Trials involving Group Randomization or Delivery of Interventions to Groups: GRTs, IRGTs, and SW-GRTs

David M. Murray, Ph.D.
Associate Director for Prevention
Director, Office of Disease Prevention
National Institutes of Health

2019 OBSSR Summer Institute
July 15, 2019

National Institutes of Health
Office of Disease Prevention

Today's Presentation

- Design and analytic issues
- Planning the trial
- Analytic methods
- Power
- The new Application Guide and Review Criteria
- Non-randomized design alternatives

Three Kinds of Randomized Trials

- Randomized Clinical Trials (RCTs)
  - Individuals randomized to study conditions with no interaction among participants after randomization (no group sessions, virtual interaction, or shared intervention agent)
    - Most drug trials
- Individually Randomized Group Treatment Trials (IRGTs)
  - Individuals randomized to study conditions with interaction among participants after randomization or with a shared intervention agent
    - Many surgical trials
    - Many behavioral trials
- Group-Randomized Trials (GRTs)
  - Groups randomized to study conditions with interaction among the members of the same group before and after randomization
    - Many trials conducted in communities, worksites, schools, clinics, etc.

Two Kinds of Group-Randomized Trials

- Parallel GRT
  - Separate but parallel intervention and control conditions throughout the trial, with no crossover.
- Stepped Wedge GRT
  - All groups start in the control condition.
  - All groups crossover to the intervention condition, but in a random order and on a staggered schedule.
  - All groups receive the intervention before the end of the study.

Alternative Labels

- Individually randomized controlled trials are also called….
  - Randomized controlled trials,
  - Randomized clinical trials,
  - Controlled clinical trials.
  - These labels are interchangeable.
- Individually randomized group treatment trials are also called…
  - Partially nested designs or partially clustered designs.
  - IRGT is the more general label.
- Group-randomized trials are also called…
  - Cluster-randomized trials,
  - Community trials.
  - These labels are interchangeable.

Impact on the Design

- Randomized clinical trials
  - There is usually good opportunity for randomization to distribute potential confounders evenly, as most RCTS have N>100.
  - If well executed, confounding is not usually a concern.
- Individually randomized group treatment trials
  - There may be less opportunity for randomization to distribute potential confounders evenly, as many IRGTs have N<100.
  - Confounding can be more of a concern in IRGTs than in RCTs.

Impact on the Design

- Parallel group-randomized trials
  - GRTs often involve a limited number of groups, often <50.
  - In any single realization, there is limited opportunity for randomization to distribute all potential confounders evenly.
  - Confounding is a concern in GRTs if G<50.
- Stepped wedge GRTs
  - Crossing of groups with study conditions avoids most confounding.
  - However, intervention effects are confounded with calendar time, as more groups are in the intervention condition as the study progresses.
  - SW-GRTs are inherently less rigorous than parallel GRTs and should be considered only when a parallel GRT is not appropriate.

Impact on the Analysis in a GRT or IRGT

- Observations on randomized individuals who do not interact are independent and are analyzed with standard methods.
- The members of the same group in a GRT will share some physical, geographic, social or other connection.
- The members of groups in an IRGT will develop similar connections.
- Those connections will create a positive intraclass correlation that reflects extra variation attributable to the group.

$$ICC_{m:g:c} = corr\left(y_{i:k:l}, y_{i':k:l}\right)$$

- The positive ICC reduces the variation among the members of the same group so the within-group variance is:

$$\sigma_e^2 = \sigma_y^2\left(1 - ICC_{m:g:c}\right)$$

Impact on the Analysis in a GRT or IRGT

- The between-group component is the one's complement:

$$\sigma^2_{g:c} = \sigma^2_y \left(ICC_{m:g:c} \right)$$

- The total variance is the sum of the two components:

$$\sigma^2_y = \sigma^2_e + \sigma^2_{g:c}$$

- The intraclass correlation is the fraction of the total variation in the data that is attributable to the unit of assignment:

$$ICC_{m:g:c} = \frac{\sigma^2_{g:c}}{\sigma^2_e + \sigma^2_{g:c}}$$

Impact on the Analysis in a GRT or IRGT

- Given m members in each of g groups...
  - When group membership is established by random assignment,

$$\sigma^2_{\bar{y}_g} = \frac{\sigma^2_y}{m}$$

  - When group membership is not established by random assignment,

$$\sigma^2_{\bar{y}_g} = \frac{\sigma^2_e}{m} + \sigma^2_g$$

  - Or equivalently,

$$\sigma^2_{\bar{y}_g} = \frac{\sigma^2_y}{m}\left(1 + (m-1)\,ICC\right)$$

Impact on the Analysis in a GRT or IRGT

- Nested factors must be modeled as random effects (Zucker, 1990).
- The variance of any group-level statistic will be larger.
- The df to estimate the group-level component of variance will be based on the number of groups, and so is often limited.
  - This is almost always true in a GRT, can be true in an IRGT.
- Any analysis that ignores the extra variation or the limited df will have a Type I error rate that is inflated, often badly.
  - Type I error rate may be 30-50% in a GRT, even with small ICC
  - Type I error rate may be 15-25% in an IRGT, even with small ICC
- Extra variation and limited df always reduce power.

Zucker DM. An analysis of variance pitfall: The fixed effects analysis in a nested design. Educ and Psych Measurement. 1990;50(4):731-8.

Impact on the Analysis: GRT, IRGT

- Scott & Holt (1982) estimate the effect of the ICC as:

$$DEFF = 1 + (m - 1) ICC_y ICC_x$$

  - DEFF is the ratio of the variance as observed to the variance under simple random sampling.
  - ICC_y is the ICC for the dependent variable.
  - ICC_x is the ICC for the independent variable.

Scott AJ, Holt D. The effect of two-stage sampling on ordinary least squares methods. JASA. 1982;77(380):848-54.

Impact on the Analysis: GRT, IRGT

- For most health related outcomes, ICC values are …
  - 0.00-0.05 for large aggregates (e.g., schools, worksites),
  - 0.05-0.25 for small aggregates (e.g., classrooms, departments),
  - 0.25-0.75 for very small aggregates (e.g., families, spouse pairs).
- ICCs tend to be larger for knowledge and attitudes, smaller for behaviors, and smaller still for physiologic measures.
- If the groups are crossed with the levels of the exposure of interest (most observational studies, SW-GRTs), $ICC_x \approx ICC_y$.
- If the groups are nested within the levels of the exposure of interest (IRGTs, GRTs), $ICC_x=1$, because all members of a group will have the same value for exposure.

Impact on the Analysis: GRT, IRGT

- Given the ICC and m per group, DEFF is…

Surveys			IRGTs			GRTs		
	$ICC_y=ICC_x$			$ICC_x=1$			$ICC_x=1$	
m	0.05	0.01	m	0.25	0.10	m	0.05	0.01
50	1.12	1.00	10	3.25	1.90	20	1.95	1.19
100	1.25	1.01	20	5.75	2.90	100	5.95	1.99
200	1.50	1.02	40	10.75	4.90	500	25.95	5.99

- The usual F-test, corrected for the ICC, is:

$$F_{corrected} = \frac{F_{uncorrected}}{DEFF}$$

Impact on the Analysis for SW-GRTs

- Crossing of groups with study conditions often reduces the impact of the ICC compared to a parallel GRT, either improving power or allowing a smaller study.
- There are other potential sources of bias in the SW-GRT:
  - The intervention is confounded with time.
  - The intervention effect may vary over time.
  - The intervention effect may vary by group.
  - Patterns of correlation may vary over time.
- Any analysis that assumes that the intervention effect is constant over time and across groups, and that the pattern of correlation is constant, may be biased.
- Compared to a parallel GRT, SW-GRTs are at greater risk to the effects of external events that affect the outcomes of the trial.

The Warning

Randomization by cluster accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception, however, and should be discouraged.

Cornfield (1978)

- Though Cornfield's remarks were addressed only to GRTs, they also apply to IRGTs, and to SW-GRTs

Cornfield J. Randomization by group: a formal analysis. Am J Epi. 1978;108(2):100-2.

The Need for GRTs, IRGTs, and SW-GRTs

- An RCT is the best comparative design when individual randomization is possible without post-randomization interaction.
- An IRGT is the best comparative design whenever...
  - Individual randomization is possible but there are good reasons to deliver the intervention in a group format or through a shared interventionist
- A GRT is the best comparative design whenever the investigator wants to evaluate an intervention that…
  - Manipulates the social or physical environment or cannot be delivered to individuals without risk of contamination
- An SW-GRT is an alternative to a parallel GRT if…
  - Preliminary evidence makes it unethical to withhold the intervention.
  - It is impossible to implement the intervention in all groups simultaneously.
  - External events are unlikely to affect the outcomes before the end of the trial.

