Essential Ingredients and Innovations in the Design and Analysis of Group-Randomized Trials

David M. Murray,1 Monica Taljaard,2,3 Elizabeth L. Turner,4 and Stephanie M. George1

1Office of Disease Prevention, National Institutes of Health, North Bethesda, Maryland 20892, USA; email: david.murray2@nih.gov, stephanie.george@nih.gov
2Clinical Epidemiology Program, Ottawa Hospital Research Institute, The Ottawa Hospital, Civic Campus, Ottawa, Ontario K1Y 4E9, Canada; email: mtaljaard@ohri.ca
3School of Epidemiology and Public Health, University of Ottawa, Ottawa, Ontario K1Y 4E9, Canada
4Department of Biostatistics and Bioinformatics, and Duke Global Health Institute, Duke University, Durham, North Carolina 27710, USA; email: liz.turner@duke.edu

Keywords

group-randomized trial, cluster-randomized trial, individually randomized group treatment trial, stepped-wedge group-randomized trial, stepped-wedge cluster-randomized trial

Abstract

This article reviews the essential ingredients and innovations in the design and analysis of group-randomized trials. The methods literature for these trials has grown steadily since they were introduced to the biomedical research community in the late 1970s, and we summarize those developments. We review, in addition to the group-randomized trial, methods for two closely related designs, the individually randomized group treatment trial and the stepped-wedge group-randomized trial. After describing the essential ingredients for these designs, we review the most important developments in the evolution of their methods using a new bibliometric tool developed at the National Institutes of Health. We then discuss the questions to be considered when selecting from among these designs or selecting the traditional randomized controlled trial. We close with a review of current methods for the analysis of data from these designs, a case study to illustrate each design, and a brief summary.
INTRODUCTION

Randomized trials are the gold standard in biomedical research because they provide the strongest evidence for causal inference by minimizing the risk of bias in the assignment of participants to treatments. This article reviews the development of methods for the design and analysis of group-randomized trials (GRTs). It also reviews developments for two closely related designs, the individually randomized group treatment (IRGT) trial and the stepped-wedge group-randomized trial (SW-GRT). Their key features are described in the Essential Ingredients section. The most important reports for the development of these methods are described in the Evolution of Methods for Design and Analysis section. The issues that guide choices among these three designs and the more familiar individually randomized controlled trial (RCT) are identified in the Decision Points section. The analytic methods that are now employed with these designs are described in the Current Analytic Strategies section. Illustrative examples are provided in the Case Studies section. We close with a brief summary.

ESSENTIAL INGREDIENTS

Group-Randomized Trials

A GRT differs from the RCT by randomizing groups rather than individuals to study conditions, with outcomes measured in participants from each group (105). In the traditional GRT considered here, there is no crossover of groups to different study conditions. Also called a cluster-randomized trial (16, 29, 33, 55), the GRT is the best comparative design available if there is a strong rationale for randomization of groups rather than individuals, such as concern for contamination or use of a group-level intervention.

The key feature of the GRT is the randomization of groups to study conditions. Outcomes on participants from the same group are likely to be correlated as a result of shared experience, common exposures, or participant interaction (83). This correlation violates the independence of errors assumption that underlies the analytic methods for RCTs, requiring methods that are appropriate to the nested design (16, 29, 33, 55, 105). The correlation is often measured by the intraclass or intracluster correlation (ICC); we use ICC to refer quite generally to this within-group correlation.

Murray et al. (111) characterized the design characteristics for recent GRTs involving cancer or cancer-related outcomes. Most compared two study conditions. Most employed a pretest-posttest design, though some used a posttest-only design and others included multiple pretest and/or posttest measurements. Most employed a cohort design observing the same participants from each group at each measurement occasion; others employed a cross-sectional design observing different participants from each group at each measurement occasion. Most employed some form of restricted randomization, including stratification, matching, or constrained randomization.

In spite of the abundance of materials on the design and analytic methods for GRTs, state-of-the-practice reviews have repeatedly shown that a large proportion of GRTs fail to account for the ICC both in sample size calculations and in the analysis, in spite of the adverse consequences for power and type 1 errors (e.g., 2, 13, 25, 26, 28, 32, 34, 40, 43, 69, 71, 75, 76, 100, 110, 111, 118, 123, 126, 129, 140, 145).

Individually Randomized Group Treatment Trials

An IRGT differs from the RCT in that ICC is induced in IRGTs by the method of intervention delivery (116). ICC can happen if participants receive some portion of their intervention in a
group format (e.g., exercise class), if multiple participants share the same intervention agent (e.g., the same instructor, therapist, or surgeon), or if participants interact in some other way related to the method of intervention delivery (e.g., through a virtual chat room created for participants in the same study condition). The IRGT is the best comparative design available if randomization of individuals is possible, but ICC is nevertheless induced by features of the method of intervention delivery.

The key feature of the IRGT is that the method of intervention delivery leads to correlation among outcomes taken on groups of participants in the same study condition, creating the same type of ICC seen in GRTs. Investigators must account for the ICC in the sample size to avoid low power and in the analysis to avoid type 1 errors (5, 12, 78, 116, 124, 141). The situation is even more complex in IRGTs if the method of intervention delivery creates multiple overlapping groups or if the group structure changes over time (3, 8, 125).

The methods literature for IRGTs is much more limited than for GRTs, and the issues inherent in this design are poorly understood. Recent reviews suggest that most investigators who employ the IRGT design are not aware of these issues and do not use appropriate methods for sample size or analysis (6, 89, 116, 117).

**Stepped-Wedge Group-Randomized Trials**

An SW-GRT is a type of GRT that has seen increasing use over the last decade and is now commonly used to evaluate service delivery interventions (73). However, SW-GRTs are more complex than GRTs or IRGTs and face greater risks for bias.

The key feature of the SW-GRT is the crossover of all groups from the intervention condition to the control condition in a random order and on a staggered schedule. As in other GRTs, the observations from participants from the same group will be correlated; however, the impact of the ICC is reduced in the SW-GRT compared with that in the GRT or IRGT because groups are crossed with rather than nested within study conditions. At the same time, the intervention effect is confounded with calendar time by design (61, 73). Moreover, the effect of the intervention may vary depending on how much time has passed since the intervention was introduced (61, 72). Finally, the pattern of correlation over time can be complex (80) because SW-GRTs involve repeated measurements on the same groups and sometimes on the same participants.

Copas et al. (19) recently summarized the design characteristics used in SW-GRTs. In a complete SW-GRT, data are collected initially when all groups are in the control condition, again in each group at each crossover point or step, and often again after all groups are in the intervention condition; in an incomplete SW-GRT, data are not collected from all groups at all steps. SW-GRTs vary in the number of groups that cross over in each step, the number of steps, and the time between crossovers. They may employ restricted randomization, as described above, to balance groups in sets to be randomized to the crossover schedule. In the continuous recruitment short exposure design, participants are recruited over time and exposed for a short period; participants may be measured only once or repeatedly. In the closed cohort design, participants are identified at the beginning of the study, participate throughout, and are measured repeatedly. In the open cohort design, many participants are identified at the beginning, but some may leave, whereas other participants are recruited over time; participants may be measured only once or multiple times.

