ePCP Design and Analysis: Trials involving Group Randomization or Delivery of Interventions to Groups

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

NIH ePCP Training Workshop
March 5, 2020

National Institutes of Health
Office of Disease Prevention

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 intervention-induced correlation among observations taken from participants who receive part of their intervention in the same group or through a shared interventionist (live or virtual)
    - Many surgical and 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.

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 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)

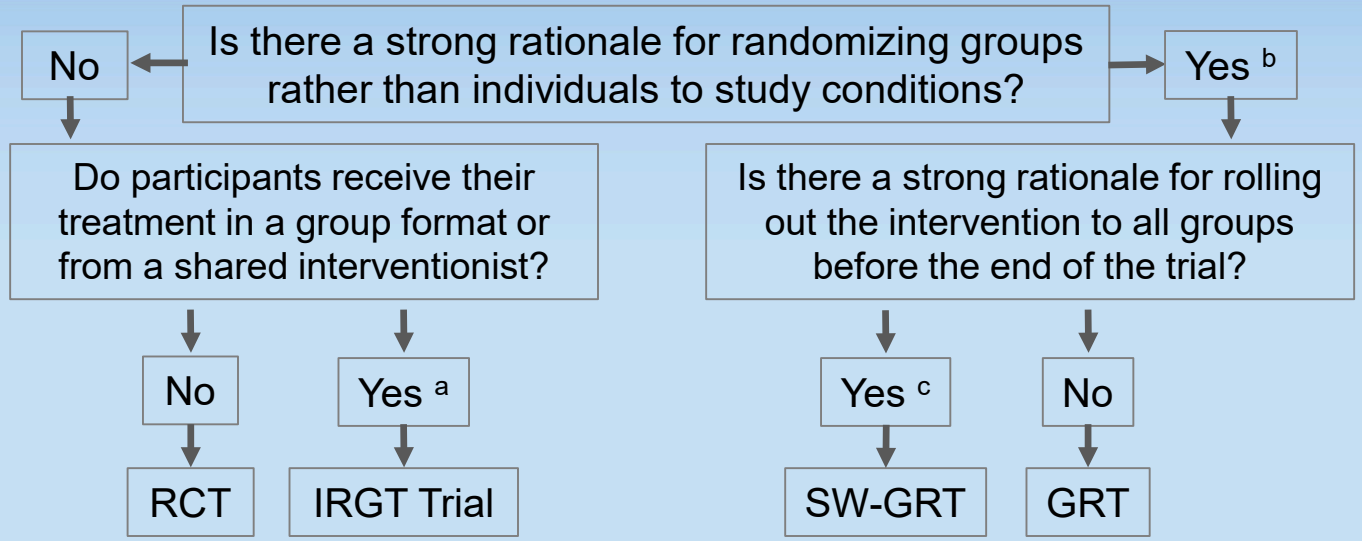
- Though Cornfield's remarks were addressed only to GRTs, they also apply to IRGTs, and to SW-GRTs

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.

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.

Single Factor and Factorial GRTs

- 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 in GRTs

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

Cross-Sectional and Cohort Designs in GRTs

- 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 in GRTs

- 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 in GRTs

- 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.