The Challenge

- The challenge is to create trials that are:
  - Rigorous enough to avoid threats to validity of the design,
  - Analyzed to avoid threats to statistical validity,
  - Powerful enough to provide an answer to the question,
  - And inexpensive enough to be practical.

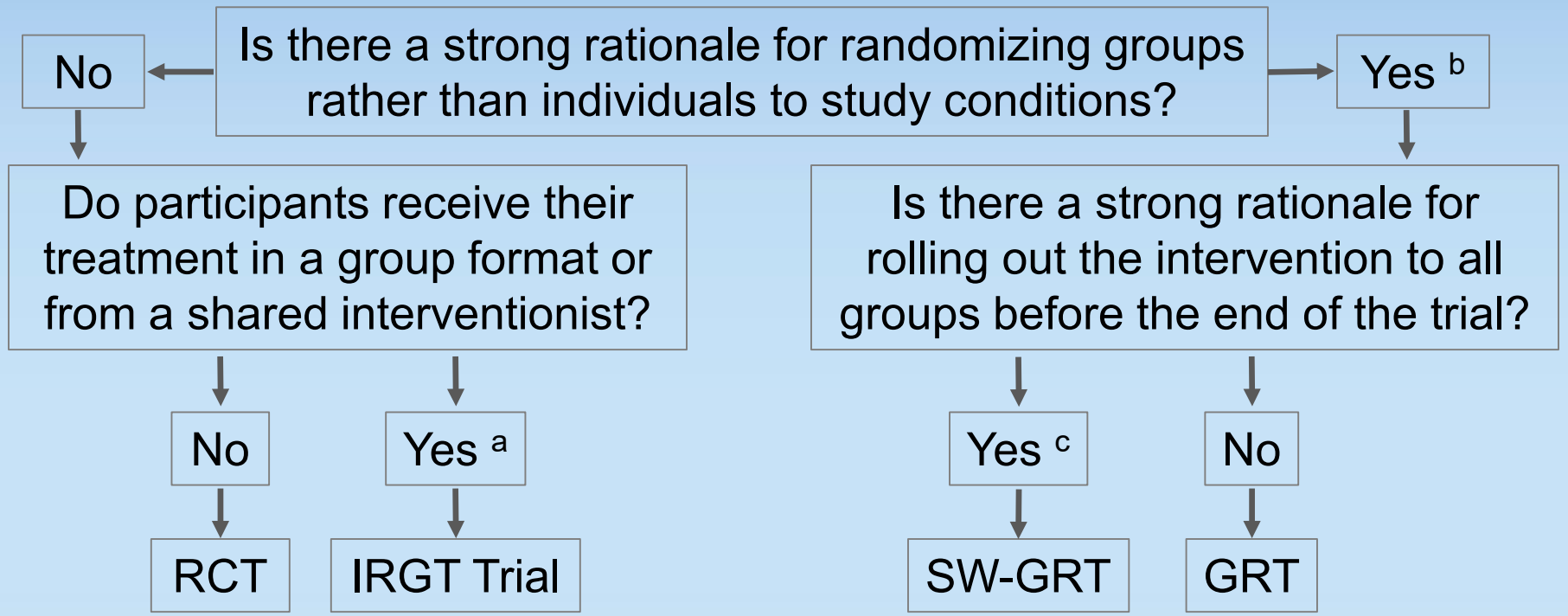
Planning the Trial

- The driving force must be the research question.
  - The question will identify the target population, the setting, the endpoints, and the intervention.
  - Those factors will shape the design and analytic plan.
- The primary criteria for choosing that question should be:
  - Is it important enough to do?
    - Will the trial address an important public health question?
    - Will the results advance the field?
  - Is this the right time to do it?
    - Is there preliminary evidence of feasibility and efficacy for the intervention?
    - Are there good estimates for the parameters needed to size the study?
- The investigators should proceed only if the answer to both questions is yes, and keep the question in mind.

Fundamentals of Research Design

- The goal in any comparative trial is to allow valid inference that the intervention as implemented caused the result as observed.
- Three elements are required:
  - Control observations
  - A minimum of bias in the estimate of the intervention effect
  - Sufficient precision for that estimate
- The three most important tools to limit bias and improve precision in any comparative trial are:
  - Randomization
  - Replication
  - Variance reduction

Choosing Among These Designs



[a] If the intervention is delivered through a physical 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.

Parallel Group-Randomized Trial Designs

- Single factor and factorial designs
- Time as a factor
- Cross-sectional and cohort designs
- A priori matching and stratification
- Constrained randomization

Single Factor and Factorial Designs

- Most involve only one treatment factor.
  - Condition
- Most have only two levels of that treatment factor.
  - Intervention vs. control.
- Most cross Condition with Time.
  - Nested cohort designs
  - Nested cross-sectional designs
- Some GRTs include stratification factors.
  - Multi-center GRTs cross Condition with Field Center.
  - Single-center GRTs often stratify on factors related to the outcome or to the ease of implementation of the intervention.
- Some IRGTs have post-randomization interaction in one condition only, others have it in both.

Time as a Factor

- Posttest-only design
- Pretest-posttest design
- Extended designs
  - Additional discrete time intervals before and/or after intervention
  - Continuous surveillance

Cross-Sectional and Cohort Designs

- Nested cohort design
  - The research question involves change in specific members.
  - Measure the same sample at each time data are collected.
- Nested cross-sectional design
  - The research question involves change in an entire population.
  - Select a new sample each time data are collected.

Cross-Sectional and Cohort Designs

- Strengths and weaknesses

Cross-section	*Cohort*
in and out migration	mortality
group change	individual change
recruitment costs	tracking and follow-up costs
less powerful?	more powerful?
full dose?	full dose?

A Priori Matching and Stratification

- Rationale
  - Either can be used if the investigators want to ensure balance on a potential source of bias.
  - *A priori* stratification is preferred if the investigators expect the intervention effect to be different across strata.
  - *A priori* matching is useful if the matching factors are well-correlated with the primary endpoint.
  - The choice of matching vs. stratification will often depend on the number of groups available and on the expected correlation.
  - Work by Donner et al. (2007) favors stratification when m<100.

Donner A, Taljaard M, et al. The merits of breaking the matches: a cautionary tale. Stat Med. 2007;26(9):2036-51.

Constrained Randomization

- Stratification and matching are difficult if there are multiple factors and a limited number of groups to be randomized.
- Constrained randomization has been suggested as a solution (Raab and Butcher, 2001).
  - Generate all possible allocations.
  - Identify those that are sufficiently well balanced across conditions on key covariates.
  - Choose one allocation at random to use for the trial.
- Li et al. (2016, 2017) reported constrained randomization improved power and maintained the type 1 error rate.

Raab GM, Butcher I. Balance in cluster randomized trials. Stat Med. 2001;20(3):351-365. PMID11180306.

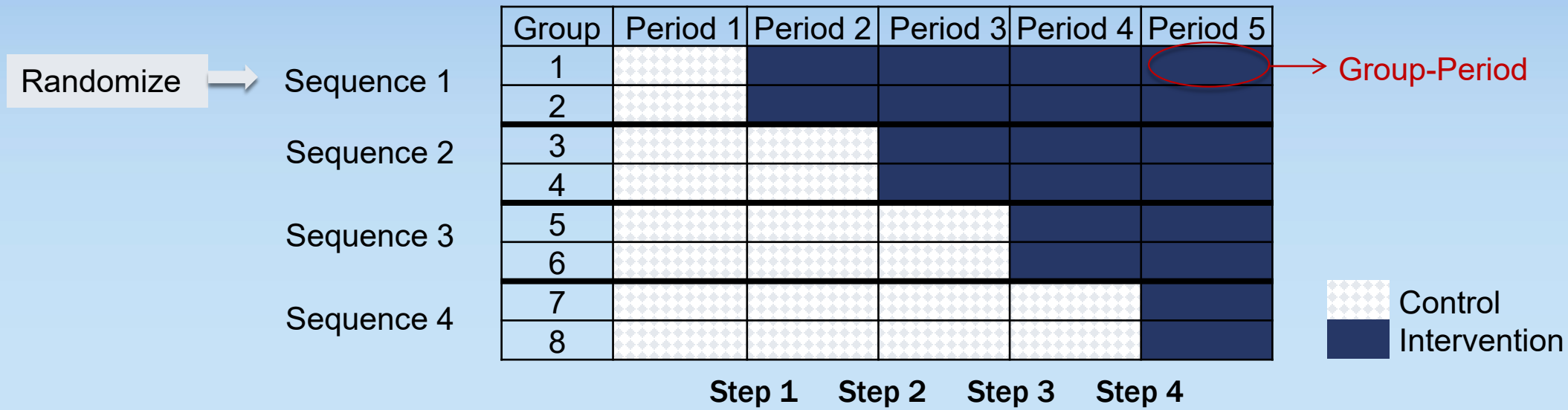
Li F, Lokhnygina Y, et al. An evaluation of constrained randomization for the design and analysis of group-randomized trials. Stat Med. 2016;35(10):1565-79. PMID26598212.