The methods literature on SW-GRTs has developed rapidly over the last decade. Reviews have noted deficiencies in the reporting of study design, sample size, analytic methods, and ethical conduct (7, 10, 14, 22, 52, 86, 98, 101, 133).
EVOLUTION OF METHODS FOR DESIGN AND ANALYSIS

A primary objective of this article is to identify the most influential reports in the evolution of methods for the design and analysis of GRTs, IRGTs, and SW-GRTs. Toward that end, PubMed was searched to identify all methods papers related to GRTs, IRGTs, and SW-GRTs (search parameters are included at the end of Supplemental Table 1). The results were augmented by articles, books, chapters, and other reports known to the authors. PubMed was then searched for other papers by any of the authors of the identified reports, yielding a total of 4,514 candidate items. The lead author reviewed each title and excluded items that were not focused on design or analytic methods for GRTs, IRGTs, or SW-GRTs, leaving 926 items; 799 focused on GRTs, 49 on IRGTs, and 74 on SW-GRTs, and 4 addressed all three designs.

The relative citation ratio (RCR) and citation counts were used to assess the influence of each item. The RCR is a metric that uses relative citation rates to measure influence at the article level (74), standardized across fields of study. The RCR was available for 791 items; for the remaining 135 items, the Web of Science, Scopus, and Google Scholar were used to obtain citation counts. GRT items \((N = 25)\) with an RCR \(\geq 7.98\) (99th percentile for all PubMed entries) or without an RCR but with \(\geq 200\) citations\(^1\) were retained. IRGT \((N = 9)\) items and items addressing all three designs \((N = 4)\) with an RCR \(\geq 3.45\) (95th percentile) or without an RCR but with \(\geq 100\) citations were retained. SW-GRT items \((N = 13)\) with an RCR \(\geq 4.91\) (97.50 percentile) or without an RCR but with \(> 150\) citations were retained. This approach provided 50 methods reports related to these designs that were deemed influential; unless otherwise noted,\(^2\) the items cited in this section comprise the set of 50 influential reports. Supplemental Table 1 presents the full list of 926 items, ranked within design by RCR and citations, and highlights the 50 items identified as influential.

Although these items were judged to be influential, they represent the history of the development of these methods and not necessarily the current state of the science. We summarize the state of the science for analysis in the Current Analytic Strategies section.

**Items Relevant to All Three Designs**

In 1978, Cornfield (20) identified the two penalties associated with group randomization: extra variation associated with the group and limited degrees of freedom for the test of the intervention effect. These two penalties must be addressed in the design and analysis of any GRT, IRGT, or SW-GRT.

Turner et al. (137, 138) published reviews of the methods for design and analysis of GRTs, IRGTs, and SW-GRTs in 2017.

**Group-Randomized Trials**

Several influential papers addressed general issues for GRTs. Those included a 1997 commentary drawing attention to the methodological issues inherent in GRTs (11), a widely cited 1997 paper on optimal design (121), a 2003 paper on the use of GRTs for evaluating the effectiveness of change and improvement strategies (31), and a 2013 paper on methods for process evaluation (50).

Others presented methods for sample size calculations. Donner et al. (27), Murray & Hannan (106), Gulliford et al. (53), Adams et al. (1), and Hedges & Hedberg (59) presented methods based on the ICC, while Hayes & Bennett (54) presented methods based on the coefficient of variation.

---

\(^1\)The citation count thresholds generally discriminated between the items that fell above and below the RCR threshold for each design. The sliding scale reflected the number of items for each design.

\(^2\)An asterisk indicates that the citation was not among the 50 influential reports.
Others focused on specific analytic methods. Those included methods for random coefficient or growth curve models (38), methods for binary data (120), methods for ordinal data (57, 58), methods for Poisson regression (148), methods for meta-analysis (30), and methods for mediation analysis (36, 87, 88).

Murray (105) published the first textbook on the design and analysis of GRTs followed two years later by the text from Donner & Klar (29). Subsequent textbooks were published by Hayes & Moulton (55) and Eldridge & Kerry (33).

In 2004, the first CONSORT statement on GRTs appeared (17), providing a checklist to identify the methodological information to include in trial reports that is deemed essential to interpreting the trial. An update was published in 2012 (18).

Also in 2004, Murray et al. (113) published a review of methodological issues in GRTs, summarizing work on both design and analytic methods.

Individually Randomized Group Treatment Trials

Several influential reports have addressed the risk of type 1 errors in studies in which participants receive their treatment in a group format or from a shared interventionist. These papers have appeared in the biomedical (78, 143), psychological (6, 21, 99), and educational literature (115).

In 2005, Roberts & Roberts (124) addressed the analytic challenges specific to IRGTs. The investigators noted that mixed models that specified the groups used to deliver the intervention as levels of a random effect provided an appropriate analysis.

In 2008, Boutron et al. (12) extended the CONSORT statement to nonpharmacologic interventions, which include many IRGTs. They pointed to the need to provide details on how the intervention is delivered (e.g., individually, in groups, via a common interventionist) and to address the implications for analysis of having nonindependent observations within one or more study conditions.

In 2017, Heo et al. (68) described sample size methods for IRGTs, focusing on partially nested designs, which reflect a common situation in an IRGT in which participants in the intervention condition receive their treatment in a group format while participants in the control condition do not. Also in 2017, Sterba (132) reviewed modeling developments for a variety of IRGT designs.

Stepped-Wedge Group-Randomized Trials

Hussey & Hughes (73) first described methods for the SW-GRT design in 2007. Copas et al. (19) later identified three main types of SW-GRTs and discussed design choices with regard to the number and length of the steps, incomplete and complete designs, and randomization methods, including restricted randomization methods. Hemming et al. (61) outlined the rationale, design, analysis, and reporting of SW-GRTs. They noted that the SW-GRT is particularly well suited for evaluations of health service delivery interventions.

Several influential papers have addressed sample size methods for SW-GRTs. Woertman et al. (144) described an approach based on a design effect and noted that the sample size will depend on group size, the ICC, the number of steps, the number of baseline measurements, and the number of measurements between steps; a subsequent letter (60) corrected an important error in this paper, which was accepted by the authors (23). Baio et al. (4) presented simulation methods for sample size estimation. Hemming et al. (63) considered power for several different SW-GRT designs, including incomplete cross-sectional SW-GRTs, SW-GRTs with multiple levels of nesting, and complete SW-GRTs. Girling & Hemming (45) proposed an algorithm to optimize the design of SW-GRTs and showed that for large studies the best design may be a hybrid of GRT and
SW-GRT design components. Hemming & Taljaard (64) compared power for GRTs and SW-GRTs when the number of groups is fixed, and they reported that the GRT tends to be more efficient when the ICC is small and that the SW-GRT tends to be more efficient when the ICC is large, dependent on group size.

Scott et al. (128) addressed the use of small-sample corrected generalized estimating equations (GEE) for SW-GRTs. They demonstrated the viability of a marginal model in studies having at least 10 groups.

Several state-of-the-practice reviews of SW-GRTs have been published (10, 14, 22, 101), reporting wide variation in data analytic methods and reporting standards. The recent CONSORT statement for SW-GRTs (67) provides reporting guidelines.

More General Material

Though not focused specifically on GRTs, IRGTs, or SW-GRTs, and so not included in the set of 50 influential reports, several well-known books have contributed substantially to the development of design and analytic methods for these designs. In 1953, Lindquist (94) published the first design and analytic text that addressed nested designs. In 1987, Goldstein (46) published the first text on multilevel models for educational and social research, including methods that have come to be used widely in GRTs and IRGTs; subsequent editions have elaborated on those methods (47–49). In 1992, Bryk & Raudenbush (15) published their text on hierarchical linear models, with a second edition in 2002 (122).