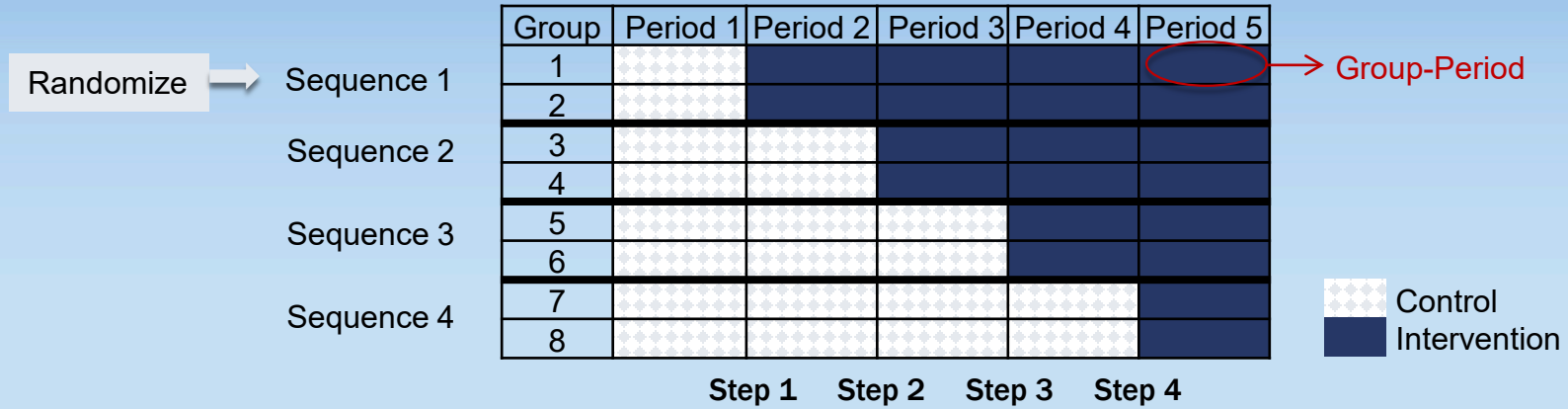
Li F, Lokhnygina Y, et al. An evaluation of constrained randomization for the design and analysis of group-randomized trials. Stat Med. 2016;35(10):1565-79. PMID26598212.

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.

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.

The Basic Stepped Wedge-GRT Design



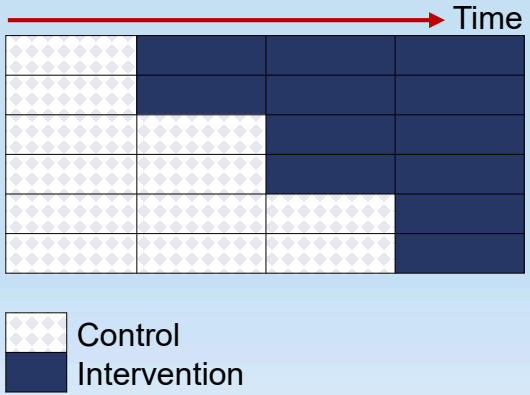
- Groups are randomized to sequences.
  - This is where matching, stratification, or constrained stratification would be used to improve comparability of the sequences.
- Groups cross to intervention sequentially and in random order, either individually or in sets.
- Outcomes are assessed repeatedly 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

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.

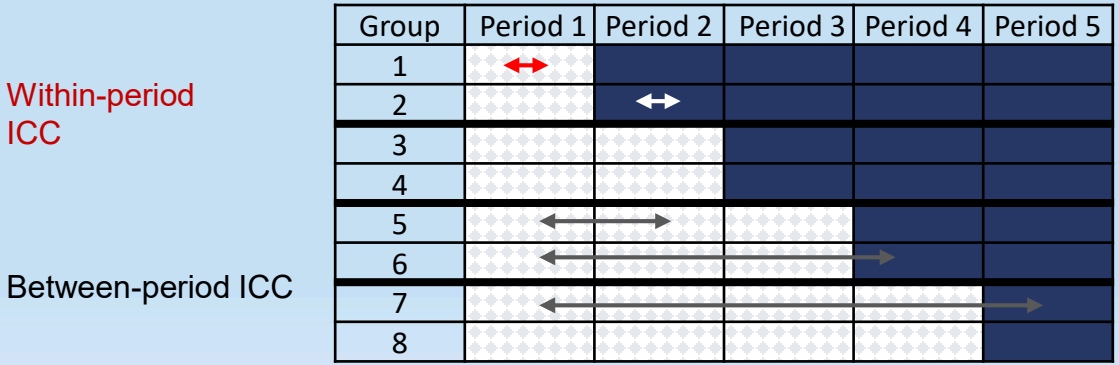
Effect Heterogeneity

- 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.



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

Strategies to Protect Statistical Validity

- 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 Statistical Validity

- 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.