Li F, Turner EL, et al. An evaluation of constrained randomization for the design and analysis of group-randomized trials with binary outcomes. Stat Med. 2017;36(24):3791-806. PMID28786223.

Stepped Wedge Group-Randomized Trial Designs

- The basic stepped wedge design
- Main types of stepped wedge designs
- Key methodological considerations
  - Confounding by time
  - Contamination
  - Time-varying intervention effects
  - Effect heterogeneity
  - Complex correlations

The Basic Design



- Groups are randomized to sequences.
  - This is where matching, stratification, or constrained stratification would be used to improve comparability of the sequences.
- Groups cross to intervention sequentially and in random order, either individually or in sets.
- Outcomes are assessed repeatedly in each group over time.
- All groups provide both intervention and control data.

Main Types of Stepped Wedge Designs

- Cross-sectional design
  - Different individuals are measured each time.
- Cohort design
  - The same individuals are measured each time.
  - Closed cohort: no individuals may join during the trial
  - Open cohort: some individuals may leave and others may join during the trial

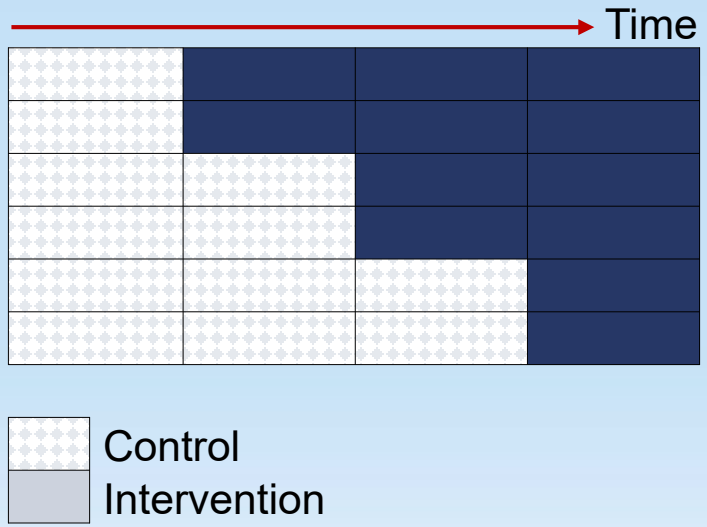
Key Methodological Considerations

- SW-GRTs have several key characteristics that complicate their design and analysis.
  - May increase the risks of bias
  - Need careful justification for the use of this design
  - Need special care in reporting

Hemming, K, Taljaard M, et al. Reporting of The CONSORT extension for Stepped-Wedge Cluster Randomised Trials: Extension of the CONSORT 2010 statement with explanation and elaboration. BMJ. 2018;363:k1614. PMID30413417.

Confounding by Time

- Intervention effect is partially confounded with time.
  - Due to staggered implementation, time is correlated with intervention.
  - Time may also be correlated with outcome (“secular trend”).
- Analysis must always adjust for time (even if not significant).



Chen et al. Secular trends and evaluation of complex interventions: the rising tide phenomenon. BMJ Qual Saf. 2016 May;25(5):303-10.

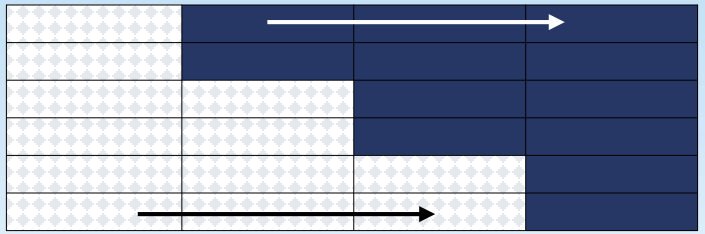
Contamination

- Increased risk of within-group contamination
  - Groups may implement intervention earlier than planned (they can't wait).
  - Groups may implement intervention later than planned (difficulties in implementation).
- As long as contamination is observed and recorded, an "as treated" analysis is possible (but deviates from "Intention-To-Treat").

Copas AJ et al. (2015) Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. Trials; 16:352(1):352. PMID26279154.

Time-Varying Intervention Effects

- Effect of intervention may vary depending on
  - Calendar time
    - Seasonal variation, external events
  - Time since the intervention was introduced
    - Response may increase with more experience.
    - Response may weaken over time (training is forgotten, decrease in adherence).
- An analysis which assumes a constant intervention effect may be biased.



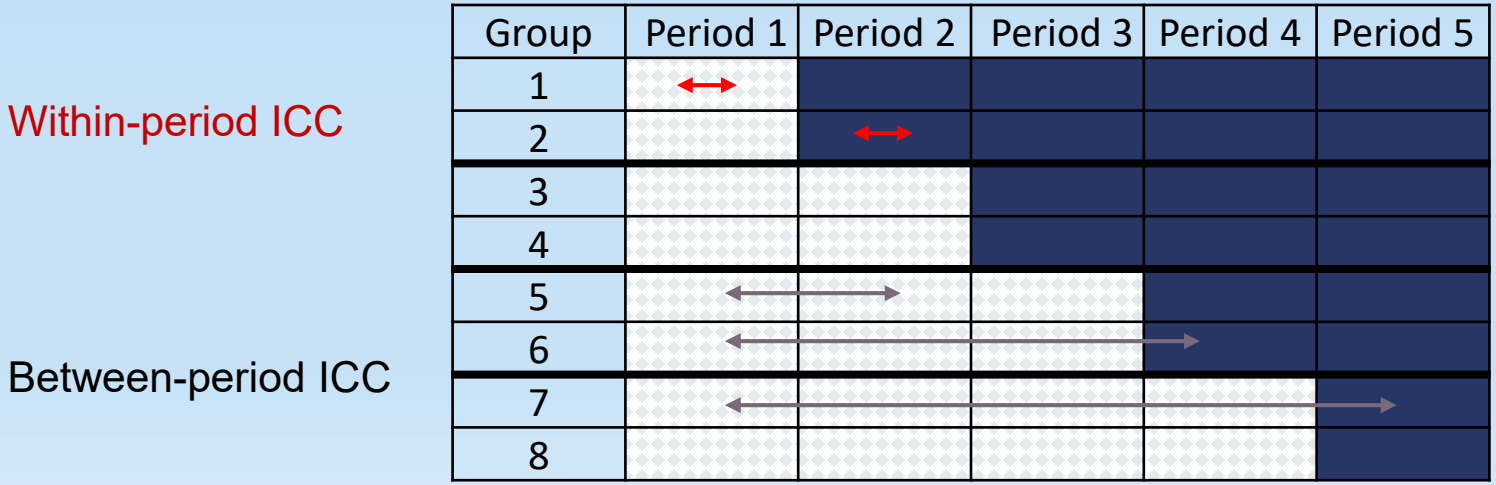
Effect Heterogeneity

- Treatment effect may vary across groups.
  - Variation in quality of implementation, fidelity, other factors
- An analysis which assumes a homogeneous intervention effect across groups may be biased.
- Heterogeneity can reduce power.

Hughes JP, Granston TS, et al. (2015) On the design and analysis of stepped wedge trials. Contem Clinl Trials. 45(Pt A):55-60

Complex Correlations

- Repeated measures on same groups (and possibly same participants)
- Need to account for within-period ICCs as well as between-period ICCs
- Bias can be introduced by mis-specifying the correlation structure.



Individually Randomized Group Treatment Trial Designs

- Post-randomization interaction in one condition
  - Creates a heterogeneous correlation structure
- Post-randomization interaction in both conditions
  - Creates a correlation structure similar to a GRT
- The design features available for GRTs are also available for IRGTs.

Threats to Internal Validity

- Four primary threats in a trial are:
  - Selection refers to pre-existing differences between the study conditions associated with the groups or members that are nested within conditions.
  - Differential history is any external influence other than the intervention that can affect the outcome and that affects one condition more than the other.
    - This is particularly a threat in a stepped wedge group-randomized trial.
  - Differential maturation reflects growth or development at the group or member level that can affect the outcome and that affects one condition more than the other.
  - Contamination exists when important components of the intervention find their way into the control condition, either directly, or indirectly.

