UC Irvine
UC Irvine Previously Published Works

Title

Accounting for Context in Randomized Trials after Assignment

Permalink

Journal

ISSN

Authors

Brown, C Hendricks
Hedeker, Donald
Gibbons, Robert D
et al.

Publication Date

DOI

Copyright Information

Peer reviewed

Accounting for Context in Randomized Trials after Assignment

C. Hendricks Brown[1] · Donald Hedeker[2] · Robert D. Gibbons[2] · Naihua Duan[3] · Daniel Almirall[4] · Carlos Gallo[1] ·
Inger Burnett-Zeigler[1] · Guillermo Prado[5] · Sean D. Young[6] · Alberto Valido[7] · Peter A. Wyman[8]

Abstract
Many preventive trials randomize individuals to intervention condition which is then delivered in a group setting. Other trials randomize higher levels, say organizations, and then use learning collaboratives comprised of multiple organizations to support improved implementation or sustainment. Other trials randomize or expand existing social networks and use key opinion leaders to deliver interventions through these networks. We use the term contextually driven to refer generally to such trials (traditionally referred to as clustering, where groups are formed either pre-randomization or post-randomization — i.e., a cluster-randomized trial), as these groupings or networks provide fixed or time-varying contexts that matter both theoretically and practically in the delivery of interventions. While such contextually driven trials can provide efficient and effective ways to deliver and evaluate prevention programs, they all require analytical procedures that take appropriate account of non-independence, something not always appreciated. Published analyses of many prevention trials have failed to take this into account. We discuss different types of contextually driven designs and then show that even small amounts of non-independence can inflate actual Type I error rates. This inflation leads to rejecting the null hypotheses too often, and erroneously leading us to conclude that there are significant differences between interventions when they do not exist. We describe a procedure to account for non-independence in the important case of a two-arm trial that randomizes units of individuals or organizations in both arms and then provides the active treatment in one arm through groups formed after assignment. We provide sample code in multiple programming languages to guide the analyst, distinguish diverse contextually driven designs, and summarize implications for multiple audiences.

Keywords Individually randomized group treated (IRGT) trials · Partially nested designs · Contextually driven designs · Mixed effects modeling · Generalized estimating equations · Spillover trials · Multiplicative implementation strategies · Learning collaboratives · Clustering · Cluster-randomized trials

✉ C. Hendricks Brown
hendricks.brown@northwestern.edu

1 Department of Psychiatry and Behavioral Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

2 Center for Health Statistics, The University of Chicago, Chicago, IL, USA

3 Department of Psychiatry, Columbia University Medical Center, New York, NY, USA

4 Institute for Social Research and Department of Statistics, University of Michigan, Ann Arbor, MI, USA

5 Graduate School, University of Miami, Miami, FL, USA

6 Department of Emergency Medicine, School of Medicine, Department of Informatics, Bren School of Information and Computer Sciences, University of California, Irvine, CA, USA

7 School of Education, University of North Carolina at Chapel Hill, Chapel Hill, Orange, NC, USA

8 Department of Psychiatry, University of Rochester School of Medicine, Rochester, NY, USA

Introduction

There are strong theoretical as well as practical reasons for delivering many preventive interventions in group or network settings. In every stage of life, each natural social field (e.g., family of origin at birth, elementary school classroom) provides powerful influences in shaping and reinforcing behavior (Kellam & Brown, 1986; Szapocznik & Coatsworth, 1999), developing skills, providing prosocial or limiting antisocial opportunities (Botvin, 2000; Hawkins et al., 2005, 2015), and affecting later life course outcomes (Kellam et al., 1994). Preventive interventions often integrate their core elements within the contexts of family, school, faith centers, work, or community settings (Kellam et al., 1999), as well as through newer social media and networks (Hunter et al., 2019; Young et al., 2013), and have long-term impact on a wide range of health outcomes (Kellam et al., 2008; Sandler et al., 2015). From a public health standpoint, it is more efficient to deliver universal and selective interventions in groups compared to individuals, provided there is (1) similar or improved efficacy — since delivery costs per person are lower — and (2) no group iatrogenic effects such as those in group foster care (Chamberlain & Reid, 1998), juvenile delinquency programs (Petrosino et al., 2013), or settings where maladaptive behaviors can be learned from the group (Dishion et al., 1999).

An under-recognized concern in analyzing prevention trials where an intervention is delivered in a group, network, or other contextual setting is the failure to account for non-independence. For example, when subjects are already in groups and groups are randomized to different treatments (e.g., classrooms within a school are randomized to treatments exposing all its students), or when individuals are randomized first to intervention conditions and an intervention is delivered in a group context, then analyzing the data as if they were independent inflates the Type I error rate. Also, when clustering into groups is ignored, the erroneous tests will reject the null hypothesis too often because the critical rejection value is smaller than it should be when independence is erroneously assumed. Furthermore, sample size calculation for a trial that ignores this non-independence will lead to an underpowered design.

We can see how clustering has a unique effect on intervention inferences from a simple example. Consider first comparing the proportion of a population in two different states who have a diagnosis, in each state, we randomly select 10,000 subjects. In this large, epidemiologic study the difference in the standard estimated proportions ($D = \hat{p}_1 - \hat{p}_2$) its ordinary standard error ($se = \sqrt{((\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2))/10,000}$) and z-test D/se are completely appropriate for comparing these rates. Now, consider a revised study where a statewide intervention is randomly assigned to one of these states, the other state serves as a control, and the same sampling is used.

The standard error can no longer provide a valid measure of variation in statewide intervention effects, as despite the large sample sizes, this is essentially a N of 2 design where the difference in states' observed rates might be due to existing differences between the two states and not caused by the intervention. We would need more than one state receiving this intervention to incorporate between-state variation, despite the exceptionally large sample sizes. With multiple states randomly assigned to the same intervention condition, we can use multilevel modeling to estimate between-state variation and make valid inferences about the intervention (Gibbons et al., 1988). But if we ignore states and conduct a single level analysis, we will inflate the Type I error, sometimes enormously. For example, in a trial with 10 states randomized to each arm and 10,000 per group, a nominal 0.05 level test that ignores state variation when in fact it accounts for only 0.01 of the total variation — a very small intraclass correlation, ICC — would reject the null fully 75% of the time when the null is true. The reason for this enormous inflation is that ignoring clustering within states results in a standard error that is one-third as large as it should be. The inferential concern is identical with other designs in which higher levels, such as clinics or municipal governments are randomized to conditions, and the outcome is based on data from a lower level (e.g., patients).

This last example involves grouping that *exists* at the time of randomization and is called a group randomized trial (Murray, 1998). But a similar situation occurs when groups are formed and used to deliver an intervention *after* randomization. One can see the inferential challenge most clearly in the case where all those assigned to one intervention are brought together in a single group to receive the intervention. In the other arm, each subject receives the intervention individually. Having a single group in the first arm provides no ability to measure variation in how the intervention impacts the group. For example, we would not know whether factors such as variation in attendance in the groups or variation in fidelity of intervention delivery would affect outcomes. Without knowing how outcome varies by group, we cannot form a statistical test of the intervention effect based on the difference in outcomes for the two arms relative to its standard error. On the other hand, if individuals in the first arm are distributed into multiple groups, then variation in impact across the groups can be estimated, using the variation in group means that are independent of one another. In general, an intervention that is delivered within a group *after* randomization also requires multiple groups and multilevel modeling to evaluate its impact (see further technical details in online supplement Appendix 1a).

