Design and Analysis of Group-Randomized Trials: A Review of Recent Methodological Developments

David M. Murray, PhD, Sherri P. Varnell, PhD, MS, and Jonathan L. Blitstein, MS

We review recent developments in the design and analysis of group-randomized trials (GRTs). Regarding design, we summarize developments in estimates of intraclass correlation, power analysis, matched designs, designs involving one group per condition, and designs in which individuals are randomized to receive treatments in groups. Regarding analysis, we summarize developments in marginal and conditional models, the sandwich estimator, model-based estimators, binary data, survival analysis, randomization tests, survey methods, latent variable methods and nonlinear mixed models, time series methods, global tests for multiple endpoints, mediation effects, missing data, trial reporting, and software.

We encourage investigators who conduct GRTs to become familiar with these developments and to collaborate with methodologists who can strengthen the design and analysis of their trials. (*Am J Public Health.* 2004;94:423–432)

Group-randomized trials (GRTs) are comparative studies designed to evaluate interventions that operate at a group level, manipulate the physical or social environment, or cannot be delivered to individuals.[1] Examples include school-, worksite-, and community-based studies designed to improve the health of students, employees, and residents, respectively. Just as the randomized clinical trial (RCT) is the gold standard in public health and medicine when allocation of individual participants is possible, the GRT is the gold standard when allocation of identifiable groups is necessary.

There are 4 characteristics that distinguish the GRT from the more familiar RCT. First, the unit of assignment is an identifiable group; such groups are formed not at random but rather through some physical, social, geographic, or other connection among their members. Second, different groups are assigned to each condition, creating a nested or hierarchical structure for the design and the data. Third, the units of observation are members of those groups nested within both their condition and their group. Fourth, usually only a limited number of groups are assigned to each condition.

These characteristics create several problems in the design and analysis of GRTs.[1] The major design problem is that a limited number of often heterogeneous groups makes it difficult for randomization to distribute potential sources of confounding evenly in any single realization of the experiment. This increases the need to use design strategies that will limit confounding and analytic strategies to deal with confounding when it is detected. The major analytic problem is that there is an expectation for a positive intraclass correlation (ICC) among observations of members of the same group.[2] That ICC reflects an extra component of variance attributable to the group above and beyond the variance attributable to its members. This extra variation will increase the variance of any group-level statistic beyond what would be expected with random assignment of members to conditions. Moreover, with a limited number of groups, the degrees of freedom available to estimate group-level statistics are limited. Any test that ignores either the extra variation or the limited degrees of freedom will have a type I error rate that is inflated, and this effect will only worsen as the ICC increases.[3]

Cornfield[4(p101–102)] warned of this danger 25 years ago when he noted that ignoring these problems was “an exercise in self-deception . . . and should be discouraged.” That warning was followed by a gradual increase in the number of methods papers in this area. The first comprehensive text on the design and analysis of GRTs appeared in 1998.[1] It detailed the design considerations for the development of GRTs, described the major approaches to their analysis both for Gaussian and binary data, and presented methods for power analysis applicable to most GRTs. We use that text as a point of departure for this review and assume that readers are familiar with its basic material.

Over the past 5 years, many articles have discussed the methodological issues involved in GRTs generally or in design papers describing new trials.[5–28] The second textbook on design and analysis of GRTs appeared in 2000.[29] That text provided a good history of GRTs and examined the role of informed consent and other ethical issues. It focused on extensions of classical methods, although it also included material on regression models for Gaussian, binary, count, and time-to-event data. Other textbooks on analysis methods germane to GRTs appeared during the same period,[30–33] as well as a large number of articles on new methods relevant to the design and analysis of GRTs. In the sections that follow, we bring the reader up to date on many of these developments.

DESIGN ISSUES

In 1998, Murray[1] detailed the design considerations for a GRT, whether the study was to use a nested cohort or nested cross-sectional design; whether the study was to have a posttest-only design, a pretest–posttest design, or an extended design with multiple pretest/posttest measures; and whether the design was to be completely randomized or to include matching/stratification. At that time, investigators were limited by the paucity of ICC and other parameter estimates needed to select an efficient design and to ensure that the study would have adequate power (the probability of rejecting the null hypothesis when it is false). One of the important recent developments has been the publication of papers providing estimates for those parameters. Another has been the publication of important refinements in the methods used for power analysis. There have also been important

TABLE 1—Recent Articles Presenting Intraclass Correlations and Related Parameter Estimates

Groups	Population	Type of Endpoint(s)	Source
Clinics	Adolescents	Tobacco, alcohol	Slymen and Hovell[41]
Clinics	Adults	Preventive practices	Baskerville et al.[129]
Clinics	Adults	Process and cost-effectiveness	Campbell et al.[130]
Clinics	Adults	Pregnancy	Piaggio et al.[131]
Clinics	Adults	Assessment and management of the elderly	Smeeth and Ng[132]
Communities	Adults	Tobacco, eating patterns, alcohol, weight	Gulliford et al.[133]
Communities	Adults	Heart attack delay	Murray et al.[42]
Communities	Adults, youths	Eating patterns, tobacco, alcohol	Feng et al.[35]
Schools	Adolescents	Tobacco, alcohol, other drug use	Scheier et al.[134]
Schools	Adolescents	Eating patterns, physical activity	Murray et al.[135]
Schools	Adolescents	Tobacco	Murray et al.[136]
Schools	Adolescents	Alcohol	Murray et al.[39]
Worksites	Adults	Wood dust exposure	Lazovich et al.[137]
Worksites	Adults	Tobacco, physical activity, alcohol, weight	Martinson et al.[138]

developments in several specific designs, including matched designs, designs involving 1 group per condition, and designs in which individuals are randomized to receive treatments in groups.

New Estimates of ICCs

Investigators planning a GRT should not proceed absent a good estimate of the extra variation likely to be present in their primary analysis. To do so is to risk a substantially underpowered or overpowered study. Table 1 lists articles published in the past 5 years that have reported ICC and related parameter estimates. Donner and Klar reported ICCs from a number of other studies,[29] as did Murray and Blitstein.[34] Collectively, these sources provide estimates for a wide variety of groups, members, and endpoints so that investigators now have a better opportunity of finding estimates that are well matched to the circumstances of the trial they are planning.

Murray and Blitstein[34] also reported a pooled analysis of ICCs from worksite, school, and community studies. They confirmed that the adverse impact of a positive ICC can be reduced by regression adjustment for covariates[1,35–38] or by taking advantage of overtime correlation in a repeated measures analysis.[1,35,39] Janega et al. (unpublished data, 2003) have shown that standard errors for intervention effects from end-of-study analyses that reflect these strategies are often different from the standard errors estimated from baseline analyses. Because the ICC of concern in any GRT is the ICC as it operates in the primary analysis,[1] these findings reinforce the need for investigators to use estimates in their power analyses that closely reflect the endpoints, target population, and primary analysis planned for the trial. And while the sources just cited will help considerably in this regard, we join others who have urged publication of such estimates as a routine part of reporting the results of GRTs.[40]

Power Analysis

Most of the sources that reported ICCs also showed how they could be used to size a new GRT, as did many of the papers cited earlier as general reviews. We do not repeat the standard presentation here and instead refer readers to those sources, and especially to chapter 9 in the Murray text,[1] including the examples offered at the end of that chapter. Even so, a few points bear repeating here. First, the increase in between-group variance due to the ICC in the simplest analysis is calculated as $1+(m-1)$ICC, where m is the number of members per group; as such, ignoring even a small ICC can underestimate standard errors if m is large. Second, while the magnitude of the ICC is inversely related to the level of aggregation, it is independent of the number of group members who provide data. For both of these reasons, more power is available given more groups per condition with fewer members measured per group than given just a few groups per condition with many members measured per group, no matter the size of the ICC.