Factors That Affect Precision in a Parallel GRT

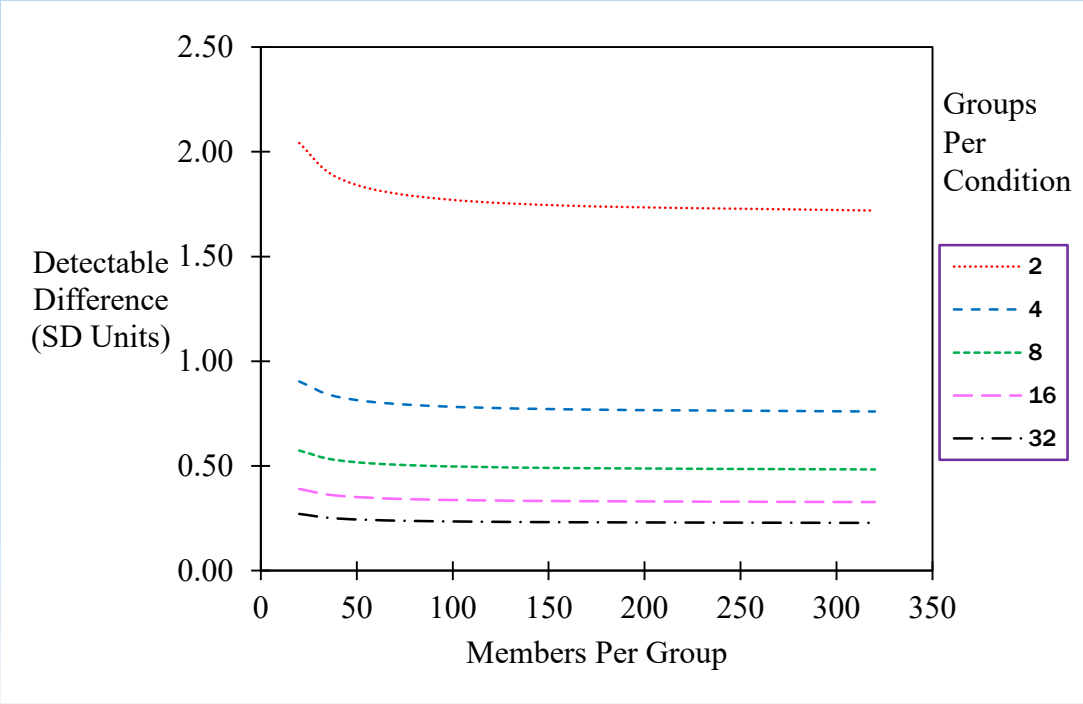
- The variance of the condition mean in a parallel GRT is:

