of deaths in the UK in 2006 (BMA, 2008). In the World Health Organisation European regions in 2002, 6.1% of all deaths and 10.7% of DALY were attributed to alcohol exposure (Rehm et al., 2006). In the USA, a high prevalence of alcohol use was found among the youngest (12–17 year olds) respondents to the National Survey on Drug Abuse, but use was highest among 25–34 year olds (Anthony and Echeagaray-Wagner, 2000). The National Institute on Alcohol Abuse and Alcoholism conducted the National Epidemiologic Survey on Alcohol and Related Conditions, reporting that 8.5% of the adults in the USA met criteria for an alcohol use disorder (USDHSS, 2008). Binge drinking by college students (Wechsler et al., 1998) and increasing alcohol consumption in rural areas (Dew et al., 2007) have also been identified as a serious public health problem.

ACKNOWLEDGEMENT OF THE NEED FOR EVIDENCE
Public health policy (Macintyre et al., 2001) and Health Promotion theory (Nutbeam, 1998) both stress the need for any efforts to address alcohol-related issues to be based on rigorous evidence (Anderson et al., 2005). In response to the severity of the problem of alcohol misuse, the BMA have extensively reviewed the evidence and produced a set of recommendations to promote the development of comprehensive and effective alcohol control policies (BMA, 2008). Even within these evidence-based recommendations, however, is the recognition of the need for further research and evaluation. Sound evidence of intervention effectiveness increases the likelihood that valuable public health resources are not wasted on ineffective interventions and promotes the implementation of interventions, which are likely to be effective in reducing the global burden of illness associated with alcohol.

WHAT CONSTITUTES EVIDENCE?
As the need for methodologically rigorous evidence to assess the effectiveness of interventions for alcohol has increased, so too has the focus in the research design literature, on the randomized controlled trial (RCT) as the gold standard research design. However, it has been argued that the RCT is not the most appropriate design for the evaluation of population-level initiatives where the unit of interest is a state, a geographic unit or a community (Sanson-Fisher et al., 2007). Increasing acknowledgement that reducing alcohol-related harm requires intervention at a system- or population-level rather than a micro- or individual-level has piqued interest.
in alternate methodologically adequate evaluation designs (Eccles et al., 2003).

Effective interventions in system-level dimensions such as those which target access, distribution and policy may have far-reaching and long-term impact. Despite the potential for system-level interventions to impact across large numbers of people, their wide-spread implementation has been limited by a lack of effective measures and evaluation designs with which to quantify their effects (Giesbrecht and Haydon, 2006). Extensive reviews of research output in the alcohol field demonstrate a continuing lack of measures and intervention research (Shakeshaft et al., 1997). The methodological quality of intervention research that has been undertaken is variable, with studies varying in design, methods to reduce bias, statistical analysis approaches and the relevance of the study to community-identified issues (Gates et al., 2006).

SYSTEM-LEVEL APPROACHES
System- or community-level interventions include approaches such as reducing alcohol outlet density (Weitzman et al., 2003), limiting trading hours for licensed premises (Stockwell, 2006), increasing police activity (Stockwell, 1997) and addressing misperceptions of alcohol consumption norms (Kypri et al., 2008). Alcohol control is a complex area which requires multifactorial interventions across sectors and geographical areas. System-level interventions take a holistic approach to intervention and evaluation, with comparisons made between systems themselves, rather than between individual outcomes. In this context, the systems examined may range from communities to treatment centres or licensed premises. It is likely that system-level predictors have substantial influence upon the outcomes achieved (Collins et al., 1985; Prentice and Miller, 1993).

Community-action approaches to reducing alcohol harm are becoming increasingly common. A recent effort to combat harmful alcohol consumption among Swedish youth targeted parents and alcohol vendors, as well as schools and police in an effort to reduce both demand for and supply of alcohol (Stafstrom et al., 2006). Outcome measures assessed in a cross-sectional survey were self-reported drinking patterns and identification of the sources from which alcohol was obtained (Stafstrom et al., 2006). In the USA, a large multi-component community prevention trial included community mobilization, responsible beverage service, drink driving, anti-drunk driving and underage drinking programmes and alcohol access components, founded upon an evidence-based conceptual model. While an adequate evaluation was built in to the design of this study, communities were chosen based on the presence of existing community coalitions which were interested in participating (Holder et al., 1997). An Australian community-action programme, the Surfers Paradise Safety Action Project, developed a risk assessment procedure and code of practice for nightclub managers, as well as improving the external regulation of licensed premises by police and liquor licensing inspectors. Evaluation included pre- and post-implementation assessment of service practices at eight venues, police data on incidents of theft, drunk and disorderly or disturbances, as well as observations of the venues for measures including the physical environment, security, the social environment, patron demographics, staff characteristics, drug and alcohol consumption and costs, responsible serving practices and violence (Homel et al., 1997).