Third, the 2 factors that largely determine power in any GRT are the ICC and the number of groups per condition. For these reasons, there is no substitute for a good estimate of the ICC for the primary endpoint, the target population, and the primary analysis planned for the trial, and it is unusual for a GRT to have adequate power with fewer than 8 to 10 groups per condition. Finally, the formula for the standard error for the intervention effect depends on the primary analysis planned for the trial, and investigators should take care to calculate that standard error, and power, based on that analysis. Chapter 9 in the Murray text[1] provides formulas for many of the common analyses, and generic formulas and examples are provided in recent work conducted by Janega et al. (unpublished data, 2003).

Several variations on the standard power analysis have appeared during the past 5 years. Slymen and Hovell presented a method that allows the investigator to compare sample size requirements for a GRT and an RCT based on the anticipated magnitude both of the ICC and of any contamination.[41] They showed that for small groups, where contamination was likely to be substantial, GRTs were a natural choice, while for large groups, where contamination was likely to be modest, RCTs were a natural choice. Hayes and Bennett presented sample size formulas for pair-matched and pair-unmatched GRTs in terms of coefficients of variation rather than ICCs for investigators more familiar with the former than the latter.[21] Murray et al. defined the design effect as it operates in a random coefficient model and presented methods for power analyses of such models.[42]

Kerry and Bland compared 3 methods for weighting group means in sample size calculations when those means are based on a variable number of observations; they reported that minimum variance weights were superior to uniform weights, particularly when clusters were small, and superior to cluster-size weights, particularly when the clusters were large.[43] Lake et al. showed how power could be improved without increasing the type I error rate using a strategy in which sample size is reestimated after the start of recruitment using the initial data.[44] This strategy has application in situations in which many groups are to be randomized and recruitment of those groups is to take place over a long period of time (e.g., some family studies). Liu et al. provided a technical discussion of sample size and power for analytic models involving differences between means, slopes, or proportions for GRTs involving repeated observations of the same groups and members[45]; less technical presentations are also available.[1,42] Raudenbush discussed sample size in GRTs accounting for the cost of recruiting members and groups and provided formulas for optimal size with and without covariate adjustment.[46]

Matched Designs

Almost half of the GRTs published in the *American Journal of Public Health* and *Preventive Medicine* during the period 1998 through 2002 involved matched designs.[47] Even so, Klar and Donner suggested that stratification may be a better design choice to ensure balance on potential confounders.[10,29,48] They argued that stratification exacted a lower price in terms of degrees of freedom, and certainly that is true. Klar and Donner also pointed out that estimation of the ICC in a matched design assumes homogeneity of effects across pairs, and they gave that as another reason to avoid a matched design. Others have argued that this assumption is often reasonable.[49]

Raab and Butcher proposed an alternative to matching[50] based on a balancing criterion calculated as a weighted sum of squared differences between the condition means on any proposed covariates. Groups would be divided into 2 sets providing a small enough value on their criterion, followed by random assignment of sets to conditions. Raab and Butcher argued that this scheme would support model-based methods because it would fulfill the conditional independence criterion. To support a randomization test, they proposed that the criterion be calculated for all possible allocations of groups to conditions, that some subset of those allocations be identified as having a small enough value on the criterion to be acceptable, and that one such allocation be chosen at random, followed by random assignment of sets to conditions.

One Group per Condition

GRTs with 1 group assigned to each condition have been criticized as unable to support a valid analysis for an intervention effect, absent strong and untestable assumptions.[1,29] Even so, these designs continue to appear, both in applications submitted to National Institutes of Health study sections and in articles in the peer-reviewed literature.[47] Varnell et al. recently provided additional documentation of the dangers of this design and urged investigators to avoid it except in the case of pilot studies.[51]

Individuals Randomized to Receive Treatments in Groups

A design intermediate between a GRT and an RCT exists in which individuals are randomized to study conditions but receive their treatment in small groups or from the same intervention team seen by other participants. Those shared experiences may result in correlated errors, just as they do in GRTs. While some may regard this as a type of “intervention effect,” it is instead a threat to the internal validity of the trial. This concern was raised nearly 20 years ago in the context of designs in which endpoints were determined jointly by patients and their providers.[52] Several recent articles have echoed that concern.[46,53–55]

Most recently, Varnell et al. compared analyses for these studies in simulations, varying the number of groups per condition, the magnitude of the ICC, and the number of conditions that received an intervention in small groups while fixing the intervention effect at zero.[56] Analyses that ignored the ICC had an inflated type I error rate, with the magnitude of the problem dependent on the size of the ICC, the number of members per group, and the number of conditions in which participants received treatment in groups. A mixed-model regression approach with the group included as a nested random effect and degrees of freedom based on the number of groups carried the nominal type I error rate. This finding confirms that allowing participants to interact with each other in small groups does not maintain the independence of observations required for the usual RCT analyses.

ANALYSIS ISSUES

Murray[1] identified several analytic approaches that can provide a valid analysis for GRTs. In each, the intervention effect is defined as a function of a condition-level statistic (e.g., difference in means, rates, or slopes) and assessed against the variation in the corresponding group-level statistic. These approaches included mixed-model analysis of variance (ANOVA)/analysis of covariance (ANCOVA) for designs having only 1 or 2 time intervals, random coefficient models for designs having 3 or more time intervals, and randomization tests as an alternative to the model-based methods. Murray[1] identified other approaches as invalid for GRTs because they ignored or misrepresented a source of random variation. These included (1) analyses that assessed condition variation against individual variation and ignored the group, (2) analyses that assessed condition variation against individual variation and included the group as a fixed effect, (3) analyses that assessed the condition variation against subgroup variation, and (4) analyses that assessed condition variation against the wrong type of group variation.

Murray[1] identified still other strategies as having limited application for GRTs. Application of fixed-effect models with post hoc correction for extra variation and limited degrees of freedom assumes that the correction is based on an appropriate ICC estimate, and in 1998 few estimates were available. Application of survey-based methods or generalized estimating equations (GEE) and the sandwich method for standard errors requires that a total of 40 or more groups be included in the study, and in 1998 most GRTs did not include 40 groups.

During the past 5 years, considerable attention has been focused on analytic issues germane to GRTs, including refinements for existing methods and development of new methods. Much of this work has occurred outside the context of GRTs but has application to GRTs, and so we include it in this review.