Strategies to Protect Internal Validity

- Randomization
- *A priori* matching, stratification, or constrained randomization
  - Of groups in GRTs and SW-GRTs, of members in IRGTs
- Objective measures
- Independent evaluation personnel who are blind to conditions
- Analytic strategies
  - Regression adjustment for covariates
  - In SW-GRTs, regression adjustment for calendar time
- Avoid the pitfalls that invite threats to internal validity
  - Testing and differential testing
  - Instrumentation and differential instrumentation
  - Regression to the mean and differential regression to the mean
  - Attrition and differential attrition

Threats to the Validity of the Analysis

- Misspecification of the analysis model
  - Ignore a measurable source of random variation
  - Misrepresent a measurable source of random variation
  - Misrepresent the pattern of over-time correlation in the data
- Low power
  - Weak interventions
  - Insufficient replication of groups and time intervals
  - High variance or intraclass correlation in endpoints
  - Poor reliability of intervention implementation

Strategies to Protect the Analysis

- Avoid model misspecification
  - Plan the analysis concurrent with the design.
  - Plan the analysis around the primary endpoints.
  - Anticipate all sources of random variation.
  - Anticipate patterns of over-time correlation.
  - Anticipate the pattern of the intervention effect over time.
    - Particularly important with repeated measures designs, including SW-GRTs
  - Assess potential confounding and effect modification.

Strategies to Protect the Analysis

- Avoid low power
  - Employ strong interventions with good reach.
  - Maintain reliability of intervention implementation.
  - Employ more and smaller groups instead of a few large groups.
  - Employ more and smaller surveys or continuous surveillance instead of a few large surveys.
  - For SW-GRTs, employ more steps.
  - Employ regression adjustment for covariates to reduce variance and intraclass correlation, and in SW-GRTs, to adjust for calendar time.

Preferred Analytic Models for Parallel GRT Designs With One or Two Time Intervals

- Mixed-model ANOVA/ANCOVA
  - Extension of the familiar ANOVA/ANCOVA based on the General Linear Model
  - Fit using the General Linear Mixed Model or the Generalized Linear Mixed Model
  - Accommodates regression adjustment for covariates
  - Can not misrepresent over-time correlation
  - Can take several forms
    - Posttest-only ANOVA/ANCOVA
    - ANCOVA of posttest with regression adjustment for pretest
    - Repeated measures ANOVA/ANCOVA for pretest-posttest design
  - Simulations have shown these methods have the nominal Type I error rate across a wide range of conditions common in GRTs.

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

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

Preferred Analytic Models for Parallel GRT Designs With More Than Two Time Intervals

- Random coefficients models
  - Also called growth curve models
  - The intervention effect is estimated as the difference in the condition mean trends.
  - Mixed-model ANOVA/ANCOVA assumes homogeneity of group-specific trends.
    - Simulations have shown that mixed-model ANOVA/ANCOVA has an inflated Type I error rate if those trends are heterogeneous (Murray et al., 1998).
  - Random coefficients models allow for heterogeneity of those trends.
  - Simulations have shown these methods have the nominal Type I error rate across a wide range of conditions common in GRTs.

Murray DM, Hannan PJ, et al. Analysis of data from group-randomized trials with repeat observations on the same groups. Stat Med. 1998;17(14):1581-600. PMID9699231.

Preferred Analytic Models for Individually Randomized Group Treatment Trials

- Analyses that ignore the ICC risk an inflated Type I error rate (cf. Pals et al., 2008; Baldwin et al., 2011).
  - Not as severe as in a GRT, but can exceed 15% under conditions common to these studies.
  - The solution is the same as in a GRT.
    - Analyze to reflect the variation attributable to the groups defined by the patterns of interaction.
    - Base df on the number of groups, not the number of members.
  - Mixed models are the most common approach.

Pals SL, Murray DM, et al. Individually randomized group treatment trials: a critical appraisal of frequently used design and analytic approaches. Am J Public Health. 2008;98(8):1418-24. PMID18556603.

Baldwin SA, Bauer DJ, et al. Evaluating models for partially clustered designs. Psych Methods. 2011;16(2):149-65. PMID21517179.

Individually Randomized Group Treatment Trials: Cross-Classification, Multiple Membership, or Dynamic Groups

- The GRT and IRGT literature assumes that each member belongs to one group and that group membership does not change over time.
  - These patterns often do not hold in practice and failure to model the correct structure can lead to an inflated type 1 error rate.
  - Roberts and Walwyn (2013), Luo et al. (2015), and Sterba (2017) describe cross-classified, multiple membership, and dynamic group models that address these complex design features.

Roberts C, Walwyn R. Design and analysis of non-pharmacological treatment trials with multiple therapists per patient. Stat Med. 2013;32(1):81-98. PMID22865729.

Luo W, Cappaert KJ, et al. Modelling partially cross-classified multilevel data. Br J Math Stat Psychol. 2015;68(2):342-62. PMID25773173.

Sterba SK. Partially nested designs in psychotherapy trials: A review of modeling developments. Psychother Res. 2017;27(4):425-36. PMID26686878.

Preferred Analytic Models for Stepped Wedge Group-Randomized Trials

- The original Hussey & Hughes (2007) approach assumed a common secular trend and an immediate and constant intervention effect.
- Hughes et al. (2015) allow the treatment effects to vary across groups.
- Hooper et al. (2016) allow the between-period ICC to be less than the within-period ICC, but allow no further decay.

Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials. 2007;28(2):182-91. PMID16829207.

Hughes JP, Granston TS, et al. Current issues in the design and analysis of stepped wedge trials. Contemp Clin Trials. 2015;45(Pt A):55-60. PMID26247569.

Hooper R, Teerenstra S, et al. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. Stat Med. 2016;35(26):4718-28. PMID27350420.

Preferred Analytic Models for Stepped Wedge Group-Randomized Trials

- Kasza et al. (2017) allow the between-period ICC to decay steadily.
- Grantham et al. (2019) allow more flexible decay models.
- Hughes et al. (2015) and Nickless et al. (2018) offer methods that model the intervention effect as a trend over time.

Kasza J, Hemming K et al. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. Stat Meth in Med Res. 2017;0(0)1-14. PMID29027505.

Grantham KL, Kasza J, et al. Accounting for a decaying correlation structure in cluster randomized trials with continuous recruitment. Stat Med. 2019;38(11):1918-34. PMID30663132.

Nickless A, Voysey M, et al. Mixed effects approach to the analysis of the stepped wedge cluster randomised trial-Investigating the confounding effect of time through simulation. PLoS One. 2018;13(12):e0208876. PMID30543671.

What About Randomization Tests?

- The intervention effect is a function of unadjusted or adjusted group-specific means, slopes or other group-level statistic.
- Under the null hypothesis of no intervention effect, the actual arrangement of those group-level statistics among the study conditions is but one of many equally likely arrangements.
- The randomization test systematically computes the effect for all possible arrangements.
- The probability of getting a result more extreme than that observed is the proportion of effects that are greater than that observed.
- No distributional or other assumptions are required.

What About Randomization Tests?

- Strengths
  - Gail et al. (1996) found that randomization tests had nominal Type I and II error rates across conditions common to GRTs.
    - Even when the member-level errors were non-normal,
    - Even when very few heterogeneous groups are assigned to each condition,
    - Even when the ICC was large or small,
    - So long as there was balance at the level of the group.
  - Programs for randomization tests are available in print and on the web.

Gail MH, Mark SD, Carroll RJ, Green SB, Pee D. On design considerations and randomization-based inference for community intervention trials. Stat Med. 1996;15(11):1069-92.

What About Randomization Tests?

- Weaknesses
  - The unadjusted randomization test does not offer any more protection against confounding than other unadjusted tests (Murray et al., 2006).
  - Randomization tests provide only a point estimate and a p-value.
  - Regression adjustment for covariates requires many of the same assumptions as the model-based tests.

Murray DM, Hannan PJ, Varnell SP, McCowen RG, Baker WL, Blitstein JL. 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. 2006;25(3):375-88.

What About Randomization Tests?

- Model-based methods provide parameter estimates, standard errors, and the nominal Type I error rate (Murray et al., 2006).
  - Even if the member- or group-level errors were non-normal, unless they were very skewed or heavy tailed (unpublished dissertation).
  - Even when few heterogeneous groups were assigned to each condition.
  - Even when the ICC was large or small.
  - So long as there was balance at the level of the group.
- Randomization tests and model-based tests perform similarly under most conditions.
- Randomization tests are preferred for very skewed or heavy tailed distributions.
- Ji et al. recently described randomization-based inference for SW-GRTs.