With increasing demand for evidence-based practice, there is a need for stringent evaluation before these approaches are more widely adopted. Evaluation of system-based interventions presents problems for any evaluation technique. There is clearly a need to improve the reliability and validity of measures that derive from routinely collected data and are applicable to system-based interventions (Breen et al., under review; Czech et al., under review). Similarly, there are difficulties in measuring the extent to which all members of a system or community are exposed to an intervention, given such exposure is highly likely to be variable in real world trials. In particular, RCTs can be problematic and possibly not feasible due to ethical, political, logistical or financial reasons (Sanson-Fisher et al., 2007).

THE TENSION IN EVALUATION DESIGN: DOES RIGOUR HAVE TO BE COMPROMISED TO EVALUATE COMMUNITY-LEVEL CHANGE?
Funding practices tend to encourage individual-level efforts due to the extensive time and funding required for community-based interventions. Thus, a tension has been created between recognition of the potential effectiveness of system-level approaches and the limited capacity of some research designs to evaluate these strategies in an acceptable manner at reasonable cost. The World Health Organisation’s European Working Group on Health Promotion Evaluation examined evaluation methods and concluded that in most cases, RCTs were ‘inappropriate, misleading, and unnecessarily expensive’ (WHO, 1998). They recommended that policy makers support the use of multiple methods to evaluate health promotion initiatives and support further research into the development of evaluation approaches (WHO, 1998). The two opposing drives, one for gold standard RCT evidence and the other for the potential effectiveness of a community-level approach, lead to a conflict in identifying the most appropriate evaluation design for community-level alcohol interventions, given both cannot be supported at one time.

The limitations of the RCT for evaluating population-level interventions have been extensively discussed (Sanson-Fisher et al., 2007). RCTs have been described as being unable to accommodate the complexity and flexibility that characterize public health and community-level intervention (Rychetnik et al., 2002). The cost of RCT evaluation designs, ethical and logistical issues associated with consent, the magnitude of sample size required (number of centres) and identification of appropriate outcome measures all limit the feasibility and effectiveness of using the RCT to study system-level initiatives (Hawkins et al., 2007; Sanson-Fisher et al., 2007).

WHAT HAVE WE LEARNT FROM HISTORY?
There are a number of examples of successful public health policies and programmes being implemented without having been subjected to rigorous evaluation from RCTs. One of the first and most well-known community-level interventions in the alcohol field was the introduction of random breath testing (RBT) to reduce alcohol-related traffic crashes. RBT laws
were introduced in all Australian states and territories between 1980 and 1983 with an evidence base limited to the use of breath test procedures introduced in the British Road Safety Act of 1967 (Homel, 1988). In the 5 years following the introduction of RBT, the number of road fatalities in Australia dropped significantly, despite increases in population and vehicle numbers. Similarly, Potts et al. (2006) cite a series of situations throughout history in which lives have been saved and other positive outcomes achieved through the implementation of interventions prior to them being evaluated in RCTs. They suggest that although RCT evaluations may be preferable, logistical and economic constraints often prevent them from occurring in a timely and rigorous manner: "...setting policies based on good science but without RCTs is often more suitable in resource poor settings" (Potts et al., 2006).

Although possible, it is not desirable that public health interventions without strong evidence for their effectiveness or cost efficiency be implemented routinely. In community- or system-level research which guides public health policy making it is, therefore, critical to find a balance between inadequate evidence and the restrictive nature of a dependence on evidence obtained from RCTs.