Conditional versus Marginal Models

Conditional or subject-specific models are typified by mixed-model regression[57] and incorporate random effects to reflect the correlation among observations made of members of the same group; the observations are considered independent conditional on those random effects. Marginal or population-averaged models are typified by GEE[58,59] and define the marginal expectation of the dependent variable as a function of the predictor variables and assume that the variance is a known function of the mean; they separately specify a correlation structure for observations made of members of the same group. In the case of Gaussian data, interpretation of the condition coefficient is the same in conditional and marginal models; however, in the case of binary data, the condition coefficient from a marginal model is smaller than that from a conditional model and has a different interpretation.

In the marginal model, the condition coefficient is the between-person difference in the log odds of the outcome comparing the effects of the intervention and control conditions as if they had been delivered to 2 different individuals. In the conditional model, the condition coefficient is the within-person change in the log odds of the outcome comparing the effect of the intervention and control conditions as if they had been delivered to the same individual. Several recent papers have recommended conditional models for GRTs focused on change within participants (e.g., preintervention vs postintervention) and marginal models for GRTs focused on differences between participants (e.g., intervention condition vs control condition). Unfortunately, both approaches have problems in certain binary data situations; because these issues affect the remainder of our presentation, we consider them first.

Limitations of the Sandwich Estimator Used in Marginal Models

One of the advantages of GEE is that it uses an estimator for variances of fixed effects that is asymptotically robust to misspecification of the correlation structure; the sandwich estimator is so named because the expression of this estimator “sandwiches” an approximate correlation matrix inside 2 outer layers of matrix algebra that otherwise define the variance of a weighted least squares estimator. Unfortunately, the sandwich estimator is biased downward when the number of groups is below 40, whether in GRTs[60–62] or in other designs involving correlated binary data.[63–65] This problem only increases as the number of groups becomes smaller.[66–68] Many investigators working in GRTs appear to be unaware of this limitation, in that there have been many applications of GEE and the sandwich estimator in GRTs involving fewer than 40 groups.[47] Thornquist and Anderson reported more than 10 years ago that this bias was corrected in a GRT by inflating the variance to reflect the uncertainty in the estimation of the fixed effects, much as restricted maximum likelihood (REML) estimation does relative to full maximum likelihood (ML) estimation. Paired with a *t* test and using degrees of freedom based on the number of groups, the size of their corrected test was at the nominal level.[60]

More recent work has also focused on the development and evaluation of correction procedures, though usually not in the context of GRTs. Long and Ervin[69] provided additional results for 3 corrections introduced earlier by MacKinnon and White[65] and reported that a jackknife estimator (a nonparametric method to estimate standard errors based on repeated subsamples) was better than the alternatives. Mancl and DeRouen reported a corrected estimator that was of nominal size even with 10 groups per condition and only 16 observations per group[67]; they also offered an SAS macro. Corcoran et al.[70] offered an exact test, but it has only narrow application to situations in which the groups represent ordered levels of an underlying factor such as dose. Fay and Graubard reported that the sandwich estimator worked well, even in small samples, so long as the usual Wald test was evaluated not as a χ^2 value but as an F ratio of the form F(1, *d*), where *d* is calculated as a function of the variance of the sandwich estimator.[71]

A similar correction provided by Kauermann and Carroll replaces the usual cutpoint in the *z* distribution with a cutpoint that is a function of the variance of the sandwich estimator; they demonstrated its utility even when the sample size was as small as 5.[72] Pan and Wall offered a correction much like that of Fay and Graubard in the form of an approximate *t* or F test, with degrees of freedom defined as a function of the variance of the sandwich estimator.[73] Bell and McCaffrey[74] offered a correction and a Satterthwaite approach to degrees of freedom that seemed to involve less bias and a better type I error rate than the sandwich estimator or the corrected estimators recommended by Long and Ervin[69] or Mancl and DeRouen.[67] Preisser et al. suggested using a model-based variance estimator in GEE, rather than the sandwich estimator, as another solution.[75]

Unfortunately, none of these corrections appear in the standard software packages, so they are relatively unavailable to investigators who analyze GRTs. Absent an effective correction, the sandwich estimator will have an inflated type I error rate in GRTs involving fewer than 40 groups, and investigators who use this approach continue to risk overstating the significance of their findings.

Limitations of Model-Based Estimators Used in Conditional Models

Rodriguez and Goldman[76] reported that multilevel analyses of binary data underestimate both fixed effects and their variances when the ICC is large (0.231 in their data) and there are few observations per group (e.g., family-based studies). With a smaller ICC (0.041 in their data), underestimation is quite modest, even with few observations per group. Breslow and Clayton[77] and Ten Have et al.[78] reported a similar problem for models fit via penalized quasi-likelihood (PQL) estimation. This led some to question the use of conditional models for GRTs involving binary data. That appears to be an overreaction, because most GRTs involve many observations per group and small ICCs; under these conditions, there is little bias. In fact, the simulation study of Hannan and Murray[79] indicated that a conditional model for Gaussian data carried the nominal type I error rate when applied to binary data with an ICC as large as 0.05, so long as there were at least 4 groups per condition and 25 observations per group.

Methods for Binary Data

Gibbons and Hedeker proposed a random-effects probit and logistic regression model for data with 3 levels of nesting based on ML estimation using numerical integration.[80] Their approach would be preferred over PQL procedures when the number of observations per group is quite small, but it is computationally intractable with more than 5 or 6 random effects; this is a problem common to methods that rely on numerical integration. Unfortunately, many models fit to longitudinal data in the context of GRTs have 5 random effects, and some stratified models have 7[1]; such models would be difficult to fit with these methods. Aitkin proposed a nonparametric method based on ML estimation[81]; he noted that this approach had been widely viewed as computationally intensive, but his method avoided that problem. The benefit of the nonparametric method is that it does not depend on correct specification of the distribution of random effects. Bellamy et al. reported a simulation study comparing mixed-model regression (using the SAS GLIMMIX macro) and GEE[68] and confirmed earlier reports that GEE was liberal with fewer than 40 total groups, while GLIMMIX was conservative when the average cluster size was quite small.

Several Bayesian approaches have also been suggested. Kleinman and Ibrahim proposed a semiparametric Bayesian approach to generalized linear mixed models but provided no simulation results to evaluate their method.[82] Ten Have and Localio[83] proposed an empirical Bayes method based on numerical integration and incorporated an adjustment for the standard error; their method performed better than PQL estimation given many small groups (100 groups with 2 observations per group) but not as well as PQL estimation with a smaller number of larger groups (20 groups and 100 observations per group). As such, their method may be useful in family-based GRTs but not in school-, worksite-, or community-based GRTs. Turner et al. discussed a Bayesian approach involving specification of an informative prior ICC distribution based on values taken from the literature[84]; as published values for ICCs become increasingly available, their approach may prove useful. A much simpler approach for binary data was reported by Hannan and Murray,[79] who indicated that the familiar conditional model for Gaussian data carried the nominal type I error rate even when applied to binary data with an ICC as large as 0.05, so long as there were at least 4 groups per condition and 25 observations per group.