DECISION POINTS

Numerous issues guide the choice among the four designs under consideration here. Because the RCT and particularly the double-blinded RCT provide the strongest evidence for causal inference among the four designs, it should be adopted if possible (16, 29, 33, 55, 105). That said, an RCT is not always viable owing to the nature of the intervention or to factors beyond the investigator’s control (e.g., 102, 130), and so it is important to understand the alternatives and the decision points to be used in choosing one design over another.

In choosing a design, the investigator should carefully consider the research question, the context of the research, and the nature of the intervention. Figure 1 presents three key questions to guide the selection process.

The first question applies to all situations: Is there a strong rationale for randomizing groups rather than individuals to study conditions? Recognizing that the RCT provides the strongest evidence, the Ottawa Statement on the ethical design and conduct of GRTs calls for investigators to provide a clear rationale for choosing randomization of groups rather than individuals (134). Acceptable reasons include the evaluation of a group-level intervention or the group effects of an individual-level intervention; concern about substantial contamination if multiple study conditions are implemented in the same group; the need to reduce costs, enhance compliance, or secure cooperation of investigators; and administrative convenience.

Contamination is often the primary reason for choosing group randomization. Contamination occurs when participants who are not selected to receive the intervention receive at least some part of it. If the risk of contamination is low, individual randomization is preferred, as an RCT or IRGT will be more efficient than a GRT or SW-GRT. If the risk of contamination is not low, individual randomization is unwise and group randomization is preferred (18, 105, 134, 136). To illustrate, consider a school-based behavioral intervention to prevent cigarette smoking among adolescents. If individual students within a school are randomized to intervention and control,
Figure 1

Three questions (blue) to guide the selection of the appropriate study design. Footnotes: (a) If the intervention is delivered through a physical group or a virtual group, or through shared interventionists who each work with multiple participants, positive ICC can develop over the course of the trial. (b) There may be logistical reasons to randomize groups, or it may not be possible to deliver the intervention to individuals without substantial risk of contamination. (c) There may be legitimate political or logistical reasons to roll out the intervention to all groups before the end of the trial. Note: Quite generally, an RCT design is preferred to an IRGT design because it is both more efficient and less subject to threats to internal validity. A parallel-arm GRT design is preferred to an SW-GRT design because the SW-GRT is subject to additional threats to internal validity. The article text provides more details. Abbreviations: GRT, group-randomized trial; ICC, intraclass or intracenter correlation; IRGT, individually randomized group treatment; RCT, randomized controlled trial; SW-GRT, stepped-wedge group-randomized trial.

If there is no strong rationale for randomizing groups rather than individuals to study conditions, the next question is, Do participants receive their treatment in a group format or from a shared interventionist? Delivery of the intervention in a group format, or through intervention agents who each interact with multiple participants, usually leads to correlated outcomes because it creates the opportunity for shared experience, common exposures, or participant interaction. In the case of a blinded placebo-controlled RCT in a large population (e.g., blood pressure–lowering medication versus placebo), there is no reason to expect that correlation of outcomes might develop, and in this case, the RCT is preferred. In contrast, when the intervention is a group-based treatment [e.g., a group-therapy session for cardiac rehabilitation (131)] or involves other interactions among intervention participants (e.g., through an online community or through a shared interventionist), then the observations taken from participants who have such shared experiences, common exposures, or interactions are expected to become correlated. The trial is an IRGT trial, and the magnitude of the correlation will depend on the frequency and intensity of those experiences. This correlation must be accounted for to avoid a type 1 error (116, 124). Correlation of outcomes may arise in one study condition or in multiple study conditions.

If there is a strong rationale for randomizing groups, the next question is, Is there a strong rationale for rolling out the intervention to all groups before the end of the trial? There may be political or logistical reasons for an affirmative answer. For example, a health care system composed of multiple clinics may decide that a change should be made in the way that a particular health service is delivered. The health system may refuse to consider a GRT in which only half of the
clinics receive the intervention. If the rollout of the intervention to the clinics can be spread over
time, and if the health system agrees that the timing of the rollout can be allocated randomly, an
SW-GRT design can provide stronger evidence for causal inference compared with a nonrandom-
ized design with a staggered rollout (19, 112). Nevertheless, given that the SW-GRT confounds
time and study condition by design, and usually requires more time and more measurements, the
GRT design should be used if possible (61, 73, 112).

Consideration of the three questions in Figure 1 will allow the investigator to select the ap-
propriate design for the research question, given the context of the study and the nature of the
intervention. It is advisable to assemble a team of both substantive and methodological experts to
consider the issues raised here when choosing among the four general designs. Once that choice
is made, many more decisions will be required on details such as the measurement schedule, how
participants will be recruited, whether the same participants will be followed over time, the risk
of selection bias due to lack of blinding, etc. The reader is referred to excellent texts for a more
comprehensive description of these issues (16, 29, 33, 55, 103, 105).

CURRENT ANALYTIC STRATEGIES

Group-Randomized Trials

Three main analytical approaches can account for the ICC in GRTs: two-stage analysis, mixed-
effects regression, and GEE. In a two-stage analysis, the multiple observations from each group
are reduced in a first stage to a single score, such as a proportion or mean, and those group scores
are compared between study conditions in a second stage using standard methods such as a two-
sample t-test, Wilcoxon rank-sum test, or permutation test (44, 56, 108, 119, 147). The two-stage
analysis is a good choice when the number of groups is small because it yields robust inferences
as long as the group sizes are neither very small nor highly variable. A disadvantage is that it is
more difficult to adjust for individual-level covariates (16, 56). The more common approach is
to analyze individual-level data in one stage using a regression model that accounts for the ICC
(111). In mixed-effects regression, the groups are modeled as random effects; in GEE, there are
no random effects, but the correlation structure is modeled directly. Mixed-effects regression and
GEE can more easily adjust for individual-level and group-level covariates and for heterogeneity
in group size.

GEE coefficients describe changes in the population average effect given changes in the co-
variates, including the intervention. Inferences for the treatment effect can be made by relying on
the sandwich variance estimator, which is robust to misspecification of the correlation structure;
however, this approach requires a sufficient number of groups—preferably, at least 50 (39, 107).
However, many GRTs randomize fewer groups than 50 (111). Methods to compensate for small
numbers of groups have been developed (37, 42, 81, 93, 97, 104) but do not always perform well
and are not universally available across statistical software packages. Li & Redden (93) found that
their performance depends not only on the number of groups, but also on the variation in group
sizes. Ford & Westgate (42) proposed methods that outperform previous methods, but they call
for further study to provide a valid bias-corrected estimator that can be applied in any situation.

Mixed-effects coefficients describe changes in the individual given changes in the covariates,
conditional on the random effects. As such, they are group-specific coefficients (commonly re-
ferred to as subject-specific coefficients in the longitudinal analysis literature); they are identical
to the population-average coefficients from GEE in a linear or log model but different in a logistic
model. Mixed-effects regression accommodates a smaller number of groups than GEE, though
the two-stage approach is preferred for small studies. In addition, mixed-effects regression can
easily be extended to account for multiple levels in the data hierarchy. A challenge is specifying the
correct model that accounts for all relevant fixed and random effects (105); another is choosing an appropriate method for calculating the degrees of freedom for testing the intervention effect, especially when the number of groups is small (82, 127). Li & Redden (92) studied the performance of a variety of methods in the case of binary outcomes and recommended the between-within method, which was robust to variation in group size and provided correct type 1 error rates even when the total number of groups was as low as ten.