Ji XY, Fink G, Robyn PJ, Small DS. Randomization Inference for Stepped-Wedge Cluster-Randomized Trials: An Application to Community-Based Health Insurance. Annals of Applied Stat. 2017;11(1):1-20.

What About a Method Like GEE That is Robust Against Misspecification?

- Methods based on GEE use an empirical sandwich estimator for standard errors.
- That estimator is asymptotically robust against misspecification of the random-effects covariance matrix.
- When the degrees of freedom are limited (<40), the empirical sandwich estimator has a downward bias.
- Recent work provides corrections for that problem; several have recently be incorporated into SAS PROC GLIMMIX (9.1.3).
- Methods that employ the corrected empirical sandwich estimator may have broad application in GRTs, IRGTs, and SW-GRTs.

What About Fixed-Effect Methods in Two Stages?

- Introduced as the a solution for nested designs in the 1950s.
  - Commonly known as the means analysis.
  - Simple to do and easy to explain.
  - Gives results identical to the mixed-model ANOVA/ANCOVA if both are properly implemented.
  - Can be adapted to perform random coefficients analyses.
  - Can be adapted to complex designs where one-stage analyses are not possible.
  - Used in several large trials, including CATCH, MHHP, REACT, CYDS, and TAAG.

- Two-staged models can be very useful in GRTs.

What About Analysis by Subgroups?

- Some have suggested analysis by subgroup rather than group, especially when the number of groups is limited.
  - Classrooms instead of schools
  - Physicians instead of clinics
- This approach rests on the strong assumption that the subgroup captures all of the variation due to the group.
- This approach has an inflated Type I error rate even when the subgroup captures 80% of the group variation (Murray et al., 1996).
- Analysis by subgroups is not recommended.

Murray DM, Hannan PJ, Baker WL. A Monte Carlo study of alternative responses to intraclass correlation in community trials: Is it ever possible to avoid Cornfield's penalties? Eval Review. 1996;20(3):313-37.

What About Deleting the Unit of Assignment From the Model if it is not Significant?

- The df for such tests are usually limited; as such, their power is usually limited.
- Standard errors for variance components are not well estimated when the variance components are near zero.
- Even a small ICC, if ignored, can inflate the Type I error rate if the number of members per group is moderate to large.

The prudent course is to retain all random effects associated with the study design and sampling plan.

Summary of Analytic Issues

- GRTs, IRGTs, and SW-GRTs require analyses that reflect their complex designs.
- Used alone and in one stage, the usual methods based on the General or Generalized Linear Model are not valid.
- Methods based on the General Linear Mixed Model and on the Generalized Linear Mixed Model are widely applicable.
- Other methods can be used effectively, with proper care, including randomization tests, GEE, and two-stage methods.

Factors That Affect Precision in a Parallel GRT

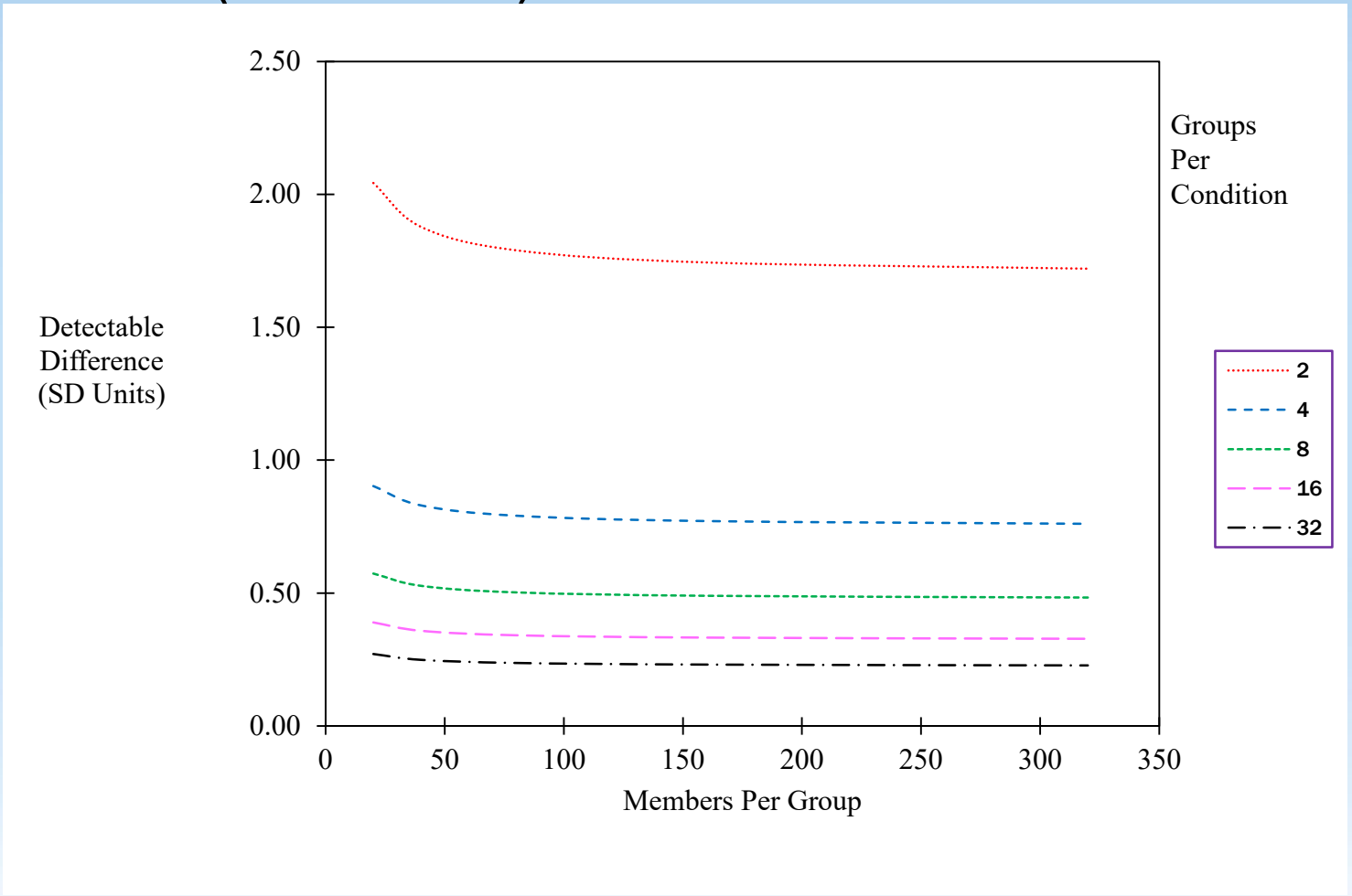
- The variance of the condition mean in a parallel GRT is:

$$\sigma^2_{\bar{y}_c} = \frac{\sigma^2_y}{mg}\left(1+(m-1)ICC\right)$$

- This equation must be adapted for more complex analyses, but the precision of the analysis will always be directly related to the components of this formula operative in the proposed analysis:
  - Replication of members and groups
  - Variation in measures
  - Intraclass correlation