Although a variety of strategies have been implemented in communities aimed at reducing alcohol-related violence, largely due to the lack of reliable evidence about which interventions are effective, recent interventions include staggered closing times for licensed premises and lock-outs for set periods prior to closing (Long, 2005; Premier’s Drug Prevention Council, 2004; Wiggers, 2007). Evaluations of these types of interventions have typically been limited to before and after assessment, usually consisting of police data relating to reported incidents (Maguire, 2003; Wiggers, 2007). These evaluation designs, however, do not provide sufficient information about intervention effectiveness. In the example of the introduction of RBT in Australia, a before-and-after evaluation showed that a 36% decline in alcohol-related road fatalities in the state of New South Wales was sustained for the first 5 years after implementation and that this was more effective than in the other states and territories (Homel, 1988). With no randomization, staged implementation or formal assessment of extraneous variables, however, it is not possible to confidently quantify the proportion of this reduction directly attributable to RBT and reasons for geographical variation in impact can only be speculated. While RCTs are unlikely to be ethically and logistically appropriate in such situations, sufficient methodological rigour could be achieved through the use of alternate research designs.

SOLUTIONS FOR EVALUATING COMMUNITY-LEVEL CHANGE: EPOC RESEARCH DESIGNS THAT PROVIDE ADEQUATE EVIDENCE AND PROVIDE FEASIBLE AND COST-EFFECTIVE EVALUATION

Heller and Page (2002) have put forward a set of principles for generating strong research evidence in the population health field which can be applied to community-level intervention efforts. These principles include: design considerations, such as the use of routinely collected data for research; statistical considerations, including the use of number needed to treat concepts and adoption of multilevel modelling to analyse clustered data appropriately; and implementation considerations, such as simple methods for data collection across a population of interest in order to calculate population measures of risk. Outcome data should also be presented to policy makers and the public in a manner that is easy to understand, and policy makers should be trained in the interpretation and use of evidence (Heller and Page, 2002).

The RCT is just one of four designs recognized by EPOC criteria (EPOC, 2002) as methodologically sound. Alternative designs are the cluster non-RCT, the before and after study and the interrupted time series (ITS) design. With comparison being the crucial element of evaluation and the capacity to reach conclusions about the effectiveness of interventions, it is important to recognize the alternative research designs which can minimize some of the problems associated with the RCT but still demonstrate that significant change has occurred and it is a consequence of the intervention.

The general principles of research design for any evaluation effort are the minimisation of bias and other threats to internal validity (confounding and chance) and maximization of generalizability. Studies should be designed in order to determine with a high level of confidence whether a change has occurred; whether that change is a result of the intervention and whether the change is significant (Hawkins et al., 2007). These goals can be achieved using alternative research designs to the RCT, which are more practical in the organizational context, and have the potential to increase the volume and quality of evaluation efforts relating to provide behaviour change.

CLUSTER NON-RANDOMIZED TRIALS

The primary difference between randomized and non-randomized controlled trials is the method of allocation to experimental or control conditions. Quasi- or non-random methods can offer logistical benefits but increase the probability of systematic bias, potentially contributing to, or negating any intervention effects (MacLehose et al., 2000).

UNCONTROLLED BEFORE AND AFTER STUDIES

These studies measure performance in a given study site or sites, before and after the introduction of an intervention. While they are relatively simple to design and conduct, and can be used with as few as two groups, the reliance on single data collection points before and after implementation eliminates any capacity to assess the impact of extraneous events on the outcomes. Secular trends or sudden changes can make it difficult to attribute observed changes to the intervention. This intrinsic methodological weakness has led to the over-estimation of intervention effects in some quality improvement interventions.

CONTROLLED BEFORE AND AFTER STUDIES

Here, pre- and post-intervention performance is compared between the study population and a control population of similar characteristics and baseline performance. While these studies, if well designed, should protect against secular trends and sudden changes, difficulties often arise in the identification
of a sufficiently similar control group. Within group, analyses are often performed to overcome the issues arising from baseline differences, where changes from pre- to post-intervention are compared between the groups. These analyses are inappropriate, however, as the two groups are unlikely to be truly comparable, with the possibility that different secular trends or sudden changes may affect them.

ITS DESIGNS

ITS designs have long been used in the evaluation of the effects of public policy change, and more recently, have been recognized for their value in community intervention research in the behavioural sciences (Biglan et al., 2000). ITS designs are appropriate whenever the object of study can be measured reliably on repeated occasions, with time points identified for data collection before and after introduction of the intervention (EPOC, 2002; NHMRC, 2000). The repeated measurements enable trends to be established in both pre- and post-intervention periods. Thus changes in alcohol consumption or incidents of alcohol-related harm can be measured over time, and underlying secular trends can be taken into consideration. Sufficient data points must be collected before the intervention to obtain a stable estimate of the underlying trend. In this design, each group acts as its own control and can therefore be used in as few as one population group. EPOC endorse the ITS approach and give two minimum criteria for inclusion of ITS designs in EPOC reviews; a clearly defined point in time when the intervention occurred, and at least three data collection points before and three after the intervention (EPOC, 2002).