For binary outcomes, there is a choice of analyzing treatment effects using absolute risk, relative risk, or odds ratio (139). Reporting of both absolute and relative effect estimates is recommended (18), but models with binomial distributions and log or identity link functions do not always converge. A log-link with a Poisson distribution and robust standard errors may be a suitable alternative (148).

Obtaining reliable ICC estimates for binary outcomes can be difficult. Mixed-effects logistic regression does not yield ICC estimates on the proportions scale—which is the scale required for most sample size formulae—and there is no single formula for converting from the logit to the proportions scale (35). GEE has the advantage of producing ICC estimates on the proportions scale directly. Wu et al. (146) compared several methods of estimating the ICC for binary outcomes and found that estimates were quite different and often negatively biased. They recommended that sample size and analytic methods should allow the ICC to vary by study condition, rather than assume a constant ICC, as is commonly done.

Hooper et al. (70) provided recommendations for analyzing GRTs that provide a baseline measurement of the outcome. Comparing study conditions in terms of changes from baseline (known as difference in difference) is popular but can be inefficient and misleading. It is more efficient to treat baseline as a covariate using an analysis of covariance (ANCOVA) (84, 105). With more than two time points in the model, we recommend random coefficients models (109). As an alternative to specifying fixed effects for study condition, time, and their interaction, omitting the study condition main effect can increase precision; the model should include random effects for group and the interaction between group and time. When using ANCOVA for a cohort design, including both individual-level and group-level versions of the baseline measurement can increase power (84).

**Individually Randomized Group Treatment Trials**

IRGTs are typically analyzed using mixed-effects regression methods to allow for different correlation structures in the study conditions (5, 9, 90, 116, 124). The early work focused on the simplest IRGT design in which it is important to allow for ICC in the intervention condition but not in the control condition. More recent work has addressed both changes in the group structure over time and participants who receive their intervention in more than one small group at the same time. Building on the early work by Goldstein (46, 47), methodologists have developed both cross-classified (95, 96, 125, 132) and multiple-membership mixed-effects models (125). Bauer et al. (8) proposed dynamic-group models that reflect changes in the structure or function of groups over time as an alternative to Goldstein’s multiple-membership model.

**Stepped-Wedge Group-Randomized Trials**

SW-GRTs are more complicated to analyze and have increased risks of bias over parallel GRTs. By design, the intervention effect is confounded with time and so the analysis must adjust for time as a covariate. The most popular method of analysis for the cross-sectional SW-GRT is mixed-effects regression with a fixed time-varying indicator for study condition and a fixed categorical effect for time as well as a random intercept for group to account for the ICC (73). However, this model
assumes a constant correlation between any two individuals in the same group, no matter how far apart in time they are measured. A more realistic correlation structure would allow the strength of the correlation to decay with increasing time separation. Kasza & Forbes (79) showed that omitting a correlation decay in an SW-GRT with continuous outcomes when such decay exists will underestimate standard errors and increase the type I error rate. Kasza et al. (80) proposed a model that allows the ICC to decay exponentially across discrete periods of time and provided SAS code. Although this model does not seem to be possible in other major statistical software packages, and computational issues can arise, it is currently recommended for analyzing data from a cross-sectional SW-GRT. It can also be used for cohort designs by including an additional random intercept for the individual. Grantham et al. (51) examined even more flexible models that allow correlations to decay as a continuous function of time. Extensions allow treatment effects to vary across groups (65, 66, 72).

Alternative SW-GRT approaches exist that model fixed time effects using parametric or semiparametric curves and allow for time-varying intervention effects (41, 65, 72, 114); however, their properties have not been extensively examined. Scott et al. (128) examined GEE with small sample corrections for cohort SW-GRTs. Additional work is required to examine SW-GRT methods for binary and time-to-event outcomes and with a small number of groups; methods for estimating ICCs and testing goodness of fit are also needed. Reporting guidelines for SW-GRTs have appeared only recently (67).

Nonparametric methods may be more robust to incorrect specification of the correlation structures and when there are few groups, Ji et al. (77) and Wang & De Gruttola (142) proposed using mixed-effects regression to obtain an estimate for the treatment effect followed by permutation tests to obtain confidence intervals and p-values. Thompson et al. (135) avoided mixed-effects regression altogether: The intervention effect was estimated using group-level data from the intervention and control conditions in vertical slices of time, and time-specific estimates were then combined as an inverse-variance weighted average; the allocation of groups to different intervention times was then randomly permuted to obtain p-values and confidence intervals.

Sample Size Estimation

Numerous resources exist to guide sample size estimation for GRTs, IRGTs, and SW-GRTs. The NIH recently provided an online sample size calculator for GRTs using continuous or binary outcomes, which includes extensive instructional material (https://researchmethodsresources.nih.gov). Kreidler et al. (85) provided a Web-based program for power computation for linear models, including those applicable to GRTs and IRGTs. Hemming et al. (62) introduced an R-Shiny application that supports sample size calculations for GRTs, IRGTs, and SW-GRTs and permits sensitivity analyses under simple and complex correlation structures; further work is required to verify that these methods work well in the case of noncontinuous outcomes and in the case of a small number of groups. Li et al. introduced alternative sample size methods based on GEE for continuous and binary outcomes (91). Moerbeek & Teerenstra’s (103) 2016 text on power analysis of trials with multilevel data addresses sample size issues for GRTs, IRGTs, and SW-GRTs and includes many ICC estimates.

CASE STUDIES

Effect of Greening Vacant Land on Mental Health of Community-Dwelling Adults

This study (130) was a GRT conducted in Philadelphia to evaluate whether interventions to green vacant urban land can improve self-reported mental health. A total of 110 noncontiguous
geographic areas containing 541 vacant lots were randomly allocated to 1 of 3 study conditions: a greening intervention, a trash cleanup intervention, or no intervention. The primary outcome was self-reported mental health, assessed using a validated questionnaire administered to residents selected at random from the area surrounding each lot. Only one participant per household was selected, for a total of 442 community-dwelling adults. Questionnaires were administered in waves during a six-month window before and after the intervention. The authors reported that sample size calculations considered the anticipated ICC and repeated measurements on participants but did not provide details about how this calculation was done. The statistical analyses included 342 residents not lost to follow-up; the authors reported using random-effects regression models that accounted for the nested design. Investigators compared each active intervention to the control. The intervention effects were modeled as the net difference over time between the control and one of the active intervention conditions. Twenty-eight results were reported for a summary measure of mental health and for six subscales. Three significant effects were reported.

This study is an example of a nested cohort design (105), with the same participants measured before and after the intervention. The paper would have been strengthened by reporting the parameter estimates used in the sample size estimation and the values later observed in the data (18), by reporting the model used for the primary analysis, and either by focusing on the primary outcome analysis or by adjusting for multiple comparisons.

The ACTonHEART Study
This study (131) was proposed as a prospective IRGT trial to evaluate the efficacy of a brief group-administered intervention integrating education on heart-healthy behaviors with acceptance and mindfulness skills to modify cardiovascular risk factors and psychological well-being among cardiac patients. Approximately 168 patients were to be recruited and randomized individually to usual care or intervention. Block randomization was to be used to balance gender and baseline cardiovascular risk. The primary outcome variables were proposed as mean changes for low-density lipoprotein (LDL) cholesterol, resting systolic blood pressure, body mass index, and psychological well-being, assessed at baseline, 6 weeks, and 6 and 12 months.