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

- 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

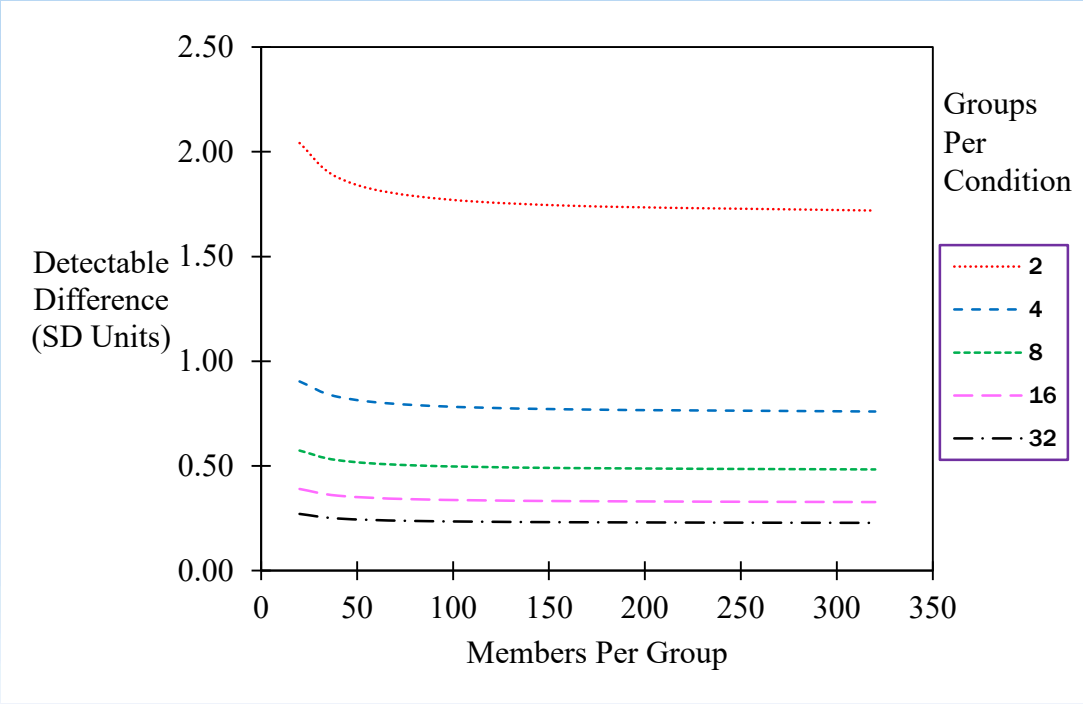
Improving Precision

- Increased replication (ICC=0.100)



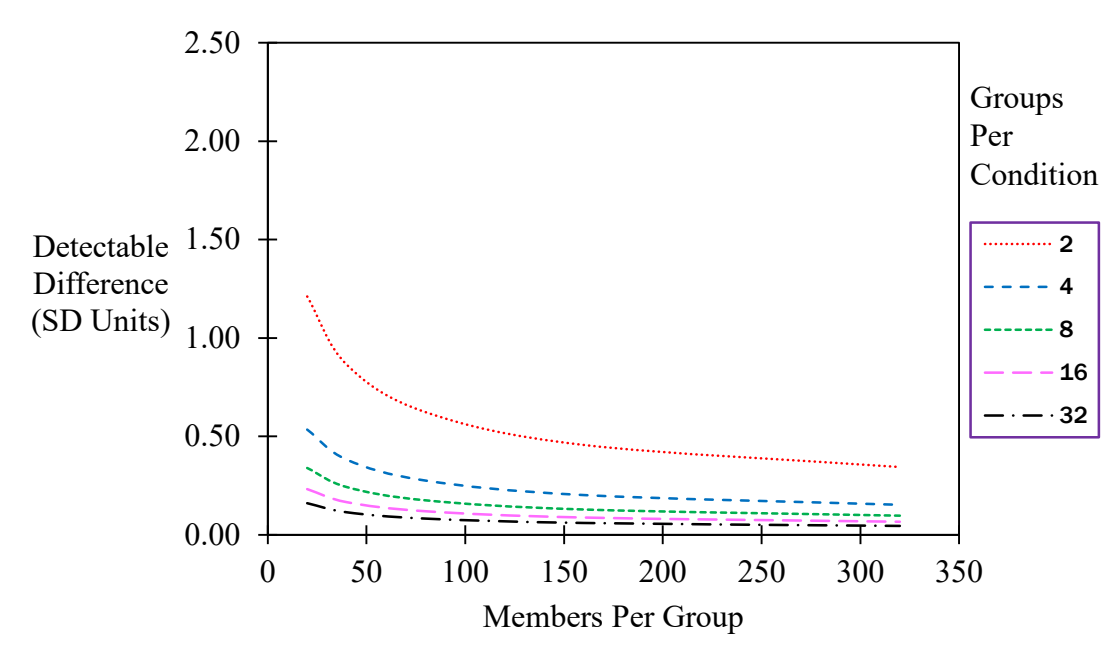
Improving Precision

- 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.

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

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.

Hemming K, Kasza J, et al. A tutorial on sample size calculation for multiple-period cluster randomised parallel, cross-over and stepped-wedge trials using the Shiny CRT Calculator. International Journal of Epidemiology. in press.

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.

Hemming K, Kasza J, et al. A tutorial on sample size calculation for multiple-period cluster randomised parallel, cross-over and stepped-wedge trials using the Shiny CRT Calculator. International Journal of Epidemiology. in press.

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.

Hemming K, Kasza J, et al. A tutorial on sample size calculation for multiple-period cluster randomised parallel, cross-over and stepped-wedge trials using the Shiny CRT Calculator. International Journal of Epidemiology. in press.

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.