In many prevention trials, the contexts differ importantly between the arms of the trial. We thus introduce a common prevention design where only one intervention condition involves delivery after randomization in groups, say for a

behavioral intervention, so non-independence occurs only in one arm of the trial. Those trials using individual level randomization with one arm delivered in a group format are known as Individual Randomized Group Treated (IRGT) trials (Pals et al., 2008), also called a partially clustered design (PCD) (Li & Hedeker, 2017).

Frequently, the published analyses of IRGTs ignore grouping or other contexts formed after randomization. We show how serious such errors can be for this common prevention trial design. This paper is not the first to point out the importance of correctly analyzing IRGT prevention and treatment trials (Andridge et al., 2014; Lee & Thompson, 2005; Li & Hedeker, 2017; Moerbeek & Wong, 2008; Murray et al., 2004; Pals et al., 2008, 2011; Roberts & Roberts, 2005; Turner et al., 2017). Nor is it the first to point out that many of the most common interventions that are used in behavioral health are delivered, at least in part, in a group context or have a nesting structure that introduces non-independence, and these contextual effects are often not accounted for in analyses or in the design of the trial (Pals et al., 2011). Our contributions to this literature, besides reinforcing these concerns, are four-fold. First, we demonstrate that naïve but incorrect specification of context in IRGT designs leads to incorrect inferences, and even when group context factors are small the inflation of Type I error can be quite large. Second, we provide explicit modeling instructions for IRGTs including coding of univariate and growth modeling of this design in six common statistical packages. Also, as two procedures have been proposed to approximate the null distributions of test statistics, we provide recommendations regarding their use in practice. Third, to recognize the full spectrum of contextually driven designs in prevention, we provide numerous examples of trials in the literature as well as introduce a new class of trials that seem well suited to strengthen prevention impact. Fourth, for diverse audiences in the prevention field, we show how contextually driven interventions can help inform prevention theory and improve population level impact of our interventions.

Consequences of Mixing up the Specification of Random Effects in a Large IRGT Trial

Central to the analysis of clustered data in general and IRGTs in particular is the idea of a random effect. The random effect describes how a particular cluster or group deviates from the overall sample mean in terms of parameter(s) related to the outcome of interest. From a statistical perspective, the inclusion of random effects in a model is complicated because as the number of clusters or groups increases, so does the parameter space. This is the so-called "nuisance parameter" problem (see Hedeker & Gibbons, 2006). The solution is to estimate the random-effect variance (over all groups) instead

of the group-specific deviations. These group-specific deviations, or random-effects, can then be estimated using Bayes or empirical Bayes methods (see Hedeker & Gibbons, 2006). For IRGTs and cluster randomized studies with a single outcome, we are typically interested in a random-intercept model, which allows each group or cluster to have its own mean value. This leads to the variance decomposition into the components of within group variation σ^2_W and between-group variation σ^2_B, leading to the intra-class correlation (ICC) which describes the proportion of total variance attributable to groups (formally defined in the following example).

The following constructed example demonstrates that ignoring or incorrectly specifying random effects in IRGT designs can readily lead to erroneous conclusions. We generated a large dataset representing an IRGT design. By using a large dataset, we directly demonstrate where ignoring or incorrectly specifying random effects leads to problems in analysis without the need to account for correcting for sample size in statistical tests (e.g., using a critical t-value rather than the traditional critical z-value of 1.96 for a Type I error of 0.05). Specifically, in this experiment shown in Table 1, there were 200 groups formed after randomization in one arm (group treatment, or Tx = 1), each having 40 subjects per group. The other arm consisted of a control condition, represented by Tx = 0, had 8000 subjects, the same number of subjects as in the arm with group-delivered intervention. For controls, the response variable Y was normally distributed with 0 mean and a within group individual-level standard deviation $\sigma_W = 1$. In the group treatment arm, the mean of the normally distributed response variable was 0.5, with an individual level residual standard deviation $\sigma_W = 1$. The groups were generated with a between level random effect standard deviation of $\sigma_B = 0.5$. This is a large intraclass correlation ICC = $\sigma_B^2 / (\sigma_B^2 + \sigma_W^2) = 0.2$, which measures the proportion of variance due to groups.

We conducted six different analyses of these data and checked $\widehat{\beta}$, the treatment effect estimate (true value of 0.5), and its standard error, allowing us to construct tests against the null value of 0 and whether a confidence interval contained the true value. We also report standard deviation estimates of the random effects (true values of $\sigma_W = 1$ and $\sigma_B = 0.5$). All of these six analyses are at face value reasonable if one does not examine the IRGT model carefully; however, only one leads to appropriate statistical inferences, and all the others produce incorrect inferences. Thus, we use this example to identify what potential errors in statistical conclusions may occur in practice by not accounting for the IRGT design appropriately.

In Table 1, we have shown in bold those standard errors, test statistics, and standard deviations of random effects that are incorrect. Note first that all six models have virtually the same point estimate $\widehat{\beta}$ of the intervention effect, and all are close to the true population value of 0.5. In Row

Table 1 Inferences of incorrect and correct analyses of a large individually randomized group treated trial (bold face estimates are erroneous)

	Name and formula $Y = \alpha + \beta Tx + \varepsilon(Tx) + \delta$	R code for fixed and random effects (R Core Team, 2021)	$\hat{\beta}$ (se) t-test for $\beta = 0$	Estimated Standard deviation of random effects or error (δ) (correlation)
1. Incorrect ignoring of grouping effects	Fixed effects model $\varepsilon(Tx) = 0$	lm (y ~ Tx)	0.456 (**.017**) $t = $ **27.4**	$\sigma_\delta = 1.075$
2. Correct specification of IRGT	IRGT model $\varepsilon(Tx = 0) = 0$ $\varepsilon(Tx = 1) \sim N(0, \sigma_1^2)$	lmer (y ~ Tx + (−1 + Tx | group))	0.456 (.035) $t = 12.865$	$\sigma_\delta = 1.002$ $\sigma_1 = 0.448$
3. Incorrect treatment of both arms as including a single grouped random intercept	Random intercept model $\varepsilon(Tx \mid \text{group } i) \sim N(0, \sigma^2)$, $i = 0, 1, \ldots, 200$	lmer (y ~ Tx + (1 | group))	0.456 (**.450**) $t = $ **1.01**	$\sigma_\delta = 1.00$ $\sigma = 0.448$
4. Incorrect treatment of arms as having different variances, but no grouping variance	Fixed effects model with $\varepsilon(Tx = 0) \sim N(0, \sigma_0^2)$ $\varepsilon(Tx = 1) \sim N(0, \sigma_1^2)$	glm (y ~ Tx, weights = varIdent(form = ~ 1|Tx))	0.456 (**0.016**) $t = $ **27.455**	$\sigma_0 = 1.007$ $\sigma_1 = 1.049*$
5. Incorrect common intercept and treatment random effects	Random intercept and treatment model $\text{Var } \varepsilon = \sigma^2_{\text{Intercept}}$ $\text{Var } \varepsilon(Tx = 1) = \sigma^2_1$	lmer (y ~ Tx + (1 + Tx |group))	0.456 (**1.088**) $t = $ **0.419**	**$\sigma_{\text{Intercept}} = 1.09$** **$\sigma_1 = 1.133$** **Corr $= -.92$** **convergence problems**
6. Incorrect inclusion of two independent random effects, one for control and one for Treatment	Distinct and uncorrelated random effects for each treatment condition $\text{Var } \varepsilon(Tx = 0) = \sigma^2_0$ $\text{Var } \varepsilon(Tx = 1) = \sigma^2_1$	lmer (y ~ Tx + (−1 + Tx0 || group) + (−1 + Tx1 | group)	0.456 (**1.008**) $t = $ **0.452**	$\sigma_0 = 1.01$ $\sigma_1 = 0.448$ **convergence problems**