These types of traditional and minimal applications of the ITS design have been used in the evaluation of behaviour changes for several decades (Baer et al., 1968; Barlow and Hersen, 1984) and do represent a good balance between methodological rigour and practical application. Their major limitation, however, is that the lack of a comparison group creates a number of threats to internal validity, including possible selection biases if the composition of the sample changes between data collection points, as is likely in community-level analyses. The design does not protect against the potential effects of simultaneously occurring, unmeasured events that may impact on outcomes (Grimshaw et al., 2000). Methodological rigour and confidence that outcomes do indeed result from an intervention (issues associated with effect heterogeneity) can be increased if multiple time series are conducted. The multiple baseline design refers to the staggered delivery of an intervention to different populations at different time points (EPOC, 2002).

THE MULTIPLE BASELINE DESIGN

The multiple baseline design is a form of ITS which is particularly effective in the development and evaluation of components of interventions. The version of this design most relevant to evaluation of strategies to reduce excessive alcohol consumption and alcohol-related harm is the ‘across subjects’ design, in which an outcome is examined in a number of different units of analysis. Interventions may be delivered to separate communities, staggered at chosen intervals. If a change is observed in the measured outcome following the implementation of the intervention, and is not coupled by changes in communities yet to receive the intervention, the change can be attributed to the intervention with a high degree of confidence. This conclusion is strengthened if the effect is replicated in all communities. Similarly, the design allows for measurement of the impact of extraneous variables, which may affect one or all communities at any given time. This is a major advantage of the multiple baseline design over previously discussed alternatives. Figure 1 illustrates a hypothetical multiple baseline study design in four communities.

The multiple baseline can be strengthened by assessing results against the basic rules of causality (Hawkins et al., 2007). By tracking changes in several populations over time, this design has the capacity to determine with confidence, not only whether the observed effects result from the intervention, but also the strength of intervention effects, the consistency of effect and the magnitude of effect in relation to the complexity of the intervention. The multiple baseline design can take a ‘mission-oriented’ approach, where a multi-component intervention is implemented with the intention to cause early change in the outcome and conduct component analysis at a later stage, by selectively removing components to assess which were most successful or engaging participating communities for feedback on the various components. Alternatively, a ‘component-oriented’ approach can be used, where individual components are added to the intervention, removing or modifying any unsuccessful components until the desired outcome is achieved.

Applying interventions of differing complexity to the separate communities may be a useful tool to assess the level of complexity that is most effective and necessary for intervention success. Consistency across communities in the time between intervention implementation and observation of effects further strengthens the conviction that effects result from the intervention. The design may also be used with differing groups or in different settings to assess the specificity of intervention effects and consistency across fields. For example, the use of a certain intervention strategies to target alcohol-related harm in a range of varied communities may be a useful tool to assess the generalizability of an intervention. Consistency in results across groups and supporting evidence from other research and data sources also add credibility to findings.

![Fig. 1. Hypothetical multiple baseline study design in four communities.](https://academic.oup.com/alcalc/article-abstract/45/5/481/184460/184460)
METHODOLOGICAL CONSIDERATIONS OF THE MULTIPLE BASELINE DESIGN IN EVALUATING COMMUNITY-LEVEL INTERVENTIONS

Sample size
As few as two units of interest such as communities or regions may be sufficient to test an intervention with the multiple baseline design, potentially reducing costs and minimizing logistical difficulties associated with recruitment and randomization of large sample sizes (Atienza and King, 2002). As with most experimental designs, however, the capacity to generalize results will be greater with increasing numbers of groups and, to a lesser extent, increased numbers of individuals within each group. Also, careful selection of participating groups can improve representativeness and generalizability of results. Random selection from matched clusters (such as communities of similar size, geographical location and infrastructure-equipped), for example, will help ensure that selected groups adequately represent the eligible sample population and reduce the possibility of selection bias (Hawkins et al., 2007; Atienza and King, 2002).

Selection bias
Once participating communities have been selected, the order in which they implement the intervention could potentially be influenced by factors such as the level of cooperation within a community or the ease of implementation in some communities over others. It is recommended that groups are randomly allocated into the order in which they will receive the intervention, in an effort to minimize selection bias (Hawkins et al., 2007).