This trial was to have two conditions, with nesting of participants within small groups in only one condition. Although patients were randomized individually, the intervention participants received the ACTonHEART intervention in 6 90-min group-therapy sessions over 6 weeks. The control group received usual care. The authors reported that the power calculation considered the anticipated ICC. Intention-to-treat analyses were proposed to examine outcome measures, and the ICC was to be considered in the multilevel analysis. The authors also reported that two-tailed tests would be used with an alpha of 0.01 (0.05/4) to account for the four primary outcomes.

This study is an example of an IRGT with nesting of participants in small groups in one condition. The paper would have been strengthened by reporting details adequate to replicate the sample size calculation, including the number of therapy groups, the value of the ICC used, and the model proposed for the primary analysis.

Effects of a Free School Breakfast Program on Children’s Attendance, Academic Achievement, and Short-Term Hunger
This study (102) was a 1-year SW-GRT in 14 New Zealand primary schools in low-socioeconomic resource areas. The study’s purpose was to evaluate the effect of introducing a free, daily school breakfast program on the primary outcome of optimal school attendance, defined as attendance of 95% or higher. Eligible schools were randomly assigned to 1 of 4 sequences (3–4 schools per
sequence). Schools randomly allocated to the first sequence implemented the intervention condition immediately upon study entry; thereafter, additional schools crossed over to the intervention condition each term, according to their random assignment. By the final term, all schools were in the intervention condition.

The study was designed to have 85% power to detect a 10% absolute change in the proportion of students with optimal attendance. Sample size methods considered the SW-GRT design and accounted for an anticipated ICC within schools of 0.05. The analysis was performed according to the intention-to-treat principle. The analysis used generalized linear mixed models for categorical outcomes and general linear mixed models for continuous outcomes. The analysis accounted for time as well as for the ICC within schools. Repeated measurements on the same child within the same school over time were also considered in the multilevel analysis. The primary outcome was not statistically significant, nor were all but one secondary outcome.

This study is an example of a closed cohort SW-GRT: All participants were identified at the beginning of the trial and participated until the end of the study, with repeated measurements on the same participants (19). The paper would have been strengthened by reporting full details for the sample size calculation to allow replication and by addressing the issue of decay of the overtime correlation, though that was not standard practice when this study was designed.

**SUMMARY**

GRTs randomize groups to study conditions and measure outcomes in participants from those groups to evaluate the effect of an intervention. The GRT is the best comparative design available if the intervention cannot be delivered separately to individuals, when there is a substantial risk of contamination, or when there are other strong reasons for preferring group randomization. The presence of positive ICC violates the independence of errors assumption that underlies the analytic methods for RCTs requiring different methods to protect the type 1 error rate (16, 29, 33, 55, 105).

IRGTs randomize individuals to study conditions; however, the intervention is delivered in a group setting or by shared interventionists, inducing correlation similar to that in a GRT. The IRGT is the best comparative design available if randomization of individuals is possible but the method of intervention delivery is expected to create positive ICC. Special analytic methods are required to account for the correlation to protect the type 1 error rate (5, 12, 78, 116, 124, 141).

SW-GRTs randomize the order in which groups cross over from the control condition to the intervention condition and provide both control and intervention observations for each group. The SW-GRT is helpful if a GRT is not possible, though the GRT is preferred because it faces fewer threats to internal validity. Because groups are crossed with conditions, the impact of the positive ICC expected within each group is diminished; thus SW-GRTs often provide more power or require fewer groups than a GRT, particularly if the ICC is large.

All three designs have a more complex correlation structure than does the traditional RCT and require sample size estimation and data analytic methods that reflect this structure. They are analyzed most often by mixed-effects regression models, but they can also be analyzed using GEE, two-stage methods, and randomization-based methods. Model-based inference depends on the correct specification of the correlation structure, whereas GEE-based inference depends on the correct specification of the data structure (24). Degrees of freedom are based on the number of groups rather than the number of participants for both GRTs and SW-GRTs; for IRGTs, degrees of freedom are based on the number of groups for any conditions that use intervention delivery methods that induce within-group correlation and on the number of participants for any conditions that do not.
The RCR (74) and citation counts were used to identify 50 influential methods reports for these designs, and we have highlighted the contributions of those reports. We offer three questions to guide the choice from among these three designs and the traditional RCT. Finally, we have summarized current analytic strategies for the three designs and presented case studies as examples.

DISCLOSURE STATEMENT
The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS
We gratefully acknowledge the work of Norma Minkoff and Luis Ganoza Caballero, who performed the bibliographic searches for this article.