*Correctly estimates Var ($y \mid Tx = 1$) but ignores group level variance

1, when we ignore the treatment arm's group level random effect entirely, we grossly underestimate the standard error and consequently overestimate the t-value for testing the null hypothesis of no difference in means ($t = 27.4$). Indeed, for this model, the standard error is so small that a 95% confidence interval formed for the difference in means, $0.456 +/- 0.017 * 1.96 = (0.423, 0.489)$, does not include the true value of 0.50. Analysis 2 shows a correct analysis; the standard error of $\hat{\beta}$ is twice as large as the one that does not account for the group level random effect (Row 1), and the standard deviation of the group-level random-effect is close to the true value of 0.5. All estimates in this correct model are within their 95% confidence intervals. The analysis in Row 3 incorrectly assumes both arms are subject to a common group-level random effect. The standard error of $\hat{\beta}$ is extremely large, 90 times as large as the appropriate value in Row 2, and even in this large study, the null hypothesis of $\beta = 0$ is not rejected ($t = 1.01$). In Row 4, we conduct a weighted least squares analysis where the variance depends on the treatment but ignores groups entirely. While one might think this analysis would be appropriate, ignoring group heterogeneity produces a much too small standard error for the difference in means. The last two analyses in Rows 5 and 6 both include two random effects that at first glance may seem adequate, but neither are appropriate for an IRGT. Row 5's fit includes a common intercept random effect and a random effect for groups in the active

intervention arm, which are allowed to be correlated. The standard error of $\hat{\beta}$ is again far too large, and consequently, the test of no difference in means is not rejected. Also, there were major convergence problems with this model, not the least of which is the high correlation between the two random effects. Row 6 models distinct standard deviations and independent random effects for each arm of the trial. While the point estimate, standard error, and test against the null are accurate, there were again convergence problems, and both standard deviations were erroneous. Convergence problems occurred in models 5 and 6 because the standard deviation for group in the $Tx = 0$ arm was based on only one homogeneous group; hence, there are 0 degrees of freedom to estimate its variance.

We conclude from this simple simulated example that it is critical to formulate the random effect to match the IRGT model. If the analysis ignores grouping effects entirely (Row 1 and Row 4), the standard error is severely biased downward, leading to too often rejecting the null hypothesis and an inflated power estimate. Alternatively, if one naively assumes a common random effect for the groups and the observations that are ungrouped (Row 3), then the standard error is severely biased upward, leading to non-rejection of the null hypothesis and a loss of power. If the model includes two random effects when only one should be present, these overdetermined models lead to standard errors that are too large and have convergence problems.

Examining Type I Error for an Individually Randomized Group Trial Using Correct and Incorrect Analyses

Above, we examined how misspecification of the random effects led to erroneous inferences under the alternative hypothesis. Here, we examine the behavior of two analyses under the null hypothesis of no treatment difference. While some simulation studies of IRGT models have been conducted (Li & Hedeker, 2017; Moerbeek & Wong, 2008), these new results on the true Type I error show how sensitive these errors are when ignoring the group level random effect (Fig. 1) and how accurate the Satterthwaite method is (Fig. 2). Figure 3 in the online supplement shows that a Kenward Roger approximation using a scaled F-test and fractional degrees of freedom is inaccurate as implemented in R.

Consider a two-arm IRGT where Tx = 0 represents the control condition and Tx = 1 represents the intervention delivered in group format. In the group-delivered arm, each of G groups has the same number of subjects, N_{Group}. There is an equal number of subjects assigned to Tx = 0 as Tx = 1, i.e., the total number of controls is $N_{Control} = G * N_{Group}$. In these simulations, we vary the number of groups, G and the ICC = $\sigma_B^2 / \sigma_B^2 + \sigma_W^2$. Choosing normally distributed outcomes, we generated datasets according to $Y_{ij} = \alpha + \beta\,Tx_i + \varepsilon_i + \delta_{ij}$.

When $i = 0$, the control arm, j ranges from 1 to $N_{Control}$. When i is between 1 and G, the intervention arm, j ranges from 1 to N_{group}. To examine Type I error, we set the coefficient $\beta = 0$ to represent the null model. We specify the within group random effects having distributions (see online supplement Appendix 1b for details),

$\delta_{ij} \sim N(0, \sigma_W^2)$, an independent error for every individual, and