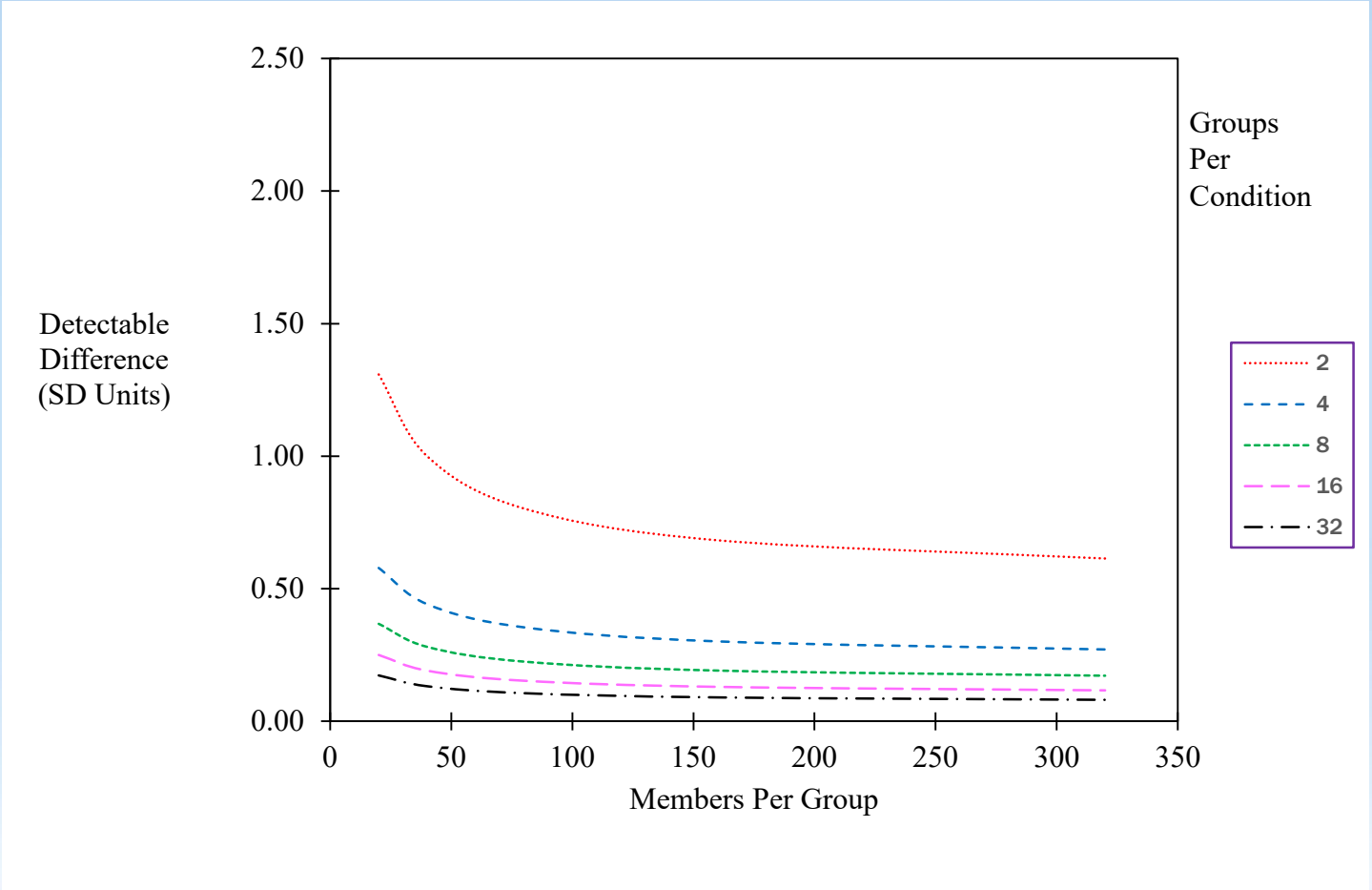
Improving Precision

- Increased replication (ICC=0.100)



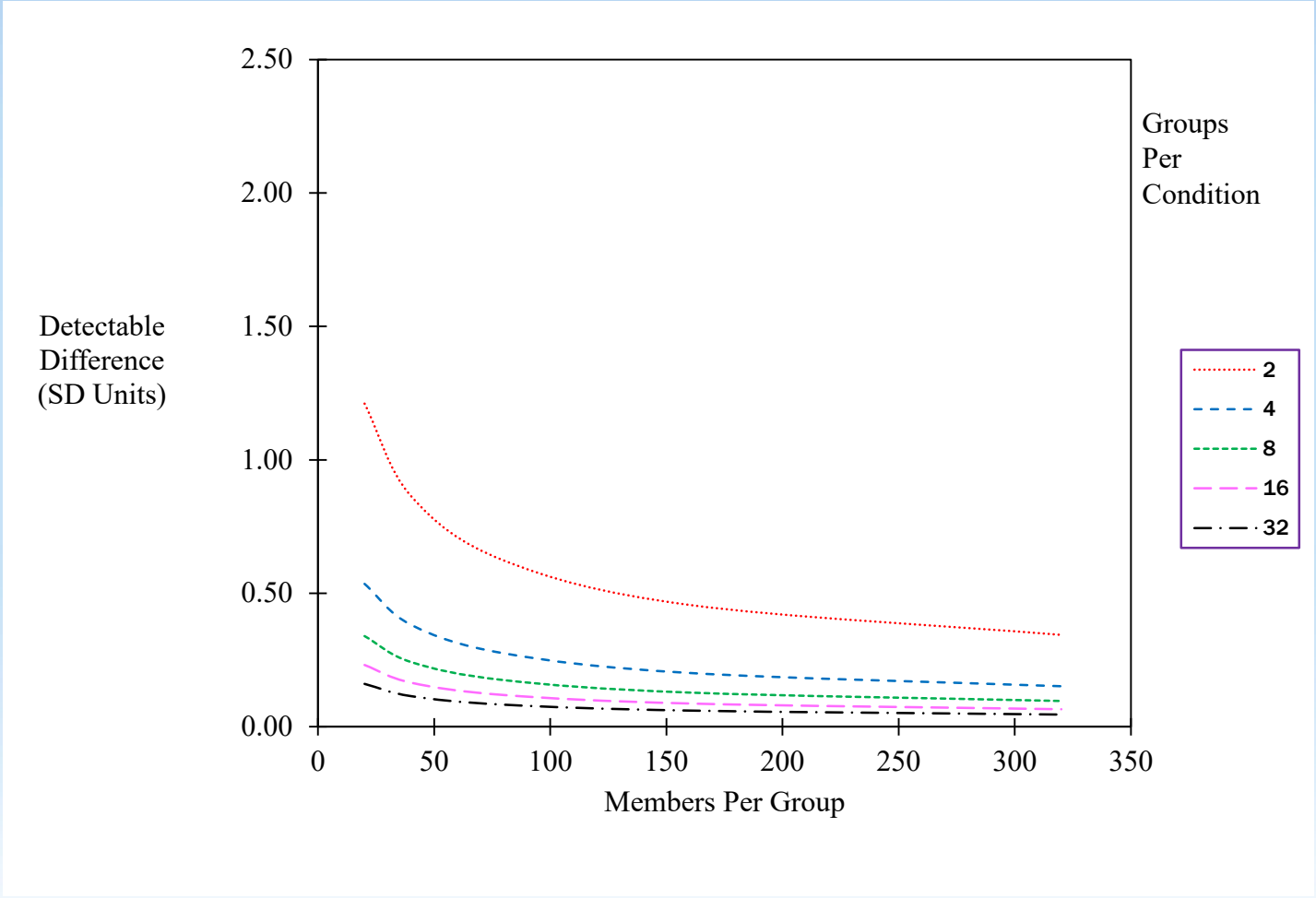
Improving Precision

- Reduced ICC (ICC=0.010)



Improving Precision

- The law of diminishing returns (ICC=0.001)



Power for Parallel GRTs

- The usual methods must be adapted to reflect the nested design
  - The variance is greater in a parallel GRT due to the expected ICC.
  - df should be based on the number of groups, not the number of members.
- Many papers now report ICCs and show how to plan a parallel GRT.
- Power in parallel GRTs is tricky, and investigators are advised to get help from someone familiar with these methods.
- A good resource is the NIH Research Methods Resources website
  - https://researchmethodsresources.nih.gov

Power for IRGTs

- Power depends heavily on the ICC, the number of groups per condition, and the number of members in the control condition for IRGTs with groups in one condition.
- Power is better in trials that do not have post-randomization interaction in the control condition.
- Methods for sample size estimation for IRGTs have been published.

Roberts C, Walwyn R. Design and analysis of non-pharmacological treatment trials with multiple therapists per patient. Stat Med. 2013;32(1):81-98. PMID22865729.

Moerbeek M, Teerenstra S. Power analysis of trials with multilevel data. Boca Raton: CRC Press; 2016.

Power for SW-GRTs

- Power depends heavily on the between- and within-period ICCs, on the number of groups, on the number of steps, and on the analytic method.
- Methods for sample size estimation for SW-GRTs have been published.

Moerbeek M, Teerenstra S. Power analysis of trials with multilevel data. Boca Raton: CRC Press; 2016.

Hemming K, Taljaard M. Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. J Clin Epi. 2016;69:137-46. PMID26344808.

Hooper R, Teerenstra S, et al. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. Stat Med. 2016;35(26):4718-28. PMID27350420.

Kasza J, Hemming K et al. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. Stat Meth in Med Res. 2017;0(0)1-14. PMID29027505.

Li F, Turner EL, et al. Sample size determination for GEE analyses of stepped wedge cluster randomized trials. Biometrics. 2018;74(4):1450-8. PMID29921006.

Unbalanced Designs

- Most of the methods for sample size estimation and data analysis assume a balanced design in terms of group size.
- As long as the ratio of the largest to the group is no worse than about 2:1, those methods are fine.
- Given more extreme imbalance reduces power and can lead to an inflated type I error rate if ignored in the analysis.

Candel MJ, Van Breukelen GJ. Varying cluster sizes in trials with clusters in one treatment arm: sample size adjustments when testing treatment effects with linear mixed models. Stat Med. 2009;28(18):2307-24.

Candel MJ, Van Breukelen GJ. Sample size adjustments for varying cluster sizes in cluster randomized trials with binary outcomes analyzed with second-order PQL mixed logistic regression. Stat Med. 2010;29(14):1488-501.