Methods for Survival Analysis

Hedeker et al. proposed a discrete-time survival model that allowed multiple random effects, operated under either the proportional hazards or proportional odds assumption, and relied on ML estimation using numerical integration.[85] Hedeker et al. did not provide simulation results for their method. Donner and Klar[29] described group-level methods that could be applied to either discrete-time or continuous-time survival data but did not allow for adjustment for individual-level factors; importantly, the unweighted form assumed that each group's survival rate was equally precise. Frailty models allow the hazard rate to vary at random among groups,[86] but their effect estimates may be difficult to interpret.[29]

Marginal survival models employ standard Cox regression methods to estimate the effect of the intervention and then use the sandwich estimator to obtain standard errors for the fixed effects[87–89]; their intervention effect estimates are readily interpretable, but caution is required if the total number of groups is less than 40. Sargent described an adaptation of the Cox model to incorporate random effects using Bayesian methods but provided no simulation data on the performance of the method.[90] Vaida and Xu[91] described a random-effects model for proportional hazards regression similar to that of Sargent, but they also did not provide simulation results.

Yau[92] proposed a 3-level proportional hazards model estimated via REML. He reported results from a simulation study involving only 10 groups with just 3 members per group and 3 repeated observations for each member; censoring varied from 30% to 60%. Yau's method provided unbiased estimates of fixed effects but slightly overestimated random effects; the overestimation of random effects was reduced with even slightly increased group size. Other advantages were that the baseline hazard function did not have to be specified and estimation did not rely on numerical integration. Cai et al.[88] proposed a transformation model with random effects based on numerical integration and showed that it was less biased than some of the earlier parametric models. Lui et al. proposed several methods for confidence interval estimation for rate ratios based on the beta-binomial distribution[93]; they reported that an interval estimator based on a log transform performed best in simulations, but their smallest study included 20 groups per condition, so the small sample properties of the estimator are unknown.

Bennett et al. presented a 2-stage approach to analysis of incidence rates based on person-year data,[94] estimating group-specific rates (for an unadjusted analysis) or residuals (for an adjusted analysis) in a first stage without regard to intervention status; these rates or residuals were used in a second stage to estimate the intervention effect and assessed via a *t* statistic with degrees of freedom based on the number of groups. Simulation studies showed this approach had nominal size even with as few as 3 groups per condition and perhaps 30 members per group. While these results are encouraging, it would be of interest to see how the method performs with smaller groups.

Randomization Tests

In a randomization test for a GRT, the data are analyzed on the basis of the actual assignment of groups to conditions and then reanalyzed for every other possible assignment of groups to conditions given the design, including any limitations in randomization due to matching, stratification, and the like. The test statistic observed on the basis of the actual assignment is referenced against the distribution of such statistics calculated from the set of all possible assignments. The 2-tailed *P* value for the observed test statistic is defined as the proportion of the possible test statistics that are as large as or larger than the observed test statistic in terms of absolute value. Randomization tests were first used in GRTs in the context of the Community Intervention Trial for Smoking Cessation (COMMIT).[95–97] Gail et al.[98] later demonstrated that randomization tests carried the desired type I error rate for the null hypothesis of no treatment effect on average so long as the number of groups assigned to each condition was the same. Given balance at the group level, randomization tests also carried the desired type

I error rate for dichotomous endpoints and for analyses that included regression adjustment for a covariate, even when the regression model was not correctly specified.[98]

At the same time, randomization tests can have less power than model-based tests when the model is correct. To address that problem, Braun and Feng[99] developed a weighted randomization test using the inverse of the total variance for each group as the weight; they showed this test to be the uniformly most powerful randomization test for Gaussian data. They also developed a locally most powerful randomization test based on a more complicated quasi-score method for non-Gaussian data. In a series of simulation studies, Braun and Feng showed that their optimal randomization test had nominal size and better power than alternative randomization tests or GEE, although it was still not as powerful as the model-based analysis when the model was specified correctly; additional research is needed to compare Braun and Feng's optimal randomization test and model-based methods under model misspecification.

Survey Methods

The clustering of data in GRTs has much in common with the clustering of data observed in complex surveys; as a result, analysis methods developed for complex surveys can have application in the analysis of data from GRTs.[100,101] Since the introduction of GEE, there has been a convergence in methods used for survey applications and for many nonsurvey applications involving correlated data, including GRTs. LaVange[100] showed that parameter estimates and standard errors from their survey logistic regression procedure were identical to those obtained with GEE under the assumption of working independence. LaVange also provided information on survey analysis procedures for proportional odds and proportional hazards regression models, which would be applicable to GRTs. The SUDAAN software package supports those models (http://www.rti.org/sudaan/home.cfm). Caution is required as with other methods that are asymptotically valid only when the total number of groups is below 40 unless special procedures are used to correct for underestimation; LaVange[100] discussed this problem and proposed a correction.

Latent Variable Methods and Nonlinear Models

Muthen[102] presented a general latent variable modeling approach that encompassed a variety of techniques used in GRTs, including mixed-model ANOVA/ANCOVA and random coefficient models. Schulenberg and Maggs observed that mixed models and latent variable models gave identical results when set up to test equivalent models.[103] Others have noted important differences between these approaches[103–106]; however, some of these differences may disappear with improvements in software.

Nonlinear mixed models are a type of mixed model in which both the fixed and random effects have a nonlinear relationship with the endpoint. They differ from the more familiar generalized linear mixed models in which the fixed and random effects are linearly related to a predictor and the predictor is related to the endpoint through a nonlinear link function. Readers are referred to Davidian and Gilinian[107] or Vonesh and Chinchilli[108] for further information.

Interrupted Time Series

Gruenewald[109] and Biglan et al.[110] suggested interrupted time series methods for the evaluation of community-level interventions. The classic time series analysis compares data in a large geographic unit before and after an intervention and evaluates the intervention effect as a change from the preintervention trend, level, or variance. It draws its strength for estimating the preintervention and postintervention time patterns from many observations, thereby providing good precision. These methods would appear to be useful for within-community comparisons but, absent a reasonably large number of communities, not for between-community comparisons. If the number of communities is limited, degrees of freedom for between-community comparisons will be limited and power will be poor; nor would asymptotically valid tests be appropriate with limited degrees of freedom.