$\varepsilon_i \sim N(0, \sigma_B^2)$ and $Tx_i = 1$ for $i = 1, \ldots, N_{Group}$,
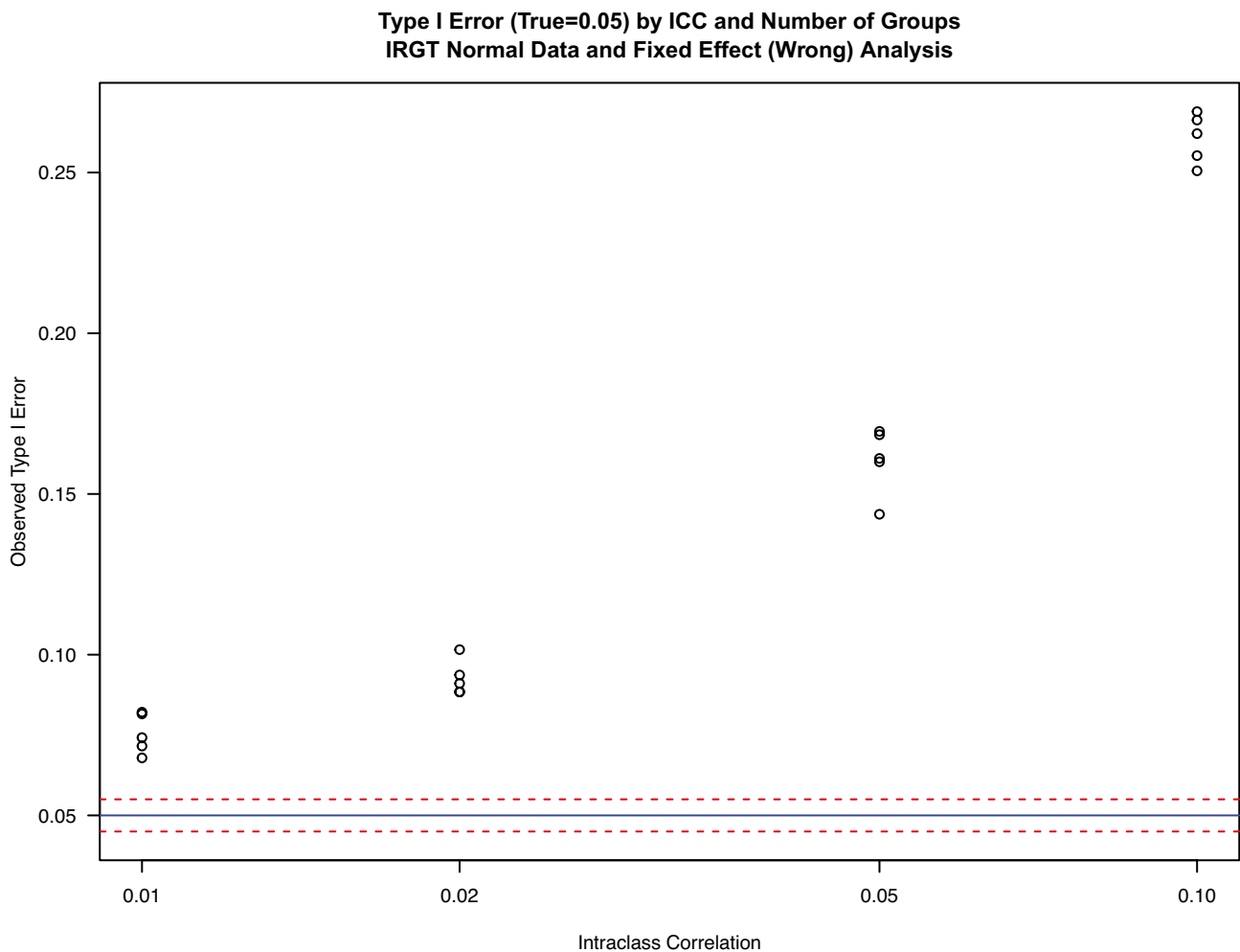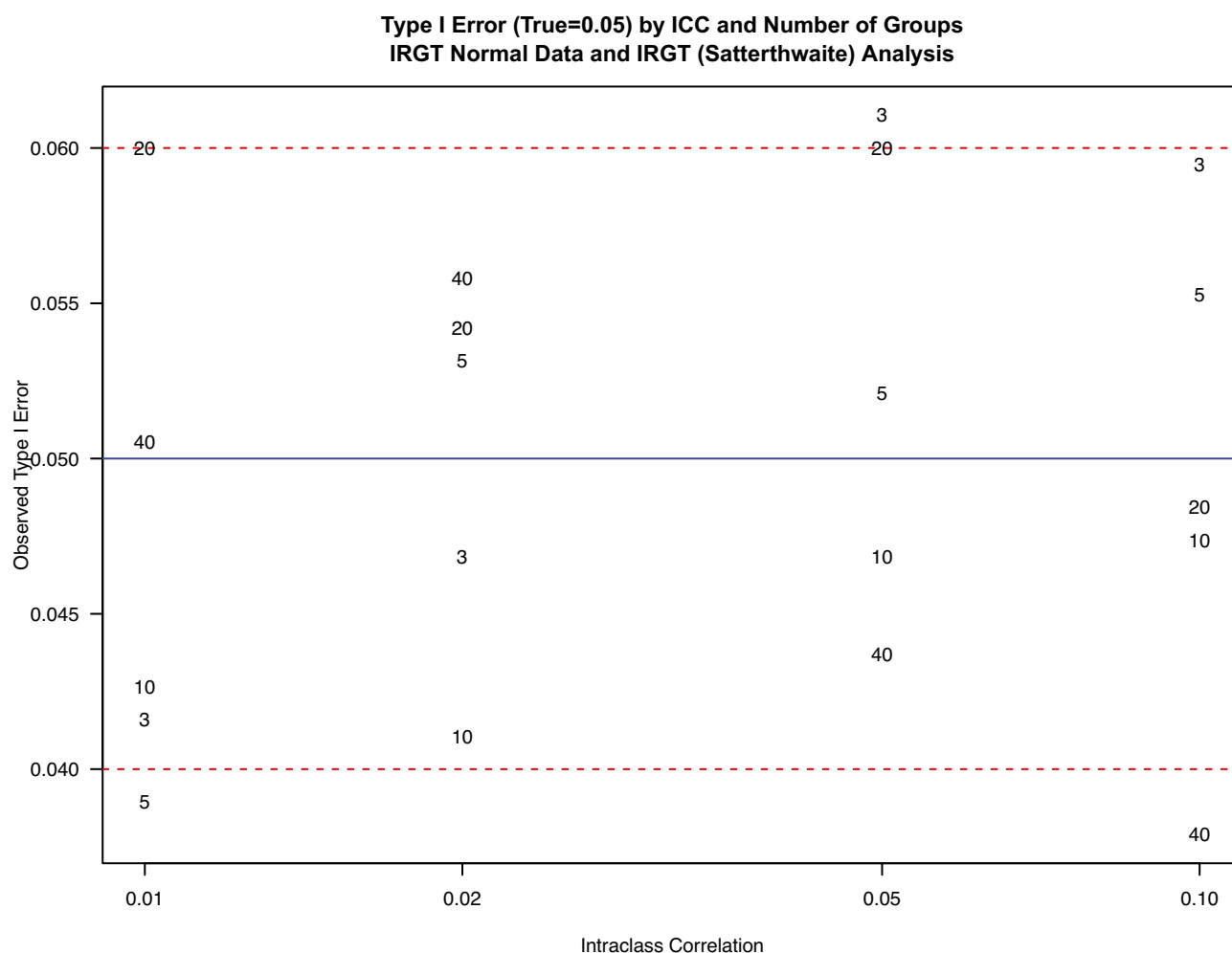$\varepsilon_i = 0$ for $i = 0$, the control subjects.

Type I Error (True=0.05) by ICC and Number of Groups
IRGT Normal Data and Fixed Effect (Wrong) Analysis

Fig. 1 Type I error (true = 0.05) by ICC and number of groups. IRGT normal data and fixed effect (wrong) analysis

Type I Error (True=0.05) by ICC and Number of Groups
IRGT Normal Data and IRGT (Satterthwaite) Analysis

Fig. 2 Type I error (true = 0.05) by ICC and number of groups. IRGT normal data and IRGT (Satterthwaite) analysis

All simulations used 1900 replications, large enough to estimate Type I error to $+/-0.01$. We fix $\sigma_W = 1$ and vary σ_B so that ICC = 0.01, 0.02, 0.05, and 0.1, ranging from very small to very large. As an example, a review of ICCs (Murray & Blitstein, 2003) identified an ICC for a school-based smoking prevention study of 0.026. A smaller ICC was found for a community-based alcohol reduction program (ICC = 0.008). Based on their review of 23 published cluster randomized studies from the fields of psychology and public health, the average ICC = 0.011 for youth studies and ICC = 0.013 for adult studies.