LITERATURE CITED
1. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. 2004. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. J. Clin. Epidemiol. 57:785–94
2. Agbla SC, DiazOrdaz K. 2018. Reporting non-adherence in cluster randomised trials: a systematic review. Clin. Trials 15:294–304
3. Andridge RR, Shoben AB, Muller KE, Murray DM. 2014. Analytic methods for individually randomized group treatment trials and group-randomized trials when subjects belong to multiple groups. Stat. Med. 33:2178–90
4. Baio G, Copas A, Ambler G, Hargreaves J, Beard E, Omar RZ. 2015. Sample size calculation for a stepped wedge trial. Trials 16:354
5. Baldwin SA, Bauer DJ, Stice E, Rohde P. 2011. Evaluating models for partially clustered designs. Psychol. Methods 16:149–65
6. Baldwin SA, Murray DM, Shadish WR. 2005. Empirically supported treatments or type I errors? Problems with the analysis of data from group-administered treatments. J. Consult. Clin. Psychol. 73:924–35
7. Barker D, McElduff P, D’Este C, Campbell MJ. 2016. Stepped wedge cluster randomised trials: a review of the statistical methodology used and available. BMC Med. Res. Methodol. 16:69
8. Bauer DJ, Gottfredson NC, Dean D, Zucker RA. 2013. Analyzing repeated measures data on individuals nested within groups: accounting for dynamic group effects. Psychol. Methods 18:1–14
9. Bauer DJ, Sterba SK, Hallfors DD. 2008. Evaluating group-based interventions when control participants are ungrouped. Multivariate Behav. Res. 43:210–36
10. Beard E, Lewis JJ, Copas A, Davey G, Osrin D, et al. 2015. Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. Trials 16:353
11. Bland JM, Kerry SM. 1997. Statistics notes. Trials randomised in clusters. BMJ 315:600
12. Bouten I, Moher D, Altman DG, Schulz KE, Ravaud P. 2008. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. Ann. Intern. Med. 148:295–309
13. Bowater RJ, Abdalmalik SM, Lilford RJ. 2009. The methodological quality of cluster randomised controlled trials for managing tropical parasitic disease: a review of trials published from 1998 to 2007. Trans. R. Soc. Trop. Med. Hyg. 103:429–36
14. Brown CA, Lilford RJ. 2006. The stepped wedge trial design: a systematic review. BMC Med. Res. Methodol. 6:54
15. Bryk AS, Raudenbush SW. 1992. Hierarchical Linear Models: Applications and Data Analysis Methods. Newbury Park, CA: Sage
16. Campbell MJ, Walters SJ. 2014. How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research. Chichester, UK: Wiley
17. Campbell MK, Elbourne DR, Altman DG. 2004. CONSORT statement: extension to cluster randomised trials. BMJ 328:702–8
18. Campbell MK, Piaggio G, Elbourne DR, Altman DG. 2012. Consort 2010 statement: extension to cluster randomised trials. BMJ 345:e5661
19. Copas AJ, Lewis JJ, Thompson JA, Davey C, Baio G, Hargreaves JR. 2015. Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. Trials 16:352
20. Cornfield J. 1978. Randomization by group: a formal analysis. J. Epidemiol. 108:100–2
21. Crites-Christoph P, Mintz J. 1991. Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. J. Consult. Clin. Psychol. 59:20–26
22. Davey C, Hargreaves J, Thompson JA, Copas AJ, Beard E, et al. 2015. Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. Trials 16:358
23. de Hoop E, Woertman W, Teerenstra S. 2013. The stepped wedge cluster randomized trial always requires fewer clusters but not always fewer measurements, that is, participants than a parallel cluster randomized trial in a cross-sectional design. In reply. J. Clin. Epidemiol. 66:1428
24. Desai M, Bryson SW, Robinson T. 2013. On the use of robust estimators for standard errors in the presence of clustering when clustering membership is misspecified. Contemp. Clin. Trials 34:248–56
25. Diaz-Ordaz K, Froud R, Sheehan B, Eldridge S. 2013. A systematic review of cluster randomised trials in residential facilities for older people suggests how to improve quality. BMC Med. Res. Methodol. 13:127
26. Diaz-Ordaz K, Kenward MG, Cohen A, Coleman CL, Eldridge S. 2014. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. Clin. Trials 11:590–600
27. Donner A, Birkett N, Buck C. 1981. Randomization by cluster: sample size requirements and analysis. Am. J. Epidemiol. 114:906–14
28. Donner A, Brown KS, Brasher P. 1990. A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979–1989. Int. J. Epidemiol. 19:795–800
29. Donner A, Klar N. 2000. Design and Analysis of Cluster Randomization Trials in Health Research. London: Arnold
30. Donner A, Klar N. 2002. Issues in the meta-analysis of cluster randomized trials, Stat. Med. 21:2971–80
31. Eccles M, Grimshaw J, Campbell M, Ramsay C. 2003. Research designs for studies evaluating the effectiveness of change and improvement strategies. Qual. Saf. Health Care 12:47–52
32. Eldridge S, Ashby D, Bennett C, Wakelin M, Feder G. 2008. Internal and external validity of cluster randomised trials: systematic review of recent trials. BMJ 336:876–80
33. Eldridge S, Kerry S. 2012. A Practical Guide to Cluster Randomized Trials in Health Research. London: Arnold
34. Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC. 2004. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. Clin. Trials 1:80–90
35. Eldridge SM, Ukoumunne OC, Carlin JB. 2009. The intra-cluster correlation coefficient in cluster randomized trials: a review of definitions. Int. Stat. Rev. 77:378–94
36. Emsley R, Dunn G, White IR. 2010. Mediation and moderation of treatment effects in randomised controlled trials of complex interventions. Stat. Methods Med. Res. 19:237–70
37. Fay M, Graubard P. 2001. Small-sample adjustments for Wald-type tests using sandwich estimators. Biometrics 57:1198–206
38. Feldman HA. 1988. Families of lines: random effects in linear regression analysis. J. Appl. Physiol. 64:1721–32
39. Feng Z, McLerran D, Grizzle J. 1996. A comparison of statistical methods for clustered data analysis with Gaussian error. Stat. Med. 15:1793–806
40. Fiero MH, Huang S, Oren E, Bell ML. 2016. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. Trials 17:72
41. Fok CC, Henry D, Allen J. 2015. Research designs for intervention research with small samples II: stepped wedge and interrupted time-series designs. *Prev. Sci.* 16:967–77
42. Ford WP, Westgate PM. 2017. Improved standard error estimator for maintaining the validity of inference in cluster randomized trials with a small number of clusters. *Biom. J.* 59:478–95
43. Froud R, Eldridge S, Diaz Ordaz K, Marinho VC, Donner A. 2012. Quality of cluster randomized controlled trials in oral health: a systematic review of reports published between 2005 and 2009. *Community Dent. Oral Epidemiol.* 40(Suppl. 1):3–14
44. Gail MH, Mark SD, Carroll RJ, Green SB, Pee D. 1996. On design considerations and randomization-based inference for community intervention trials. *Stat. Med.* 15:1069–92
45. Girling AJ, Hemming K. 2016. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Stat. Med.* 35:2149–66
46. Goldstein H. 1987. *Multilevel Models in Educational and Social Research.* New York: Oxford Univ. Press
47. Goldstein H. 1995. *Multilevel Statistical Models.* London: Edward Arnold
48. Goldstein H. 2003. *Multilevel Statistical Models.* London: Edward Arnold. 3rd ed.
49. Goldstein H. 2011. *Multilevel Statistical Modeling.* Chichester, UK: Wiley. 4th ed.
50. Grant A, Treweek S, Dreischulte T, Foy R, Guthrie B. 2013. Process evaluations for cluster-randomised trials of complex interventions: a proposed framework for design and reporting. *Trials* 14:15
51. Grantham KL, Kasza J, Heritier S, Hemming K, Forbes AB. 2019. Accounting for a decaying correlation structure in cluster randomized trials with continuous recruitment. *Stat. Med.* 38:1918–34
52. Grayling MJ, Wason JMS, Mander AP. 2017. Stepped wedge cluster randomized controlled trial designs: a review of reporting quality and design features. *Trials* 18:33
53. Gulliford MC, Ukoumunne OC, Chinn S. 1999. Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: data from the Health Survey for England 1994. *Am. J. Epidemiol.* 149:876–83
54. Hayes RJ, Bennett S. 1999. Simple sample size calculation for cluster-randomized trials. *Int. J. Epidemiol.* 28:319–26
55. Hayes RJ, Moulton LH. 2009. *Cluster Randomised Trials.* Boca Raton, FL: CRC Press
56. Hayes RJ, Moulton LH. 2017. *Cluster Randomised Trials.* Boca Raton, FL: CRC Press. 2nd ed.
57. Hedeker D, Gibbons RD. 1994. A random-effects ordinal regression model for multilevel analysis. *Biometrics* 50:933–44
58. Hedeker D, Gibbons RD. 1996. MIXOR: a computer program for mixed-effects ordinal regression analysis. *Comput. Methods Programs Biomed.* 49:157–76
59. Hedges LV, Hedberg EC. 2007. Intraclass correlation values for planning group-randomized trials in education. *Educ. Eval. Policy Anal.* 29:60–87
60. Hemming K, Girling A. 2013. The efficiency of stepped wedge versus cluster randomized trials: stepped wedge studies do not always require a smaller sample size. *J. Clin. Epidemiol.* 66:1427–28
61. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. 2015. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ* 350:h391
62. Hemming K, Kasza J, Hooper R, Forbes A, Taljaard M. 2020. A tutorial on sample size calculation for cluster randomised multiple-period parallel, cross-over and stepped-wedge trials using the Shiny CRT Calculator. *Int. J. Epidemiol.* In press
63. Hemming K, Lilford R, Girling AJ. 2015. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. *Stat. Med.* 34:181–96
64. Hemming K, Taljaard M. 2016. Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. *J. Clin. Epidemiol.* 69:137–46
65. Hemming K, Taljaard M, Forbes A. 2017. Analysis of cluster randomised stepped wedge trials with repeated cross-sectional samples. *Trials* 18:101
66. Hemming K, Taljaard M, Forbes A. 2018. Modeling clustering and treatment effect heterogeneity in parallel and stepped-wedge cluster randomized trials. *Stat. Med.* 37:883–98
67. Hemming K, Taljaard M, McKenzie JE, Hooper R, Copas A, et al. 2018. Reporting of stepped wedge cluster randomised trials: extension of the CONSORT 2010 statement with explanation and elaboration. *BMJ* 363:k1614
68. Heo M, Litwin AH, Blackstock O, Kim N, Arnsten JH. 2017. Sample size determinations for group-based randomized clinical trials with different levels of data hierarchy between experimental and control arms. *Stat. Methods Med. Res.* 26:399–413