Global Tests for Multiple Endpoints

Many GRTs have more than 1 primary endpoint, raising the issue of how to adjust the type I error rate for multiple tests. One solution is to divide the nominal type I error rate evenly among the tests. Feng and Thompson offered as an alternative a global test that functions in much the same way as a multivariate test statistic.[18]

Methods for Analysis of Mediation Effects

Krull and MacKinnon described methods for mediation analyses in GRTs using extensions of methods developed for RCTs.[111] Simulation results indicated that the mediation estimators were unbiased and that estimation of standard errors via first-order Taylor series approximation was preferred. MacKinnon et al. expanded that discussion in an application to tobacco prevention research to include a discussion of a model with multiple mediators.[112]

Missing Data

Missing data are as serious a problem in GRTs as they are in RCTs. Fortunately, methods developed for RCTs are easily adapted to GRTs. For example, Yi and Cook reported on marginal methods for missing data from clustered designs.[113] Hunsberger et al. described strategies for missing data in GRTs and identified a multiple imputation method that carried acceptable type I and type II error rates in simulations.[114]

Software

There has been substantial improvement over the past 5 years in the software available for analysis of GRTs. Zhou et al. reviewed many of these programs and reported that when they were used correctly to fit equivalent models, they gave the same results in simulation studies.[115] HLM (http://www.ssicentral.com/hlm/hlm5all.htm) provides a flexible and powerful vehicle for a variety of analyses appropriate for GRTs.[32,116] It can be used with Gaussian, binary, and Poisson data and can fit 2- and 3-level models. As such, it supports both nested cross-sectional and nested cohort designs. HLM also supports latent variable estimation, multiple imputation, GEE, and sandwich estimation for standard errors. HLM relies on REML for Gaussian endpoints and PQL for non-Gaussian endpoints. The Laplace approximation to ML is available for 2-level and 3-level Bernoulli models.

Several SAS (http://www.sas.com/) procedures support analyses for GRTs. PROC MIXED[117,118] supports models and covariance structures for Gaussian endpoints.[1,119] The GLIMMIX macro[118] supports parallel models and structures for non-Gaussian endpoints and can perform mixed-model logistic and Poisson regression. Some have criticized GLIMMIX because it uses pseudo-likelihood estimation, which is similar to PQL and so underestimates fixed effects and their standard errors under the circumstances noted earlier. However, because most GRTs do not fit those circumstances GLIMMIX continues to be a valid tool in most GRTs.

More recently, SAS introduced PROC NLMIXED, which is a nonlinear mixed-model regression procedure.[117] NLMIXED uses numerical integration for ML estimation and so is more appropriate than GLIMMIX for GRTs that involve very small groups (e.g., family studies). NLMIXED can be used with Gaussian, binomial, and Poisson distributions for mixed-model linear, logistic, and Poisson regression; users can also construct their own log-likelihood function to perform, for example, a clustered ordinal logistic regression or frailty analysis (O. Schabenberger; written communication; April 9, 2003). NLMIXED can accommodate nested designs, although the procedure will encounter computational difficulties if the number of random terms exceeds 5 or 6.[120]

The NLMIXED procedure does not support the within-group repeated measures structures available in MIXED and GLIMMIX; instead, NLMIXED assumes that repeated observations within a member or group are uncorrelated. MIXED and GLIMMIX support model-based and sandwich estimation for standard errors, while NLMIXED provides only model-based estimation. PROC PHREG and PROC GENMOD support sandwich estimation for standard errors and so can be applied to GRTs to perform Cox regression and logistic and ordinal logistic regression, respectively[121]; however, caution is required when there are fewer than 40 groups, absent a correction for the bias in the sandwich estimator.

MIXOR (http://tigger.uic.edu/~hedeker/mix.html) and its related programs[122–124] can be used with Gaussian, binary, and Poisson data to provide mixed-model linear, logistic, and Poisson regression. These programs also allow mixed-model grouped-time survival analysis,[85] mixed-model logistic or probit analysis for ordinal endpoints,[125] and mixed-model logistic regression for nominal endpoints.[124,126]

The MlwiN program (http://multilevel.ioe.ac.uk/index.html) can be used with Gaussian, Bernoulli, binomial, multinomial, and Poisson distributions and can also fit ordinal logistic models for clustered data.[127] The SUDAAN software package (http://www.rti.org/sudaan/home.cfm) (Research Triangle Institute, Research Triangle Park, NC) supports models for analysis of survey data that are often applicable to GRTs. In addition, SPSS (http://www.spss.com) has introduced a mixed-model regression program that supports several covariance structures.[128]

None of the programs just mentioned incorporate a correction for the underestimation bias in the sandwich estimator when the data are binary and there are few groups per condition. As indicated earlier, the work in that area seems to be converging on a solution, and this may encourage the developers to add such a correction to their procedures.

Recommendations for Trial Reporting

Investigators reporting on GRTs are encouraged to report their reasons for choosing group randomization; separate eligibility criteria, sampling schemes, and informed consent procedures for groups and members; justification for their sample size; ICC or variance component estimates from the analysis of intervention effects; and details of the analysis methods and software used.[1,16,29,40]

CONCLUSION

The purpose of this article has been to review the methodological developments from the past 5 years regarding the design and analysis of GRTs. The sheer volume of work is quite remarkable, and while every effort was made to provide a thorough review based on extensive searches of electronic databases and other sources, there are no doubt relevant papers that we did not include. Nonetheless, this review makes clear that there are valid methods that are readily available and well documented for the design and analysis of GRTs. We hope that this review will help investigators familiarize themselves with these methods and encourage them to collaborate with methodologists who can use these developments to strengthen the design and analysis of their trials.

Certainly, the methods required for GRTs are not as simple as those required for RCTs, and this is unfortunate. As noted 5 years ago, however:

> Whenever the investigator wants to evaluate an intervention that operates at a group level, manipulates the social or physical environment, or cannot be delivered to individuals, a group-randomized trial design is the best comparative design available.[1(p15)]

When that text appeared in 1998, it attempted to address the question of how to conduct GRTs well. Clearly the developments of the past 5 years have made it even easier to conduct GRTs well, and we simply must do a better job of taking advantage of these developments. ■

About the Authors

David M. Murray and Jonathan L. Blitstein are with the Department of Psychology, College of Arts and Sciences, University of Memphis, Memphis, Tenn. Sherri P. Varnell is with Northrop-Grumman Mission Systems, Atlanta, Ga.

Requests for reprints should be sent to David M. Murray, PhD, 3693 Norriswood, 202 Psychology Bldg, Memphis, TN 38134 (e-mail: d.murray@mail.psyc.memphis.edu).

This article was accepted September 12, 2003.

Contributors

D.M. Murray wrote the first and final drafts of the article. S.P. Varnell and J.L. Blitstein located many of the reviewed articles and edited several versions of the article.

Acknowledgments

We wish to acknowledge helpful comments from Zideng Feng, Fred Hutchinson Cancer Research Center; Barry Graubard, National Cancer Institute; Peter Hannan, University of Minnesota; Donald Hedeker, University of Illinois, Chicago; Stephen Raudenbush, University of Michigan; Oliver Schabenberger, SAS Institute Inc; and Alexander Wagenaar, University of Minnesota.

References

1. Murray DM. *Design and Analysis of Group-Randomized Trials.* New York, NY: Oxford University Press Inc; 1998.

2. Kish L. *Survey Sampling.* New York, NY: John Wiley & Sons Inc; 1965.

3. Zucker DM. An analysis of variance pitfall: the fixed effects analysis in a nested design. *Educ Psychol Meas.* 1990;50:731–738.

4. Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol.* 1978;108:100–102.

5. Campbell MK, Grimshaw JM. Cluster randomised trials: time for improvement. *BMJ.* 1998;317:1171–1172.

6. Atienza AA, King AC. Community-based health intervention trials: an overview of methodological issues. *Epidemiol Rev.* 2002;24:72–79.