Figure 1 shows the mean Type I errors in testing for a treatment effect using an incorrect model that ignores grouping effects entirely (i.e., Row 1 of Table 1). Each point on this plot represents a different number of groups (3, 5, 10, 20, 40) each with $n = 40$ subjects per group, and the ICC is given on the abscissa. A constant group size was used because power is driven largely by the number of groups and not the group size. The nominal Type I error is 0.05,

and we surround this value with 95% confidence interval lines for this size simulation at the bottom of the figure. The actual Type I errors for this linear model are far above these bounds, indicating an extreme bias of 5-times larger error rate when ICC = 0.1 and even a 50% increase when ICC is a meager 0.01. Thus, ignoring the grouping in this IRGT design leads to the use of smaller critical values than appropriate and opens the door to higher rates of rejection of the null than appropriate.

In contrast, Fig. 2 shows that the inclusion of the proper random effect to account for groups in one arm of the IRGT design does provide accurate critical values. Here, we use the Satterthwaite approximation to adjust the degrees of freedom for a Wald-type test as the standard error for testing $\beta = 0$. This test depends on a combination of the two variances σ_W^2 and σ_B^2, in contrast to standard linear models that depend on only one variance estimate for σ_W^2. Virtually all simulated means fall within or near the acceptable bounds for this simulation study across all values of ICC and numbers of groups.

Figure 3 (see online supplement) shows a similar plot with the same statistical model that accounts for a random group effect in the one arm of an IRGT design but uses the Kenward-Roger (KR) approximation to adjust for imprecision in the two variance estimates (Kenward & Roger, 1997). Unlike the Satterthwaite approximation, which compares the standard Wald-type test of the estimate to its standard error, KR uses a scaled *F*-test based on a likelihood ratio test statistic. Using the KRmodcomp() function that is recommended in R's package pbkrtest (Halekoh & Højsgaard, 2014), this instantiation of the KR test behaves poorly across low numbers of groups where the observed Type I error is substantially smaller than the true value.

In an online supplement, Appendix 2, we provide computer code in R, SAS, SPSS, STATA, Mplus, and SuperMix for analyzing IRGT trials using the best available approximation method, first using a univariate outcome followed by that for a growth model. A sample output is also provided.

Other Types of Contextually Driven Designs

This paper has focused so far on the IRGT design since this is likely the most common and often incorrectly analyzed contextually driven design in prevention research. But a wide range of other designs exists where context also matters in analysis. Here, we describe several variants and refer the reader to the online supplement Appendix 3 and Table 2 for detailed examples of each variation mentioned below and a taxonomy to help individuals recognize the diverse ways that context enters into such trial designs.

We have already examined how shared contexts for interventions that precede randomization (e.g., group-randomized trials) and those formed post randomization (e.g., IRGT designs) require handling of this non-independence in analysis. Both of these situations treat the groupings as static, either before or after randomization, whereas even in pre-existing classrooms, there are continual entrances and exits, and some interventions with groups formed after randomization are formulated to allow rolling entries and exits so that individuals are exposed to intervention components in different orders. Analyses for these rolling groups still need to account for such overlap (Brown et al., 2008a).

There are several important variations of the IRGT design. One involves randomization at units larger than the individual level to intervention but then delivers an intervention in these larger groups. A common example in implementation research is a partially clustered randomized trial in which organizations (e.g., primary care sites) are randomly assigned either to their own implementation strategy or assigned to participate in a learning collaborative consisting of similar organizations (Brown et al., 2014; Ebert et al., 2012; Saldana & Chamberlain, 2012). In the learning collaborative arm, multiple clusters of organizations participate jointly to support delivery of an intervention (e.g., leaders from different primary care sites share experiences and learn from each other). Sites in the same learning collaborative cannot be considered as having outcomes independent of one another. An analysis of a trial containing multiple groups of learning collaboratives in one arm that ignores this non-independence will also have a higher Type I error than intended and consequently reject the null too often. Another modification of the IRGT design is one where both arms are delivered in a group format, in which case random effects in both arms need to be accounted for. Finally, some interventions are delivered partially in group format while some components are individually delivered. As mentioned in online supplement, Appendix 3, mediational analyses can potentially distinguish the two components' effects. It is important to note that during the COVID pandemic, a number of group-delivered trials are being held using synchronous virtual groups, which still require accounting for non-independence. Other group interventions have been transformed to a taped setting to be viewed individually. These do not require special analysis.

Another broad category of contextually driven interventions includes those delivered through a network as opposed to a specified group, where communications vary from person to person. There are also designs that targeted one person but can have potential effects on others, say siblings in the family. Analysis of such spillover trials needs to account for such nesting. In the discussion, we extend this idea of spillover trials to an innovative new class of multiplicative implementation strategies that proactively expand the reach of most preventive interventions.

Discussion

There are a wide number of interventions that are delivered in contexts of groups or utilize a social network to deliver the intervention. Trials that test these interventions need to account for these settings even if randomization is at the individual level. It takes only a very small ICC to have a large effect on statistical tests. A tiny ICC of 0.01 can inflate the type I error rate by 50%, when it is nominally 0.05%. This is identical to using a critical value of 1.78 rather than 1.96 in standard *z*-testing, which shortens the standard confidence interval by 9%. With such a small ICC of 0.01, it would be relatively rare for a trial to detect this as significantly larger than 0. With larger ICCs of 0.02 commonly encountered in prevention trials, the reduction in the width of the confidence interval is 16% and for very large ICCs of 0.1, the reduction in confidence interval width is only 44% of what it should be.

One must be careful to specify the random effect appropriately, as simply including a random intercept in the

model to account for group effects, and pooling those in the control group into one large group, also leads to an incorrect analysis for IRGT designs. This can have the opposite effect of making the standard error much too large — in our example in Table 1 Row 3, it is nearly 13 times too large — and therefore, even large treatment effects can often be judged to be non-significant. While modern statistical packages do have the capacity to fit appropriate mixed-effects models for IRGTs, exact specification in these packages and the literature is notably absent. To aid investigators, we provide useful code for analyzing IRGTs; in our online supplement Appendix 2, we provide code for six commonly used statistical packages when there is a single outcome and when there is an underlying linear growth model. Some of the coding is subtle, but all of these packages produce the same results on test datasets, except for negligible rounding error.

This paper has limitations. While the simulation includes a wide range of ICCs and numbers of groups, we did not investigate all of the possible designs in our simulation study, so as yet we do not know how appropriate the Satterthwaite approximation is with designs that are very imbalanced, involve nonlinear models, or handle different missing data mechanisms. We have not investigated the behavior of the KR method using other statistical packages besides R. We also have not investigated their behavior when using generalized linear models (e.g., logistic regression), but the code changes in most packages are straight forward.