Baseline stability
The length of the baseline should be sufficient to include enough data points for stringent statistical analysis and account for extraneous variations. A minimum of three data points are required to plot a trend (Crosbie, 1993), and a minimum of 10 data points are necessary for procedures such as ITSACORR analysis (Biglan et al., 2000). An associated issue is the timing of introduction of the intervention in each community. The intervals between intervention implementation should be sufficient to enable monitoring of variations, while obviously considering the logistics and overall time required to complete the study (Hawkins et al., 2007).

Suitability of measures
The outcome measures used should be consistent across all groups and for the duration of the study. Collection of consistent, reliable data over a long period can be logistically difficult and expensive. Using routinely collected data is advantageous, in that baseline and post-intervention data collection can continue for extended time periods at low cost and may eliminate the need for reliability and validity studies for other data collection procedures (Hawkins et al., 2007).

Statistical analyses
Large numbers of data points are necessary in time series studies to enable rigorous statistical modelling. Grimshaw et al. report that many ITS published to date have been inappropriately analysed and often overestimate intervention effects (Grimshaw et al., 2000). Autocorrelation is the main challenge faced in analysing repeatedly measured data. This refers to the fact that values taken at any given time may be influenced by previously measured values and may impact upon estimates of the intervention effect. Auto-regressive integrated moving average modelling overcomes this problem, removing trends and random drift in baseline data by transforming the values in a time series, using the difference between successive observations (Kazdin, 1984). Once this model is applied, traditional statistical techniques can be used. This model requires extended periods of observation to gather sufficient baseline data. A less sophisticated technique appropriate for shorter time series, independent time series analysis of autocorrelated data (ITSACORR), evaluates differences in slope and intercept between baseline and intervention phases while including autocorrelation factors, leading to $F$ and $t$-test assessments for changes resulting from the intervention (Crosbie, 1993).

French and Heagerty (2008) have examined analysis methods for longitudinal data and conclude that existing approaches such as generalized linear mixed models, random-effects meta-analysis and empirical Bayes estimators are well suited to assessing the impact of policy change. They recommend a series of steps as an outline from which to approach policy change analysis. These include exploratory longitudinal analysis, unit-specific summaries, effect heterogeneity and average effect (French and Heagerty, 2008).

Feasibility
A major advantage of the multiple baseline design is that it allows an intervention to occur on a systems basis while still providing an opportunity of application of the research design which can answer the question about whether a change occurred, and whether the change was a consequence of the intervention, as opposed to some other miscellaneous event. For example, four geographically separate communities adopt a similar but complex intervention targeting alcohol consumption and levels of violence. The baseline period ranges from 6 to 18 months as adoption of the intervention is staggered across the communities. The intervention aims to reduce consumption levels and alcohol-related harm, so a challenge will be to identify accurate and consistent outcome measures. Arrests, recorded violent incidents at licensed venues, emergency department visits, alcohol-related traffic incidents and alcohol sales would all be useful measures, but these would need to be collected and recorded consistently, with the same coding and definitions across all communities. Assuming this could be achieved (and problems solved in the early part of the baseline period), the study would generate rigorous, generalizable (because the intervention can be implemented in a manner appropriate for each of the participating communities) evidence for the impact of the interventions and allowing for the impact of any extraneous events (and therefore the extent to which changes in outcomes did result from the interventions). The results, if positive and consistent across the four communities, may allow a conclusion that the intervention is effective or alternatively create the justification for an expensive and time-consuming shift to a RCT involving a cluster randomized trial of 14 or more communities.
WHERE TO FROM HERE?

The current political and social climate has recognized the need for intervention to reduce excessive alcohol consumption and alcohol-related harm. Internationally, governments are committing funds to take action at a population-level and implementing strategies which have the potential to significantly impact on this and other substantial public health issues. It is both ethically and conceptually important to rigorously evaluate the cost-effectiveness of public health policy and community-level interventions. The multiple baseline design represents perhaps the most practical and methodologically rigorous approach to the evaluation of such strategies, with the capacity to answer the key questions of evaluation without compromising on the strength of evidence obtained.

Conflict of interest statement. None declared.

REFERENCES


Cochrane Database Syst Rev CD005803.


Gill JS. (2002) Reported levels of alcohol consumption and binge drinking within the UK undergraduate student population over the last 25 years. Alcohol Alcohol 37:109–20.