7. Kerry SM, Bland JM. Analysis of a trial randomised in clusters. *BMJ.* 1998;316:54.

8. Kirkwood BR, Cousens SN, Victora CG, de Zoysa I. Issues in the design and interpretation of studies to evaluate the impact of community-based interventions. *Trop Med Int Health.* 1997;2:1022–1029.

9. Campbell MK, Mollison J, Steen N, Grimshaw JM, Eccles M. Analysis of cluster randomized trials in primary care: a practical approach. *Fam Pract.* 2000;17:192–196.

10. Donner A. Some aspects of the design and analysis of cluster randomization trials. *Appl Stat.* 1998;47:95–113.

11. Carvajal SC, Baumler E, Harrist RB, Parcel GS. Multilevel models and unbiased tests for group based interventions: examples from the Safer Choices Study. *Multivariate Behav Res.* 2001;36:185–205.

12. Kenny DA, Mannetti L, Pierro A, Livi S, Kashy DA. The statistical analysis of data from small groups. *J Pers Soc Psychol.* 2002;83:126–137.

13. Kerry SM, Bland JM. Trials which randomize practices I: how should they be analysed? *Fam Pract.* 1998;15:80–83.

14. Bloom HS, Bos JM, Lee S-W. Using cluster random assignment to measure program impacts: statistical implications for the evaluation of education programs. *Eval Rev.* 1999;23:445–469.

15. Altman DG. Statistics in medical journals: some recent trends. *Stat Med.* 2000;19:3275–3289.

16. Klar N, Donner A. Current and future challenges in the design and analysis of cluster randomization trials. *Stat Med.* 2001;20:3729–3740.

17. Feng Z, Diehr P, Peterson A, McLerran D. Selected statistical issues in group randomized trials. *Annu Rev Public Health.* 2001;22:167–187.

18. Feng Z, Thompson B. Some design issues in a community intervention trial. *Control Clin Trials.* 2002;23:431–449.

19. Bland JM. Sample size in guidelines trials. *Fam Pract.* 2000;17(suppl):S17–S20.

20. Kerry SM, Bland JM. Trials which randomize practices II: sample size. *Fam Pract.* 1998;15:84–87.

21. Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. *Int J Epidemiol.* 1999;28:319–326.

22. Resnicow K, Braithwaite R, Dilorio C, Vaughan R, Cohen MI, Uhl GA. Preventing substance use in high risk youth: evaluation challenges and solutions. *J Primary Prev.* 2001;21:399–415.

23. Zucker DM. Design and analysis of cluster randomization trials. In: Geller N, ed. *Advances in Clinical Trials Biostatistics.* New York, NY: Marcel Dekker Inc; 2003.

24. Murray DM. Statistical models appropriate for designs often used in group-randomized trials. *Stat Med.* 2001;20:1373–1385.

25. Reed JF. Eliminating bias in randomized cluster trials with correlated binomial outcomes. *Comput Methods Programs Biomed.* 2000;61:119–123.

26. Brown CH, Liao J. Principles for designing randomized preventive trials in mental health: an emerging developmental epidemiology paradigm. *Am J Community Psychol.* 1999;27:673–710.

27. Hayes RJ, Alexander NDE, Bennett S, Cousens SN. Design and analysis issues in cluster-randomized trials of interventions against infectious diseases. *Stat Methods Med Res.* 2000;9:95–116.

28. Loeys T, Vansteelandt S, Goetghebeur E. Accounting for correlation and compliance in cluster randomized trials. *Stat Med.* 2001;20:3753–3767.

29. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research.* London, England: Arnold; 2000.

30. McCulloch CE, Searle SR. *Generalized, Linear and Mixed Models.* New York, NY: John Wiley & Sons Inc; 2001.

31. Brown H, Prescott R. *Applied Mixed Models in Medicine.* Chichester, England: John Wiley & Sons Inc; 1999.

32. Raudenbush SW, Bryk AS. *Hierarchical Linear Models.* 2nd ed. Thousand Oaks, Calif: Sage Publications; 2002.

33. Kreft I, De Leeuw J. *Introducing Multilevel Modeling.* London, England: Sage Publications; 1998.

34. Murray DM, Blitstein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. *Eval Rev.* 2003;27:79–103.

35. Feng Z, Diehr P, Yasui Y, Evans B, Beresford S, Koepsell TD. Explaining community-level variance in group randomized trials. *Stat Med.* 1999;18:539–556.

36. Murray DM, Short BJ. Intraclass correlation among measures related to alcohol use by school aged adolescents: estimates, correlates, and applications in intervention studies. *J Drug Educ.* 1996;26:207–230.

37. Murray DM, Short BJ. Intraclass correlation among measures related to alcohol use by young adults: estimates, correlates and applications in intervention studies. *J Stud Alcohol.* 1995;56:681–694.

38. Murray DM, Short BJ. Intraclass correlation among measures related to tobacco use by adolescents: estimates, correlates, and applications in intervention studies. *Addict Behav.* 1997;22:1–12.

39. Murray DM, Clark MH, Wagenaar AC. Intraclass correlations from a community-based alcohol prevention study: the effect of repeat observations on the same communities. *J Stud Alcohol.* 2000;61:881–890.

40. Elbourne DR, Campbell MK. Extending the CONSORT statement to cluster randomized trials: for discussion. *Stat Med.* 2001;20:489–496.

41. Slymen DJ, Hovell MF. Cluster versus individual randomization in adolescent tobacco and alcohol studies: illustrations for design decisions. *Int J Epidemiol.* 1997;26:765–771.

42. Murray DM, Feldman HA, McGovern PG. Components of variance in a group-randomized trial analyzed via a random-coefficients model: the REACT Trial. *Stat Methods Med Res.* 2000;9:117–133.

43. Kerry SM, Bland JM. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. *Stat Med.* 2001;20:377–390.

44. Lake S, Kaumann E, Klar N, Betensky R. Sample size re-estimation in cluster randomization trials. *Stat Med.* 2002;21:1337–1350.

45. Liu A, Shih WJ, Gehan E. Sample size and power determination for clustered repeated measurements. *Stat Med.* 2002;21:1787–1801.

46. Raudenbush SW. Statistical analysis and optimal design in cluster randomized trials. *Psychol Methods.* 1997;2:173–185.

47. Varnell S, Murray DM, Janega JB, and Blitstein BL. Design and analysis of group-randomized trials: a review of recent practices. *Am J Public Health.* 2004;94:393–399.

48. Klar N, Donner A. The merits of matching in community intervention trials: a cautionary tale. *Stat Med.* 1997;16:1753–1764.

49. Thompson SG. The merits of matching in community intervention trials: a cautionary tale [letter]. *Stat Med.* 1998;17:2147–2152.

50. Raab GM, Butcher I. Balance in cluster randomized trials. *Stat Med.* 2001;20:351–365.

51. Varnell SP, Murray DM, Baker WL. An evaluation of analysis options for the one group per condition design: can any of the alternatives overcome the problems inherent in this design? *Eval Rev.* 2001;25:440–453.