An alternative general approach to account for clustering that covers most of the examples in Table 1, which we have not investigated, is the generalized estimating equations (GEE) approach. In particular, for the classic GEE1 approach, the point estimates ignore clustering completely and then adjust the standard errors for non-independence using "sandwich type estimators" (Liang & Zeger, 1986; Rubin, 1980). This provides an alternative approach to account for clustering than that using random effects. For linear models, these two approaches should provide similar findings. However, in nonlinear models, these two estimates can differ and need to be interpreted differently (Fitzmaurice et al., 2012). Bayesian approaches are also appropriate but not discussed here due to the added complexity and less well-known statistical programs that are available. Details on causal inference assumptions and their violations for these context-driven designs are beyond the scope of this paper.

Conclusions

We close by providing recommendations for diverse audiences. Trialists have a responsibility to conduct a trial that is not only ethical but also likely to produce scientifically useful information. Without accounting for group context in the study design, the study can be woefully underpowered. Simply increasing group size or size of the control group without increasing the number of groups has very limited effects on power. It is important to calculate statistical power based on the critical value for the test statistic that corresponds to the specified Type I error rate. In the simulations that we have done, we found the Satterthwaite approximation does an excellent job and therefore recommend its use.

For developers of implementation strategies for either behavioral or biomedical interventions, we call for a reconceptualization of group- and network-based interventions that could lead to an expanded effect of prevention across a larger population and persist for a longer time period. Two major challenges in scaling up effective preventive interventions have been recognized (Chambers et al., 2013). Voltage drop implies that intervention effects often weaken as they move from efficacy to effectiveness to wide-scale use. Program drift is a phenomenon that often affects programs over time as adaptations that naturally occur in a manualized intervention lead to weakening of its impact. An appropriate implementation design that uses context proactively could potentially reverse these diminishing forces.

In particular, we introduce the new term *multiplicative implementation strategy* to represent strategies whose components are deliberately designed to target individuals beyond those who would ordinarily be direct recipients and thereby to extend the intervention's reach, effectiveness, or system-level sustainment and scale-up. For example, peer-based delivery of behavioral interventions for HIV prevention has had a long history of success, with peer leaders being identified sociometrically as those most influential within their social networks (Amirkhanian et al., 2005; Kelly et al., 1997). These trusted leaders then receive up to 14 h of training in communicating within their networks on how to deliver messages regarding safer sex behaviors (Amirkhanian et al., 2005) or use pre-exposure prophylaxis (PrEP) to prevent HIV infections among those at risk (Kelly et al., 2020). Because PrEP users are highly likely to be networked with other PrEP users (Schueler et al., 2019), it is feasible for one peer-leading PrEP user to recruit others in her network to become leaders as well. To convert a peer delivered to a multiplicative implementation, a community-based organization could not only identify and train peer leaders to talk to others about using a preventive intervention (e.g., PrEP) but also to train them on how to select, motivate, and train a next generation to become peer leaders themselves.

We believe that such a strategy to activate existing friendship networks could be effective when the target population is difficult for the research team to reach directly, when the evidence-based intervention has clear benefit, but it has low use in a segment of the population who could benefit. These conditions all occur for PrEP, which has a 99% success rate when taken regularly (Grant et al., 2014) and is applicable to

young African American/Black men who have sex with men, who as a group are among those having the highest risk for infection and are not often reached by our current medical system (Ezennia et al., 2019).

Multiplicative implementation strategies could be applied to other prevention challenges involving difficult to reach networks for which research has had limited success. These include strategies for preventing adolescent deaths or suicide attempts where youth are often made aware of their friend's suicidality, and these friends could be instrumental in getting help from trusted adults (Pickering et al., 2018). Likewise, for opioid overdose, getting naloxone, a highly effective rescue medication, into the hands of friends and family members, could be an important community strategy to protect those who are unlikely to receive this rescue medication from emergency medical services or police (Irvine et al., 2019). A strategy pertaining to COVID-19 vaccination involves focusing on the head of a multigenerational family to have all family members receive a vaccine at the same time. These multiplicative approaches could well have applicability beyond peer-based implementation policy strategies. If program developers and implementation researchers were to examine how our existing group or network based prevention programs could expand their current reach, we could enlarge our field's population preventive effect (Faraone et al., 2002).

Moving to another audience, meta-analysts and other synthesis analysts need to recognize that reported standard errors for IRGT interventions are often biased and therefore could lead to overly optimistic conclusions, especially when using fixed-effects meta-analysis, which are highly sensitive to a very limited number of small standard errors, rather than mixed-effects meta-analysis models, which are much less sensitive (Brown et al., 2008b).

The other type of synthesis approach is known as integrative data analysis (IDA), individual participant data meta-analysis, individual patient data meta-analysis, or individual level meta-analysis. By their names, this synthesis requires combining individual level data from all the trials, and it too requires accounting for nonindependence wherever it occurs. Compared to meta-analysis, IDA synthesis that combines individual level data from multiple trials has much greater precision to assess intervention effects across all subjects and within distinct subgroups. While IDA for multiple trials routinely incorporates random-effects representing each trial's distinct intercept and growth pattern (Brincks et al., 2018), it is reasonable to account for cross-condition differences in variance structure due to some trials being conducted as IRGTs. Regarding an IDA involving multiple trials with different types of non-independence, it may be quite challenging to estimate as well as even specify separate variances of all the required random effects together in one analysis given current statistical packages. We suggest two alternatives to the random effect modeling approach throughout most

of this paper. The simplest is to compute individual level effect sizes and appropriate standard errors for each single trial and then combine as one would using standard meta-analysis. A more integrative approach would be to use a generalized estimating equation approach (GEE1), whereby the estimator of, say, overall slope difference over time for interventions versus control, is computed by ignoring all clustering, while its variance is computed using the so-called sandwich-type estimator that does take into account non-independence. Both the point estimate and its variance are accounted for in the analysis (Brincks et al., 2018). This computational approach may be most useful when examining moderator effects across a collection of interventions. In particular, when a synthesis is examining differential impact between one smaller subpopulation having few subjects per trial, and the remaining larger population, GEE1 may be the only practical way to account for clustering.

Accounting for contextual effects in intervention trials is particularly important when considering and conducting an IDA that uses an intersectional perspective where group membership may include small numbers of subjects sharing the intersections of race/ethnicity, gender identity, sexual orientation, and other socio-demographic characteristics. Such clustering also violates the assumptions of independence highlighted in this paper, given that individuals within these intersections may share similar characteristics which, if not accounted for, can result in high intraclass correlations and biased estimates of intervention effects. Greater attention to the methodological implications of intersectionality in intervention research is needed to elucidate the true impact of interventions for marginalized communities and to obtain accurate estimates of intervention effects (Schrager et al., 2019).

Science policy makers and journal editors should be aware that trials that fail to account for variation in groups constructed after assignment will produce too many significant findings similar to group-based randomized trials that fail to account for grouping. Funders need to be aware that the many advantages of interventions that are delivered in group or network settings come at the price of larger and sometimes more expensive designs. Science writers and also journalists should have basic awareness of this issue when describing findings to the public.