You Z, Williams OD, Aban I, Kabagambe EK, Tiwari HK, Cutter G. Relative efficiency and sample size for cluster randomized trials with variable cluster sizes. Clinical Trials. 2011;8(1):27-36.

Candel MJ, Van Breukelen GJ. Repairing the efficiency loss due to varying cluster sizes in two-level two-armed randomized trials with heterogeneous clustering. Stat Med. 2016;35(12):2000-15.

Moerbeek M, Teerenstra S. Power analysis of trials with multilevel data. Boca Raton: CRC Press; 2016.

The New Application Guide and Review Criteria

- The Application Guide instructions for the new FORMS-E include several changes to alert investigators to the methodological issues inherent in GRTs and IRGTs.
  - PHS 398 Research Plan Form
  - PHS 398 Career Development Award Supplemental Form
  - PHS Fellowship Supplemental Form
- The clinical trials specific Review Criteria include similar changes.

Application Guide FORMS-E
Applications Due on or After 01-25-18

- PHS 398 Research Plan Form
- 3. Research Strategy
- 3.3. Approach
  - For trials that randomize groups or deliver interventions to groups, describe how your methods for analysis and sample size are appropriate for your plans for participant assignment and intervention delivery. These methods can include a group- or cluster-randomized trial or an individually randomized group-treatment trial. Additional information is available at the Research Methods Resources webpage.

Application Guide FORMS-E
Applications Due on or After 01-25-18

- PHS Human Subjects and Clinical Trials Information
- 4.2. Study Design
- 4.2.a. Narrative Study Description
  - Enter a narrative description of the protocol. Studies differ considerably in the methods used to assign participants and deliver interventions. Describe your plans for assignment of participants and delivery of interventions. You will also need to show that your methods for sample size and data analysis are appropriate given those plans. For trials that randomize groups or deliver interventions to groups, special methods are required; additional information is available at the Research Methods Resources webpage.

Application Guide FORMS-E
Applications Due on or After 01-25-18

- PHS Human Subjects and Clinical Trials Information
- 4.4. Statistical Design and Power
  - You will need to show that your methods for sample size and data analysis are appropriate given your plans for assignment of participants and delivery of interventions. For trials that randomize groups or deliver interventions to groups, special methods are required; additional information is available at the Research Methods Resources webpage.

NOT-OD-17-118 New Review Criteria for Research Projects Involving Clinical Trials

- Approach
  - Study Design. Is the study design justified and appropriate to address primary and secondary outcome variable(s)/endpoints that will be clear, informative and relevant to the hypothesis being tested? Is the scientific rationale/premise of the study based on previously well-designed preclinical and/or clinical research? Given the methods used to assign participants and deliver interventions, is the study design adequately powered to answer the research question(s), test the proposed hypothesis/hypotheses, and provide interpretable results? Is the trial appropriately designed to conduct the research efficiently? Are the study populations (size, gender, age, demographic group), proposed intervention arms/dose, and duration of the trial, appropriate and well justified?

NOT-OD-17-118 New Review Criteria
for Research Projects Involving Clinical Trials

- Approach
  - Data Management and Statistical Analysis. Are planned analyses and statistical approach appropriate for the proposed study design and methods used to assign participants and deliver interventions? Are the procedures for data management and quality control of data adequate at clinical site(s) or at center laboratories, as applicable? Have the methods for standardization of procedures for data management to assess the effect of the intervention and quality control been addressed? Is there a plan to complete data analysis within the proposed period of the award?

NIH Resources

- Pragmatic and Group-Randomized Trials in Public Health and Medicine
  - https://prevention.nih.gov/grt
  - 7-part online course on GRTs and IRGTs
- Mind the Gap Webinars
  - https://prevention.nih.gov/education-training/methods-mind-gap
    - SW-GRTs for Disease Prevention Research (Monica Taljaard, July 11, 2018)
    - Design and Analysis of IRGTs in Public Health (Sherri Pals, April 24, 2017)
    - Research Methods Resources for Clinical Trials Involving Groups or Clusters (David Murray, December 13, 2017)
- Research Methods Resources Website
  - https://researchmethodsresources.nih.gov/
  - Material on GRTs and IRGTs and a sample size calculator for GRTs.

What About Alternative Designs?

- Many alternatives to GRTs have been proposed.
  - Multiple baseline designs
  - Time series designs
  - Quasi-experimental designs
  - Regression discontinuity designs
- Murray et al. (2010) compared these alternatives to GRTs for power and cost in terms of sample size and time.

Murray DM, Pennell M, Rhoda D, Hade EM, Paskett ED. Designing studies that would address the multilayered nature of health care. J Nat Cancer Institute Monographs. 2010(40):90-6. PMC3482955.

See also Shadish WR, Cook TD, Campbell DT. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston, MA: Houghton Mifflin Company; 2002.

Multiple Baseline Designs

- Intervention introduced into groups one by one on a staggered schedule
  - Measurement in all groups with each new entry.
  - Often used with just a few groups, e.g., 3-4 groups.
  - Data examined for changes associated with the intervention.

Multiple Baseline Designs

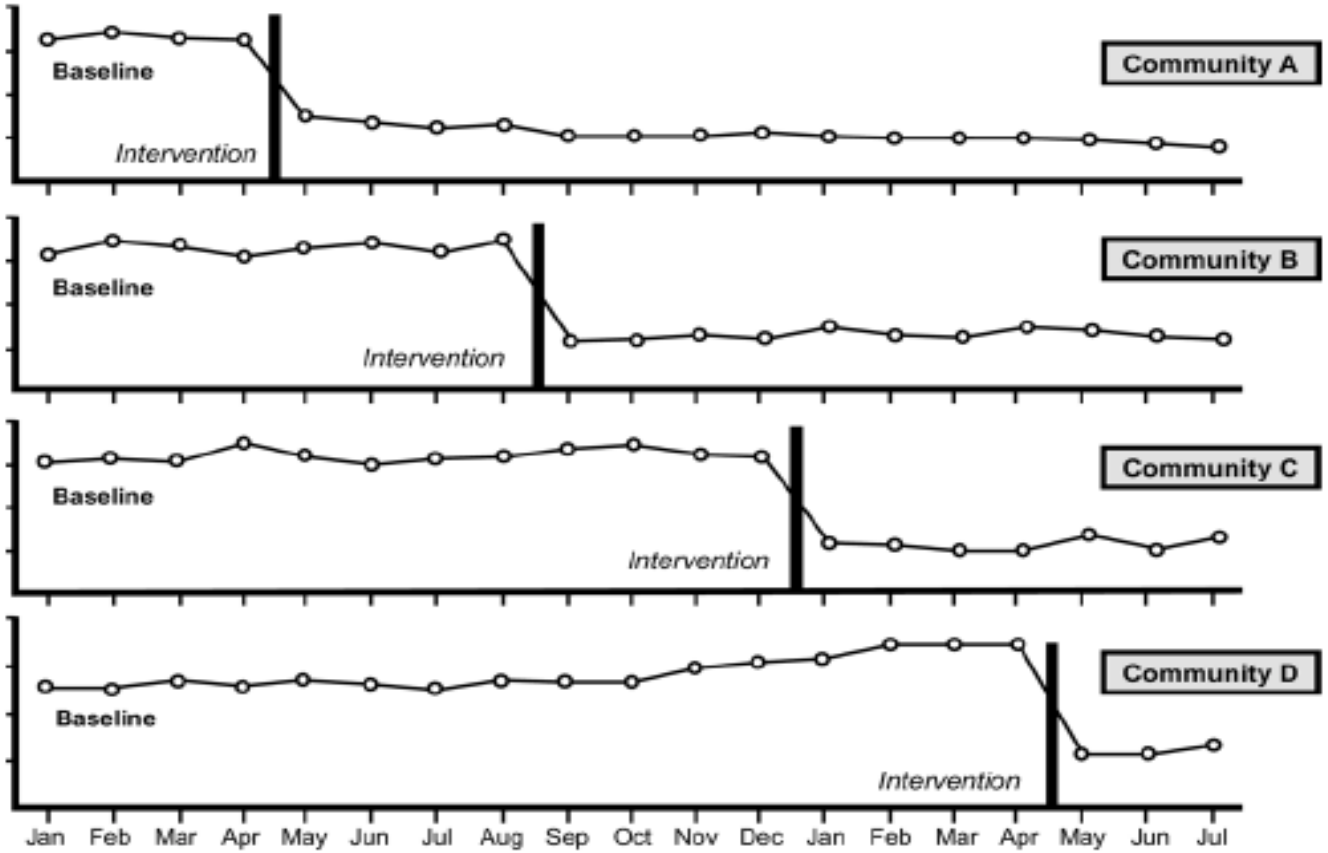


Figure 1. Hypothetical example of a multiple baseline design used to assess behavior change following an intervention in four communities.