52. Whiting-O'Keefe QE, Henke C, Simborg DW. Choosing the correct unit of analysis in medical care experiments. *Med Care.* 1984;22:1101–1114.

53. Roberts C. The implications of variation in outcome between health professionals for the design and analysis of randomized controlled trials. *Stat Med.* 1999;18:2605–2615.

54. Schnurr PP, Friedman MJ, Lavori PW, Hsieh FY. Design of Department of Veterans Affairs Cooperative Study No. 420: group treatment of posttraumatic stress disorder. *Control Clin Trials.* 2001;22:74–88.

55. Hoover DR. Clinical trials of behavioral interventions with heterogeneous teaching subgroup effects. *Stat Med.* 2002;21:1351–1364.

56. Varnell S, Murray DM, Hannan PJ, Baker WL. Intraclass correlation at the level of the unit of intervention in a randomized clinical trial: implications for analysis. Paper presented at: Annual Meeting of the American Evaluation Association, November 7–10, 2001, St. Louis, Mo.

57. Harville DA. Maximum likelihood approaches to variance component estimation and to related problems. *J Am Stat Assoc.* 1977;72:320–338.

58. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika.* 1986;73:13–22.

59. Zeger SL, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics.* 1986;42:121–130.

60. Thornquist MD, Anderson GL. Small sample properties of generalized estimating equations in group-randomized designs with Gaussian response. Paper presented at: 120th Annual Meeting of the

American Public Health Association, October 8–12, 1992, Washington, DC.

61. Feng Z, McLerran D, Grizzle J. A comparison of statistical methods for clustered data analysis with Gaussian error. *Stat Med.* 1996;15:1793–1806.

62. Murray DM, Hannan PJ, Baker WL. A Monte Carlo study of alternative responses to intraclass correlation in community trials: is it ever possible to avoid Cornfield's penalties? *Eval Rev.* 1996;20:313–337.

63. Emrich LJ, Piedmonte MR. On some small sample properties of generalized estimating equation estimates for multivariate dichotomous outcomes. *J Stat Computation Simulation.* 1992;41:19–29.

64. Lipsitz SR, Fitzmaurice GM, Orav EJ, Laird NM. Performance of generalized estimating equations in practical situations. *Biometrics.* 1994;50:270–278.

65. MacKinnon JG, White H. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *J Econometrics.* 1985;29: 305–325.

66. Murray DM, Hannan PJ, Wolfinger RD, Baker WL, Dwyer JH. Analysis of data from group-randomized trials with repeat observations on the same groups. *Stat Med.* 1998;17:1581–1600.

67. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics.* 2001;57:126–134.

68. Bellamy SL, Gibberd R, Hancock L, et al. Analysis of dichotomous outcome data for community intervention studies. *Stat Methods Med Res.* 2000;9:135–159.

69. Long JS, Ervin LH. Using heteroscedasticity consistent standard errors in the linear regression model. *Am Statistician.* 2000;54:217–224.

70. Corcoran C, Ryan L, Senchaudhuri P, Mehta C, Patel N, Molenberghs G. An exact trend test for correlated binary data. *Biometrics.* 2001;57:941–948.

71. Fay M, Graubard P. Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics.* 2001;57:1198–1206.

72. Kauermann G, Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. *J Am Stat Assoc.* 2001;96:1387–1396.

73. Pan W, Wall MM. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Stat Med.* 2002;21:1429–1441.

74. Bell RM, McCaffrey DF. Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology.* 2002;28:169–181.

75. Preisser JS, Young ML, Zaccaro DJ, Wolfson M. An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. *Stat Med.* 2003;22:1235–1254.

76. Rodriguez G, Goldman N. An assessment of estimation procedures for multilevel models with binary responses. *J R Stat Soc.* 1995;158:73–89.

77. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc.* 1993;88:9–25.

78. Ten Have TR, Kunselman A, Zharichenko E. Accommodating negative intracluster correlation with a mixed effects logistic model for bivariate binary data. *J Biopharm Stat.* 1998;8:131–149.

79. Hannan PJ, Murray DM. Gauss or Bernoulli? A Monte Carlo comparison of the performance of the linear mixed model and the logistic mixed model analyses in simulated community trials with a dichotomous outcome variable at the individual level. *Eval Rev.* 1996;20:338–352.

80. Gibbons RD, Hedeker D. Random effects probit and logistic regression models for three-level data. *Biometrics.* 1997;53:1527–1537.

81. Aitkin M. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics.* 1999;55:117–128.

82. Kleinman KP, Ibrahim JG. A semi-parametric Bayesian approach to generalized linear mixed models. *Stat Med.* 1998;17:2579–2596.

83. Ten Have TR, Localio AR. Empirical Bayes estimation of random effects parameters in mixed effects logistic regression models. *Biometrics.* 1999;55: 1022–1029.

84. Turner RM, Omar RZ, Thompson SG. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Stat Med.* 2001;20:453–472.

85. Hedeker D, Siddiqui O, Hu FB. Random-effects regression analysis of correlated grouped-time survival data. *Stat Methods Med Res.* 2000;9:161–179.

86. Ross EA, Moore D. Modeling clustered, discrete, or grouped time survival data with covariates. *Biometrics.* 1999;55:813–819.

87. Segal MR, Neuhaus JM, James IR. Dependence estimation for marginal models of multivariate survival data. *Lifetime Data Analysis.* 1997;3:251–268.

88. Cai T, Cheng SC, Wei LJ. Semiparametric mixed-effects models for clustered failure time data. *J Am Stat Assoc.* 2002;95:514–522.

89. Gray RJ, Li Y. Optimal weight functions for marginal proportional hazards analysis of clustered failure time data. *Lifetime Data Analysis.* 2002;8:5–19.

90. Sargent DJ. A general framework for random effects survival analysis in the Cox proportional hazards setting. *Biometrics.* 1998;54:1486–1497.

91. Vaida F, Xu R. Proportional hazards model with random effects. *Stat Med.* 2000;19:3309–3324.

92. Yau KK. Multilevel models for survival analysis with random effects. *Biometrics.* 2001;57:96–102.

93. Lui K-J, Mayer JA, Eckhardt L. Confidence intervals for the risk ratio under cluster sampling based on the beta-binomial model. *Stat Med.* 2000;19: 2933–2942.

94. Bennett S, Parpia T, Hayes R, Cousens S. Methods for the analysis of incidence rates in cluster randomized trials. *Int J Epidemiol.* 2002;31:839–846.

95. Gail MH, Byar D, Pechacek TF, Corle D. Aspects of statistical design for the Community Intervention Trial for Smoking Cessation (COMMIT). *Control Clin Trials.* 1992;13:6–21.

96. COMMIT Research Group. Community Intervention Trial for Smoking Cessation (COMMIT): I. Cohort results from a four-year community intervention. *Am J Public Health.* 1995;85:183–192.

97. COMMIT Research Group. Community Intervention Trial for Smoking Cessation (COMMIT): II. Changes in adult cigarette smoking prevalence. *Am J Public Health.* 1995;85:193–200.