69. Heo M, Nair SR, Wylie-Rosett J, Faith MS, Pietrobelli A, et al. 2018. Trial characteristics andappropriateness of statistical methods applied for design and analysis of randomized school-based studies addressing weight-related issues: a literature review. *J. Obes.* 2018:8767315

70. Hooper R, Forbes A, Hemming K, Takeda A, Beresford L. 2018. Analysis of cluster randomised trials with an assessment of outcome at baseline. *BMJ* 360:k1121

71. Huang S, Fiero MH, Bell ML. 2016. Generalized estimating equations in cluster randomized trials with a small number of clusters: review of practice and simulation study. *Clin. Trials* 13:445–49

72. Hughes JP, Granston TS, Heagerty PJ. 2015. Current issues in the design and analysis of stepped wedge trials. *Contemp. Clin. Trials* 45:55–60

73. Hussey MA, Hughes JP. 2007. Design and analysis of stepped wedge cluster randomized trials. *Contemp. Clin. Trials* 28:182–91

74. Hutchins BI, Yuan X, Anderson JM, Santangelo GM. 2016. Relative citation ratio (RCR): a new metric that uses citation rates to measure influence at the article level. *PLOS Biod.* 14:e1002541

75. Ivers NM, Halperin IJ, Barnsley J, Grimshaw JM, Shah BR, et al. 2012. Allocation techniques for balance at baseline in cluster randomized trials: a methodological review. *Trials* 13:120

76. Ivers NM, Taljaard M, Dixon S, Bennett C, McRae A, et al. 2011. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000–8. *BMJ* 343:d5886

77. Ji XY, Fink G, Robyn PJ, Small DS. 2017. Randomization inference for stepped-wedge cluster-randomized trials: an application to community-based health insurance. *Ann. Appl. Stat.* 11:1–20

78. Kahan BC, Morris TP. 2013. Assessing potential sources of clustering in individually randomised trials. *BMC Med. Res. Methodol.* 13:58

79. Kasza J, Forbes AB. 2019. Inference for the treatment effect in multiple-period cluster randomised trials when random effect correlation structure is misspecified. *Stat. Methods Med. Res.* 28:3112–22

80. Kasza J, Hemming K, Hooper R, Matthews J, Forbes AB, et al. 2019. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. *Stat. Methods Med. Res.* 28:703–16

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

82. Kenward MG, Roger JH. 1997. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 53:983–97

83. Kish L. 1965. *Survey Sampling.* New York: Wiley

84. Klar N, Darlington G. 2004. Methods for modelling change in cluster randomization trials. *Stat. Med.* 23:2341–57

85. Kreidler SM, Muller KE, Grunwald GK, Ringham BM, Coker-Dukowitz ZT, et al. 2013. *GLIMMPSE*: online power computation for linear models with and without a baseline covariate. *J. Stat. Softw.* 54:i10

86. Kristunas C, Morris T, Gray L. 2017. Unequal cluster sizes in stepped-wedge cluster randomised trials: a systematic review. *BMJ Open* 7:e017151

87. Krull JL, MacKinnon DP. 1999. Multilevel mediation modeling in group-based intervention studies. *Eval. Rev.* 23:418–44

88. Krull JL, MacKinnon DP. 2001. Multilevel modeling of individual and group level mediated effects. *Multivariate Behav. Res.* 36:249–77

89. Lee KJ, Thompson SG. 2005. Clustering by health professional in individually randomised trials. *BMJ* 330:142–44

90. Lee KJ, Thompson SG. 2005. The use of random effects models to allow for clustering in individually randomised trials. *Clin. Trials* 2:163–73

91. Li F, Turner EL, Preisser JS. 2018. Sample size determination for GEE analyses of stepped wedge cluster randomized trials. *Biometrics* 74:1450–58
92. Li P, Redden DT. 2015. Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. *BMC Med. Res. Methodol.* 15:38

93. Li P, Redden DT. 2015. Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Stat. Med.* 34:281–96

94. Lindquist EF. 1953. *Design and Analysis of Experiments in Psychology and Education.* Boston: Houghton Mifflin

95. Lohr S, Schochet PZ, Sanders E. 2014. Partially nested randomized controlled trials in education research: a guide to design and analysis. NCER 2014-2000, Natl. Cent. Educ. Res., Inst. Educ. Sci., US Dep. Educ., Washington, DC. [https://ies.ed.gov/ncer/pubs/20142000/pdf/20142000.pdf](https://ies.ed.gov/ncer/pubs/20142000/pdf/20142000.pdf)

96. Luo W, Cappaert KJ, Ning L. 2015. Modelling partially cross-classified multilevel data. *Br. J. Math. Stat. Psychol.* 68:342–62

97. Mancl LA, DeRouen TA. 2001. A covariance estimator for GEE with improved small-sample properties. *Biometrics* 57:126–34

98. Martin J, Taljaard M, Girling A, Hemming K. 2016. Systematic review finds major deficiencies in sample size methodology and reporting for stepped-wedge cluster randomised trials. *BMJ Open* 6:e010166

99. Martindale C. 1978. *The therapist-as-fixed-effect fallacy in psychotherapy research.* *J. Consult. Clin. Psychol.* 46:1526–30

100. Mdege ND, Brabyn S, Hewitt C, Richardson R, Torgerson DJ. 2014. The 2 × 2 cluster randomized controlled factorial trial design is mainly used for efficiency and to explore intervention interactions: a systematic review. *J. Clin. Epidemiol.* 67:1083–92

101. Mdege ND, Man MS, Taylor Nee Brown CA, Torgerson DJ. 2011. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *J. Clin. Epidemiol.* 64:936–48

102. Mhurchu CN, Gorton D, Turley M, Jiang Y, Michie J, et al. 2013. Effects of a free school breakfast programme on children’s attendance, academic achievement and short-term hunger: results from a stepped-wedge, cluster randomised controlled trial. *J. Epidemiol. Community Health* 67:257–64

103. Moerbeek M, Teerenstra S. 2016. *Power Analysis of Trials with Multilevel Data.* Boca Raton, FL: CRC

104. Morel JG, Bokossa MC, Neerchal NK. 2003. Small sample correction for the variance of GEE estimators. *Biom. J.* 45:395–409

105. Murray DM. 1998. *Design and Analysis of Group-Randomized Trials.* New York: Oxford Univ. Press

106. Murray DM, Hannan PJ. 1990. Planning for the appropriate analysis in school-based drug-use prevention studies. *J. Consult. Clin. Psychol.* 58:458–68

107. Murray DM, Hannan PJ, Baker WL. 1996. A Monte Carlo study of alternative responses to intraclass correlation in community trials. Is it ever possible to avoid Cornfield’s penalties? *Evol. Rev.* 20:313–37

108. Murray DM, Hannan PJ, Pals SP, McGowen RG, Baker WL, Bliststein JL. 2006. A comparison of permutation and mixed-model regression methods for the analysis of simulated data in the context of a group-randomized trial. *Stat. Med.* 25:375–88