Methodologists have a major opportunity to expand our causal modeling approaches; the vast majority of this work is based on a stable unit treatment value assumption (SUTVA) (Rubin, 1980), which never holds in group-based interventions. For interventions that are only partly delivered in group settings, and ones where group composition changes with individuals entering and exiting over time, general statistical inference frameworks are in development (Basse & Feller, 2018; Basse et al., 2019; Benjamin-Chung et al., 2018; Hudgens & Halloran, 2008; VanderWeele &

Christakis, 2019; Vanderweele et al., 2013). While methodologic work has guided the analysis of secondary effects in spillover trials (Vanderweele et al., 2013), methods for evaluation of multiplicative implementation strategies, which proactively deliver their interventions, are less developed.

Declarations

Ethics Approval NA.

Consent to Participate NA.

Conflict of Interest Robert Gibbons and Don Hedeker receive royalties from Scientific Software International for their authorship of the Super-Mix computer program. All other authors declare no competing interests.

References

Amirkhanian, Y. A., Kelly, J. A., Kabakchieva, E., Kirsanova, A. V., Vassileva, S., Takacs, J., ... Mocsonaki, L. (2005). A randomized social network HIV prevention trial with young men who have sex with men in Russia and Bulgaria. *AIDS, 19*(16), 1897–1905.

Andridge, R. R., Shoben, A. B., Muller, K. E., & Murray, D. M. (2014). Analytic methods for individually randomized group treatment trials and group-randomized trials when subjects belong to multiple groups. *Statistics in Medicine, 33*, 2178–2190. https://doi.org/10.1002/sim.6083

Basse, G. W., & Feller, A. (2018). Analyzing two-stage experiments in the presence of interference. *Journal of the American Statistical Association, 113*, 41–55.

Basse, G. W., Feller, A., & Toulis, P. (2019). Randomization tests of causal effects under interference. *Biometrika, 106*, 487–494. https://doi.org/10.1093/biomet/asy072

Benjamin-Chung, J., Arnold, B. F., Berger, D., Luby, S. P., Miguel, E., Colford, J. M., Jr., & Hubbard, A. E. (2018). Spillover effects in

epidemiology: Parameters, study designs and methodological considerations. *International Journal of Epidemiology, 47*, 332–347.

Botvin, G. J. (2000). Preventing drug abuse in schools: Social and competence enhancement approaches targeting individual-level etiologic factors. *Addictive Behaviors, 25*, 887–897.

Brincks, A., Montag, S., Howe, G. W., Huang, S., Siddique, J., Ahn, S., ... Brown, C. H. (2018). Addressing methodologic challenges and minimizing threats to validity in synthesizing findings from individual level data across randomized trials. *Prevention Science, 19*(Supplement 1), 60–73. https://doi.org/10.1007/s11121-017-0769-1

Brown, C. H., Chamberlain, P., Saldana, L., Padgett, C., Wang, W., & Cruden, G. (2014). Evaluation of two implementation strategies in 51 child county public service systems in two states: Results of a cluster randomized head-to-head implementation trial. *Implementation Science, 9*, 134. https://doi.org/10.1186/s13012-014-0134-8

Brown, C. H., Wang, W., Kellam, S. G., Muthén, B. O., Petras, H., Toyinbo, P., … Prevention Science and Methodology Group. (2008a). Methods for testing theory and evaluating impact in randomized field trials: Intent-to-treat analyses for integrating the perspectives of person, place, and time. *Drug and Alcohol Dependence, 95*, S74–S104. https://doi.org/10.1016/j.drugalcdep.2007.11.013

Brown, C. H., Wang, W., & Sandler, I. (2008b). Examining how context changes intervention impact: The use of effect sizes in multilevel meta-analysis. *Child Development Perspectives, 2*, 198–205. https://doi.org/10.1111/j.1750-8606.2008.00065.x

Chamberlain, P., & Reid, J. B. (1998). Comparison of two community alternatives to incarceration for chronic juvenile offenders. *Journal of Consulting & Clinical Psychology, 66*, 624–633.

Chambers, D. A., Glasgow, R., & Stange, K. (2013). The dynamic sustainability framework: Addressing the paradox of sustainment amid ongoing change. *Implementation Science, 8*, 117. https://doi.org/10.1186/1748-5908-8-117

Dishion, T. J., McCord, J., & Poulin, F. (1999). When interventions harm. Peer groups and problem behavior. *American Psychologist, 54*(9), 755–764.

Ebert, L., Amaya-Jackson, L., Markiewicz, J., & Fairbank, J. A. (2012). Development and application of the NCCTS learning collaborative model for the implementation of evidence-based child trauma treatment. In R. K. McHugh & D. H. Barlow (Eds.), *Dissemination and implementation of evidence-based psychological interventions* (pp. 97–123). Oxford University Press.

Ezennia, O., Geter, A., & Smith, D. K. (2019). The PrEP care continuum and black men who have sex with men: A scoping review of published data on awareness, uptake, adherence, and retention in PrEP care. *AIDS and Behavior, 23*, 2654–2673. https://doi.org/10.1007/s10461-019-02641-2

Faraone, S. V., Brown, C. H., Glatt, S. J., & Tsuang, M. T. (2002). Preventing schizophrenia and psychotic behaviour: Definitions and methodological issues. *Canadian Journal of Psychiatry, 47*, 527–537.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis* (vol. 998): John Wiley & Sons.

Gibbons, R. D., Hedeker, D., Waternaux, C., & Davis, J. M. (1988). Random regression models: A comprehensive approach to the analysis of longitudinal psychiatric data. *Psychopharmacology Bulletin, 24*, 438–443.

Grant, R. M., Anderson, P. L., McMahan, V., Liu, A., Amico, K. R., Mehrotra, M., ... Montoya, O. (2014). Uptake of pre-exposure prophylaxis, sexual practices, and HIV incidence in men and transgender women who have sex with men: A cohort study. *The Lancet Infectious Diseases, 14*(9), 820–829.

Halekoh, U., & Højsgaard, S. (2014). A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models – The R package pbkrtest. *Journal of Statistical Software, 59*(9), 1–30. Retrieved January 20, 2022, from https://www.jstatsoft.org/v59/i09/

Hawkins, J. D., Jenson, J. M., Catalano, R., Fraser, M. W., Botvin, G. J., Shapiro, V., ... Stone, S. (2015). *Unleashing the power of prevention:*

Discussion paper. Retrieved from Washington, DC. Retrieved January 20, 2022, from https://nam.edu/perspectives-2015-unleashing-the-power-of-prevention/

Hawkins, J. D., Kosterman, R., Catalano, R. F., Hill, K. G., & Abbott, R. D. (2005). Promoting positive adult functioning through social development intervention in childhood: Long-term effects from the Seattle Social Development Project. *Archives of Pediatrics & Adolescent Medicine, 159*, 25–31.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Wiley-Interscience.

Hudgens, M. G., & Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association, 103*, 832–842.

Hunter, R. F., de la Haye, K., Murray, J. M., Badham, J., Valente, T. W., Clarke, M., & Kee, F. (2019). Social network interventions for health behaviours and outcomes: A systematic review and meta-analysis. . *PLoS Medicine, 16*(9).

Irvine, M. A., Kuo, M., Buxton, J. A., Balshaw, R., Otterstatter, M., Macdougall, L., ... Tyndall, M. (2019). Modelling the combined impact of interventions in averting deaths during a synthetic-opioid overdose epidemic. *Addiction, 114*(9), 1602–1613.