98. Gail MH, Mark SD, Carroll RJ, Green SB, Pee D. On design considerations and randomization-based inference for community intervention trials. *Stat Med.* 1996;15:1069–1092.

99. Braun T, Feng Z. Optimal permutation tests for the analysis of group randomized trials. *J Am Stat Assoc.* 2001;96:1424–1432.

100. LaVange LM, Koch GG, Schwartz TA. Applying sample survey methods to clinical trials data. *Stat Med.* 2001;20:2609–2623.

101. Korn EL, Graubard BI. *Analysis of Health Surveys.* New York, NY: John Wiley & Sons Inc; 1999.

102. Muthen BO. Beyond SEM: general latent variable modeling. *Behaviormetrika.* 2002;29:81–117.

103. Schulenberg J, Maggs JL. Moving targets: modeling developmental trajectories of adolescent alcohol misuse, individual and peer risk factors, and intervention effects. *Appl Dev Sci.* 2001;5:237–253.

104. Muthen BO, Curran PJ. General longitudinal modeling of individual differences in experimental designs: a latent variable framework for analysis and power estimation. *Psychol Methods.* 1997;2:371–402.

105. Curran PJ, Muthen BO. The application of latent curve analysis to testing developmental theories in intervention research. *Am J Community Psychol.* 1999;27: 567–595.

106. Hser Y-I, Shen H, Chuang C-P, Messer SC, Anglin MD. Analytic approaches for assessing long-term treatment effects. *Eval Rev.* 2001;25:233–262.

107. Davidian M, Giltinan DM. *Nonlinear Models for Repeated Measurement Data.* London, England: Chapman & Hall; 1995.

108. Vonesh EF, Chinchilli VM. *Linear and Nonlinear Models for the Analysis of Repeated Measurements.* New York, NY: Marcel Dekker; 1997.

109. Gruenewald PJ. Analysis approaches to community evaluation. *Eval Rev.* 1997;21:209–230.

110. Biglan A, Ary D, Wagenaar AC. The value of interrupted time-series experiments for community intervention research. *Prev Sci.* 2000;1:31–49.

111. Krull J, MacKinnon DP. Multilevel mediation modeling in group-based intervention studies. *Eval Rev.* 1999;23:418–444.

112. MacKinnon DP, Taborga MP, Morgan-Lopez AA. Mediation designs for tobacco prevention research. *Drug Alcohol Depend.* 2002;68:S69–S83.

113. Yi GY, Cook RJ. Marginal methods for incomplete longitudinal data arising in clusters. *J Am Stat Assoc.* 2002;97:1071–1080.

114. Hunsberger S, Murray DM, Davis CE, Fabsitz R. Imputation strategies for missing data in a school based multicenter study: the Pathways study. *Stat Med.* 2001; 20:305–316.

115. Zhou Z-H, Perkins AJ, Hui SL. Comparisons of software packages for generalized linear multilevel models. *Am Statistician.* 1999;53:282–290.

116. Bryk AS, Raudenbush SW. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, Calif: Sage Publications; 1992.

117. *SAS/STAT User's Guide, Version 8.* Cary, NC: SAS Institute Inc; 1999.

118. Littell RC, Milliken GA, Stroup WW, Wolfinger RD. *SAS System for MIXED Models.* Cary, NC: SAS Institute Inc; 1996.

119. Singer J. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *J Educ Behav Stat.* 1998;24:322–354.

120. Wolfinger RD. Fitting nonlinear mixed models with the new NLMIXED procedure. Paper presented

at: 24th Annual SAS Users Group International Conference, April 1999, Miami, Fla.

121. *SAS/STAT Software: Changes and Enhancements, Release 8.1.* Cary, NC: SAS Institute Inc; 2000.

122. Hedeker D, Gibbons RD. MIXOR: a computer program for mixed-effects ordinal regression analysis. *Comput Methods Programs Biomed.* 1996;49:157–176.

123. Hedeker D, Gibbons RD. MIXREG: a computer program for mixed-effects regression analysis with autocorrelated errors. *Comput Methods Programs Biomed.* 1996;49:229–252.

124. Hedeker D. MIXNO: a computer program for mixed-effects nominal logistic regression. *J Stat Software.* 1999;4(5):1–92.

125. Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. *Biometrics.* 1994;50:933–944.

126. Hedeker D. A mixed-effects multinomial logistic regression model. *Stat Med.* 2003;22:1433–1446.

127. Goldstein H, Browne W, Rasbash J. Multilevel modelling of medical data. *Stat Med.* 2002;21: 3291–3315.

128. MIXED. In: *SPSS 11.0 Syntax Reference Guide.* Chicago, Ill: SPSS Inc; 2002:136–151.

129. Baskerville NB, Hogg W, Lemelin J. The effect of cluster randomization on sample size in prevention research. *J Fam Pract.* 2001;50:242–246.

130. Campbell MK, Mollison J, Grimshaw JM. Cluster trials in implementation research: estimation of intracluster correlation coefficients and sample size. *Stat Med.* 2001;20:391–399.

131. Piaggio G, Carroli G, Villar J, et al. Methodological considerations on the design and analysis of an equivalence stratified cluster randomization trial. *Stat Med.* 2001;20:401–416.

132. Smeeth L, Ng ES-W. Intraclass correlation coefficients for cluster randomized trials in primary care: data from the MRC trial of the assessment and management of older people in the community. *Control Clin Trials.* 2002;23:409–421.

133. Gulliford MC, Ukoumunne OC, Chinn S. Components of variance and intraclass correlations for the design of community-based surveys and intervention studies. *Am J Epidemiol.* 1999;149:876–883.

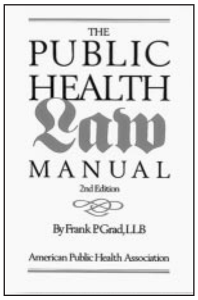
134. Scheier LM, Griffin KW, Doyle MM, Botvin GJ. Estimates of intragroup dependence for drug use and skill measures in school-based drug abuse prevention trials: an empirical study of three independent samples. *Health Educ Behav.* 2002;29:83–101.

135. Murray DM, Phillips GA, Birnbaum AS, Lytle LA. Intraclass correlation for measures from a school-based nutrition intervention study: estimates, correlates and applications. *Health Educ Behav.* 2001;28:666–679.

136. Murray DM, Alfano CM, Zbikowski SM, Padgett LS, Robinson LA, Klesges R. Intraclass correlation among measures related to cigarette use by adolescents: estimates from an urban and largely African American cohort. *Addict Behav.* 2002;27:509–527.

137. Lazovich D, Murray DM, Brosseau LM, Parker DL, Milton FT, Dugan SK. Sample size considerations for studies of intervention efficacy in the occupational setting. *Ann Occup Hyg.* 2002;46:219–227.

138. Martinson BC, Murray DM, Jeffery RW, Hennrikus DJ. Intraclass correlation for measures from a worksite health promotion study: estimates, correlates and applications. *Am J Health Promotion.* 1999;13:347–357.



THE
PUBLIC
HEALTH
Law
MANUAL
2nd Edition
By Frank P. Grad, LLB
American Public Health Association