109. Murray DM, Hannan PJ, Wolfinger RD, Baker WL, Dwyer JH. 1998. Analysis of data from group-randomized trials with repeat observations on the same groups. *Stat. Med.* 17:1581–600

110. Murray DM, Pals SL, Bliststein JL, Alfano CM, Lehman J. 2008. Design and analysis of group-randomized trials in cancer: a review of current practices. *J. Natl. Cancer Inst.* 100:483–91

111. Murray DM, Pals SL, George SM, Kuzmichev A, Lai GY, et al. 2018. Design and analysis of group-randomized trials in cancer: a review of current practices. *Prev. Med.* 111:241–47

112. Murray DM, Pennell M, Rhoda D, Hade EM, Paskett ED. 2010. Designing studies that would address the multilayered nature of health care. *J. Natl. Cancer Inst. Monogr.* 2010:90–96

113. Murray DM, Varnell SP, Bliststein JL. 2004. Design and analysis of group-randomized trials: a review of recent methodological developments. *Am. J. Public Health* 94:423–32

114. Nickless A, Voysey M, Geddes J, Yu LM, Fanshawe TR. 2018. Mixed effects approach to the analysis of the stepped wedge cluster randomised trial—investigating the confounding effect of time through simulation. *PLOS ONE* 13:e0208876
115. Nye B, Konstantopoulos S, Hedges L. 2004. How large are teacher effects? Educ. Eval. Policy Anal. 26:237–57
116. Pals SL, Murray DM, Alfano CM, Shadish WR, Hannan PJ, Baker WL. 2008. Individually randomized group treatment trials: a critical appraisal of frequently used design and analytic approaches. Am. J. Public Health 98:1418–24
117. Pals SL, Wiegand RE, Murray DM. 2011. Ignoring the group in group-level HIV/AIDS intervention trials: a review of reported design and analytic methods. AIDS 25:989–96
118. Puffer S, Torgerson D, Watson J. 2003. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. BMJ 327:785–89
119. Raab GM, Butcher I. 2005. Randomization inference for balanced cluster-randomized trials. Clin. Trials 2:130–40
120. Rao JN, Scott AJ. 1992. A simple method for the analysis of clustered binary data. Biometrics 48:577–85
121. Raudenbush SW. 1997. Statistical analysis and optimal design for cluster randomized trials. Psychol. Methods 2:173–85
122. Raudenbush SW, Bryk AS. 2002. Hierarchical Linear Models: Applications and Data Analysis Methods. Thousand Oaks, CA: Sage
123. Richardson M, Garner P, Donegan S. 2016. Cluster randomised trials in Cochrane Reviews: evaluation of methodological and reporting practice. PLOS ONE 11:e0151818
124. Roberts C, Roberts SA. 2005. Design and analysis of clinical trials with clustering effects due to treatment. Clin. Trials 2:152–62
125. Roberts C, Walwyn R. 2013. Design and analysis of non-pharmacological treatment trials with multiple therapists per patient. Stat. Med. 32:81–98
126. Rutterford C, Taljaard M, Dixon S, Copas A, Eldridge S. 2015. Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review. J. Clin. Epidemiol. 68:716–23
127. Satterthwaite FE. 1946. An approximate distribution of estimates of variance components. Biometrics 2:110–14
128. Scott JM, deCamp A, Juraska M, Fay MP, Gilbert PB. 2017. Finite-sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials. Stat. Methods Med. Res. 26:583–97
129. Simpson JM, Klar N, Donner A. 1995. Accounting for cluster randomization—a review of primary prevention trials, 1990 through 1993. Am. J. Public Health 85:1378–83
130. South EC, Hohl BC, Kondo MC, MacDonald JM, Branas CC. 2018. Effect of greening vacant land on mental health of community-dwelling adults: a cluster randomized trial. JAMA Netw. Open. 1:e180298
131. Spatola CA, Manzoni GM, Castelnuovo G, Malfatto G, Facchini M, et al. 2014. The ACTonHEART study: rationale and design of a randomized controlled clinical trial comparing a brief intervention based on acceptance and commitment therapy to usual secondary prevention care of coronary heart disease. Health Qual. Life Outcomes 12:22
132. Sterba SK. 2017. Partially nested designs in psychotherapy trials: a review of modeling developments. Psychother. Res. 27:425–36
133. Taljaard M, Hemming K, Shah L, Giraudureau B, Grimshaw JM, Weijer C. 2017. Inadequacy of ethical conduct and reporting of stepped wedge cluster randomized trials: results from a systematic review. Clin. Trials 14:333–41
134. Taljaard M, Weijer C, Grimshaw JM, Eccles MP, Ott. Ethics Cluster Randomised Trials Consens. Group. 2013. The Ottawa Statement on the ethical design and conduct of cluster randomised trials: precis for researchers and research ethics committees. BMJ 346:f2838
135. Thompson JA, Davey C, Fielding K, Hargreaves JR, Hayes RJ. 2018. Robust analysis of stepped wedge trials using cluster-level summaries within periods. Stat. Med. 37:2487–500
136. Torgerson DJ. 2001. Contamination in trials: Is cluster randomisation the answer? BMJ 322:355–57
137. Turner EL, Li F, Gallis JA, Prague M, Murray DM. 2017. Review of recent methodological developments in group-randomized trials: part 1-design. Am. J. Public Health 107:907–15
138. Turner EL, Prague M, Gallis JA, Li F, Murray DM. 2017. Review of recent methodological developments in group-randomized trials: part 2-analysis. *Am. J. Public Health* 107:1078–86
139. Ukoumunne OC, Forbes AB, Carlin JB, Gulliford MC. 2008. Comparison of the risk difference, risk ratio and odds ratio scales for quantifying the unadjusted intervention effect in cluster randomized trials. *Stat. Med.* 27:5143–55
140. Varnell SP, Murray DM, Janega JB, Blitstein JL. 2004. Design and analysis of group-randomized trials: a review of recent practices. *Am. J. Public Health* 94:393–99
141. Walwyn R, Roberts C. 2010. Therapist variation within randomised trials of psychotherapy: implications for precision, internal and external validity. *Stat. Methods Med. Res.* 19:291–315
142. Wang R, De Gruttola V. 2017. The use of permutation tests for the analysis of parallel and stepped-wedge cluster-randomized trials. *Stat. Med.* 36:2831–43
143. Whiting-O’Keefe QE, Henke C, Simborg DW. 1984. Choosing the correct unit of analysis in medical care experiments. *Med. Care* 22:1101–14
144. Woertman W, de Hoop E, Moerbeek M, Zuiderma SU, Gerritsen DL, Teerenstra S. 2013. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *J. Clin. Epidemiol.* 66:752–58
145. Wright N, Ivers N, Eldridge S, Taljaard M, Brenner S. 2015. A review of the use of covariates in cluster randomized trials uncovers marked discrepancies between guidance and practice. *J. Clin. Epidemiol.* 68:603–9
146. Wu S, Crespi CM, Wong WK. 2012. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contemp. Clin. Trials* 33:869–80
147. Zhang K, Traskin M, Small DS. 2012. A powerful and robust test statistic for randomization inference in group-randomized trials with matched pairs of groups. *Biometrics* 68:75–84
148. Zou GY, Donner A. 2013. Extension of the modified Poisson regression model to prospective studies with correlated binary data. *Stat. Methods Med. Res.* 22:661–70