Hawkins NG, Sanson-Fisher RW, Shakeshaft A, D'Este C, Green LW. The multiple baseline design for evaluating population-based research. Am J Prev Med. 2007;33(2):162-68.

Multiple Baseline Designs

- Evaluation relies on logic rather than statistical evidence.
  - Replication of the pattern in each group, coupled with the absence of such changes otherwise, is taken as evidence of an intervention effect.
  - With just a few groups, there is little power for a valid analysis.
- Good choice if effects are expected to be large and rapid.
- Poor choice if effects are expected to be small or gradual.
- Very poor choice if the intervention effect is expected to be inconsistent across groups.

Rhoda DA, Murray DM, Andridge RR, Pennell ML, Hade EM. Studies with staggered starts: multiple baseline designs and group-randomized trials. Am J Pub Health. 2011;101(11):2164-9. PMC3222403.

Time Series Designs

- Often used to evaluate a policy change in a single group.
- Require repeated and reliable measurements.
  - Standard methods require ~50 observations before and again after the intervention.
- Rely on a combination of logic and statistical evidence.
  - Standard methods provide evidence for change in a single group.
  - One-group designs provide no statistical evidence for between-group comparisons.
- Best used in with an archival data collection system.
  - Could be a strong approach with archival data on many groups.
- May require several cycles of data.

Quasi-Experimental Designs

- QEs have all the features of experiments except randomization.
  - Causal inference requires elimination of plausible alternatives.
- If groups are assigned and members are observed, analysis and power issues are the same as in GRTs.
- Useful when randomization is not possible.
  - Can provide experience with recruitment, measurement, intervention.
  - Can provide evidence of treatment effects if executed properly.
- Well-designed and analyzed QEs are usually more difficult and more expensive than well-designed and analyzed GRTs.

Shadish WR, Cook TD, Campbell DT. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston, MA: Houghton Mifflin Company; 2002.

Regression Discontinuity Designs

- Individuals are assigned to conditions based on a score, often reflecting the need for the intervention (Shadish et al., 2002).
- The analysis models the relationship between the assignment variable and the outcome.
  - The difference in intercepts at the cutoff is the intervention effect.
- Several recent papers have focused on regression discontinuity designs in public health and medicine (Moscoe et al., 2015; Bor et al., 2015).

Moscoe E, Bor J, Barnighausen T. Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. J Clin Epi. 2015;68(2):122-33.

Bor J, Moscoe E, Barnighausen T. Three approaches to causal inference in regression discontinuity designs. Epi. 2015;26(2):e28-30.

Bor J, Fox MP, Rosen S, Venkataramani A, Tanser F, Pillay D, Barnighausen T. Treatment eligibility and retention in clinical HIV care: A regression discontinuity study in South Africa. PLoS Med. 2017;14(11):e1002463.

Regression Discontinuity Design

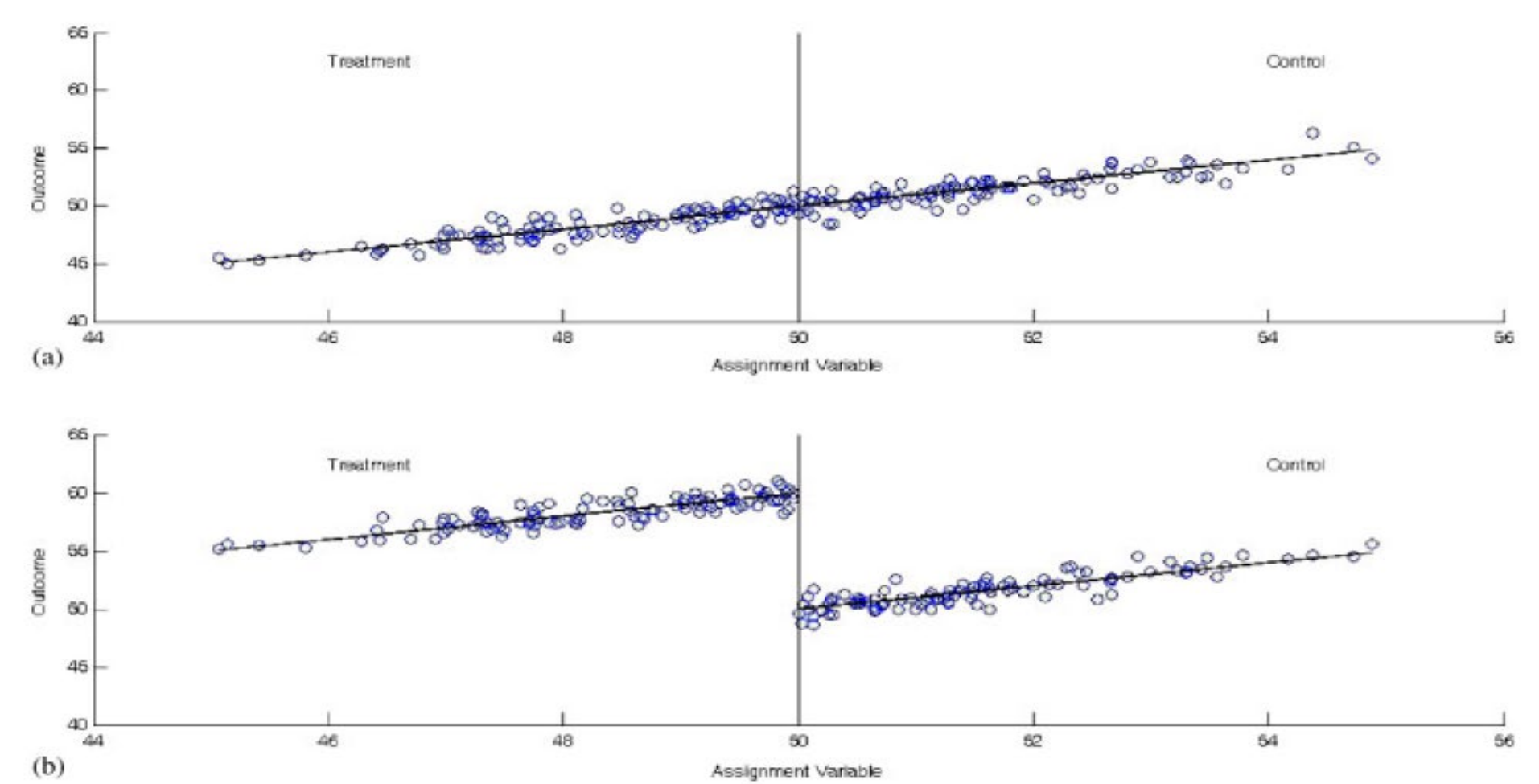


Figure 1. Hypothetical regression discontinuity experiments: (a) ineffective treatment and (b) effective treatment.

Regression Discontinuity Design

- Because assignment is fully explained by the assignment variable, proper modeling supports causal inference (Rubin, 1977).
- RDs avoid randomization, but are as valid as a RCT or GRT.
- RDs are less efficient than the standard RCT or GRT, often requiring twice as many participants.
- RDs can be used in the context of GRTs (Pennell, et al., 2011).

Pennell ML, Hade EM, Murray DM, Rhoda DA. Cutoff designs for community-based intervention studies. Stat Med. 2011;30(15):1865-82. PMC3127461.

Rubin DB. Assignment to treatment group on the basis of a covariate. J Ed Beh Stat. 1977;2(1):1-26.

Summary

- A parallel GRT remains the best comparative design available whenever the investigator wants to evaluate an intervention that…
  - operates at a group level
  - manipulates the social or physical environment
  - cannot be delivered to individuals
- Parallel GRTs provide better or equal quality evidence and are either more efficient or take less time than the alternatives.
- Even so, GRTs are more challenging than the usual RCT.
  - IRGTs present many of the same issues found in GRTs.
  - Investigators new to GRTs and IRGTs should collaborate with more experienced colleagues, especially experienced biostatisticians.

Summary

- Many alternatives to parallel GRTs have been proposed.
  - Stepped wedge designs
  - Multiple baseline designs
  - Time series designs
  - Quasi-experimental designs
  - Regression discontinuity designs
- These alternatives can provide evidence for causal inference.
  - Some rely on logic more than statistical evidence.
    - Multiple baseline designs, time-series designs
  - Others require studies as large or larger than GRTs
    - Quasi-experimental designs, regression discontinuity
  - Stepped wedge can be more efficient, but takes longer and faces more threats to internal validity