Kellam, S. G., & Brown, C. H. (1986). Social adaptational and psychological antecedents in the first grade of adolescent psychopathology ten years later. In G. L. Klerman (Ed.), *Suicide and depression among adolescents and young adults*. Washington, D.C.: American Psychiatric Press.

Kellam, S. G., Brown, C. H., Poduska, J. M., Ialongo, N. S., Wang, W., Toyinbo, P., ... Wilcox, H. C. (2008). Effects of a universal classroom behavior management program in first and second grades on young adult behavioral, psychiatric, and social outcomes. *Drug and Alcohol Dependence, 95*(Suppl 1), S5–S28. https://doi.org/10.1016/j.drugalcdep.2008.01.004

Kellam, S. G., Koretz, D., & Mościcki, E. K. (1999). Core elements of developmental epidemiologically based prevention research. *American Journal of Community Psychology, 27*, 463–482. https://doi.org/10.1023/A:1022129127298

Kellam, S. G., Rebok, G. W., Ialongo, N., & Mayer, L. S. (1994). The course and malleability of aggressive behavior from early first grade into middle school: Results of a developmental epidemiology-based preventive trial. *Journal of Child Psychology & Psychiatry & Allied Disciplines, 35*, 259–281. https://doi.org/10.1111/j.1469-7610.1994.tb01161.x

Kelly, J. A., Amirkhanian, Y. A., Walsh, J. L., Brown, K. D., Quinn, K. G., Petroll, A. E., ... Ertl, T. (2020). Social network intervention to increase pre-exposure prophylaxis (PrEP) awareness, interest, and use among African American men who have sex with men. *AIDS Care, 32*(sup2), 40–46.

Kelly, J. A., Murphy, D. A., Sikkema, K. J., McAuliffe, T. L., Roffman, R. A., Solomon, L. J., ... Collaborative, T. C. H. P. R. (1997). Randomised, controlled, community-level HIV-prevention intervention for sexual-risk behaviour among homosexual men in US cities. *The Lancet, 350*(9090), 1500–1505.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 983–997.

Lee, K. J., & Thompson, S. G. (2005). The use of random effects models to allow for clustering in individually randomized trials. *Clinical Trials, 2*, 163–173. Retrieved January 20, 2022, from http://ctj.sagepub.com/content/2/2/163.full.pdf

Li, H., & Hedeker, D. (2017). Statistical methods for continuous outcomes in partially clustered designs. *Communications in Statistics - Theory and Methods, 46*, 3915–3933.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*, 13–22. https://doi.org/10.1093/biomet/73.1.13

Moerbeek, M., & Wong, W. K. (2008). Sample size formulae for trials comparing group and individual treatments in a multilevel model. *Statistics in Medicine, 27*, 2850–2864.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*: Oxford University Press.

Murray, D. M., & Blitstein, J. L. (2003). Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review, 27*, 79–103.

Murray, D. M., Varnell, S. P., & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: A review of recent methodological developments. *American Journal of Public Health, 94*(3), 423–432. Retrieved January 20, 2022, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1448268/pdf/0940423.pdf

Pals, S. L., Murray, D. M., Alfano, C. M., Shadish, W. R., Hannan, P. J., & Baker, W. L. (2008). Individually randomized group treatment trials: A critical appraisal of frequently used design and analytic approaches. *American Journal of Public Health, 98*, 1418–1424. https://doi.org/10.2105/ajph.2007.127027

Pals, S. L., Wiegand, R. E., & Murray, D. M. (2011). Ignoring the group in group-level HIV/AIDS intervention trials: A review of reported design and analytic methods. *AIDS, 25*, 989–996. Retrieved January 20, 2022, from http://graphics.tx.ovid.com/ovftpdfs/FPDDNCFBIAEHIO00/fs047/ovft/live/gv024/00002030/00002030-201104240-00012.pdf

Petrosino, A., Turpin-Petrosino, C., Hollis-Peel, M. E., & Lavenberg, J. G. (2013). Scared straight and other juvenile awareness programs for preventing juvenile delinquency: A systematic review. *Campbell Systematic Reviews, 9*, 1–55.

Pickering, T. A., Wyman, P. A., Schmeelk-Cone, K., Hartley, C., Valente, T. W., Pisani, A. R., ... LoMurray, M. (2018). Diffusion of a peer-led suicide preventive intervention through school-based student peer and adult networks. *Frontiers in Psychiatry, 9*, 598–598. https://doi.org/10.3389/fpsyt.2018.00598

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Roberts, C., & Roberts, S. A. (2005). Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials, 2*, 152–162. Retrieved January 20, 2022, from http://ctj.sagepub.com/content/2/2/152.full.pdf

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association, 75*, 591–593.

Saldana, L., & Chamberlain, P. (2012). Supporting implementation: The role of community development teams to build infrastructure. *American Journal of Community Psychology, 50*, 334–346. https://doi.org/10.1007/s10464-012-9503-0;10.1007/s10464-012-9503-0

Sandler, I. N., Ingram, A., Wolchik, S., Tein, J. Y., & Winslow, E. (2015). Long-term effects of parenting-focused preventive interventions to promote resilience of children and adolescents. *Child Development Perspectives, 9*, 164–171.

Schrager, S. M., Steiner, R. J., Bouris, A. M., Macapagal, K., & Brown, C. H. (2019). Methodological considerations for advancing research on the health and wellbeing of sexual and gender minority youth. *LGBT Health, 6*, 156–165. https://doi.org/10.1089/lgbt.2018.0141

Schueler, K., Ferreira, M., Nikolopoulos, G., Skaathun, B., Paraskevis, D., Hatzakis, A., ... Schneider, J. A. (2019). Pre-exposure prophylaxis (PrEP) awareness and use within high HIV transmission networks. *AIDS and Behavior, 23*(7), 1893–1903.

Szapocznik, J., & Coatsworth, J. D. (1999). An ecodevelopmental framework for organizing the influences on drug abuse: A developmental model of risk and protection. In *Drug abuse: Origins & interventions* (pp. 331–366). Washington, DC, US: American Psychological Association.

Turner, E. L., Li, F., Gallis, J. A., Prague, M., & Murray, D. M. (2017). Review of recent methodological developments in group-randomized

trials: Part 1-design. *American Journal of Public Health, 107*, 907–915. https://doi.org/10.2105/ajph.2017.303706

VanderWeele, T. J., & Christakis, N. A. (2019). Network multipliers and public health. *International Journal of Epidemiology, 48*, 1032–1037. https://doi.org/10.1093/ije/dyz010

Vanderweele, T. J., Hong, G., Jones, S. M., & Brown, J. L. (2013). Mediation and spillover effects in group-randomized trials: A case study of the 4Rs educational intervention. *Journal of the American Statistical Association, 108*, 469–482.

Young, S. D., Cumberland, W. G., Lee, S. J., Jaganath, D., Szekeres, G., & Coates, T. (2013). Social networking technologies as an emerging tool for HIV prevention: A cluster randomized trial. *Annals of Internal Medicine, 159*, 318–324. https://doi.org/10.7326/0003-4819-159-5-201309030-00005