



Health Care Systems Research Collaboratory

*Rethinking Clinical Trials®*

# Design and Analysis of Embedded Pragmatic Trials

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

***Essentials of Embedded Pragmatic Clinical Trials Seminar***



## Learning goals

- ePCT Design: Learn about group- or cluster-randomized trials, individually randomized group-treatment trials, and stepped wedge group- or cluster-randomized trials
- ePCT Analysis: Learn about the special analytic requirements for these designs and about the current recommendations for their analysis



## Important things to know

- Studies that randomize groups or deliver interventions to groups face special analytic challenges not found in traditional randomized controlled trials
- Failure to address these challenges will result in an underpowered study and/or an inflated type 1 error rate
- We won't advance the science by using inappropriate methods

# Methods for pragmatic trials

- Pragmatic trials do not require a completely different set of research designs, measures, analytic methods, etc.
- As always, the choice of methods depends on the research question.
- The research question dictates
  - the intervention, target population, and variables of interest,
  - which dictate the setting, research design, measures, and analytic methods.
- Randomized trials will provide the strongest evidence.
  - What kind of randomized trial depends on the research question and how the intervention will be delivered.
- Alternatives to randomized trials are available, but not included in this presentation.

# Three kinds of randomized trials

- Randomized Controlled Trials (RCTs)
  - Individuals randomized to study conditions, no interaction among participants after randomization
    - Most drug trials
- Individually Randomized Group Treatment Trials (IRGTs)
  - Individuals randomized to study conditions, interaction among participants post randomization in at least one condition
    - Many surgical trials
    - Many behavioral trials
- Group-Randomized Trials (GRTs)
  - Groups randomized to study conditions, interaction among members of the same group before and after randomization
    - Many trials conducted in communities, worksites, schools, 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 trial.

# 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 controlled 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.



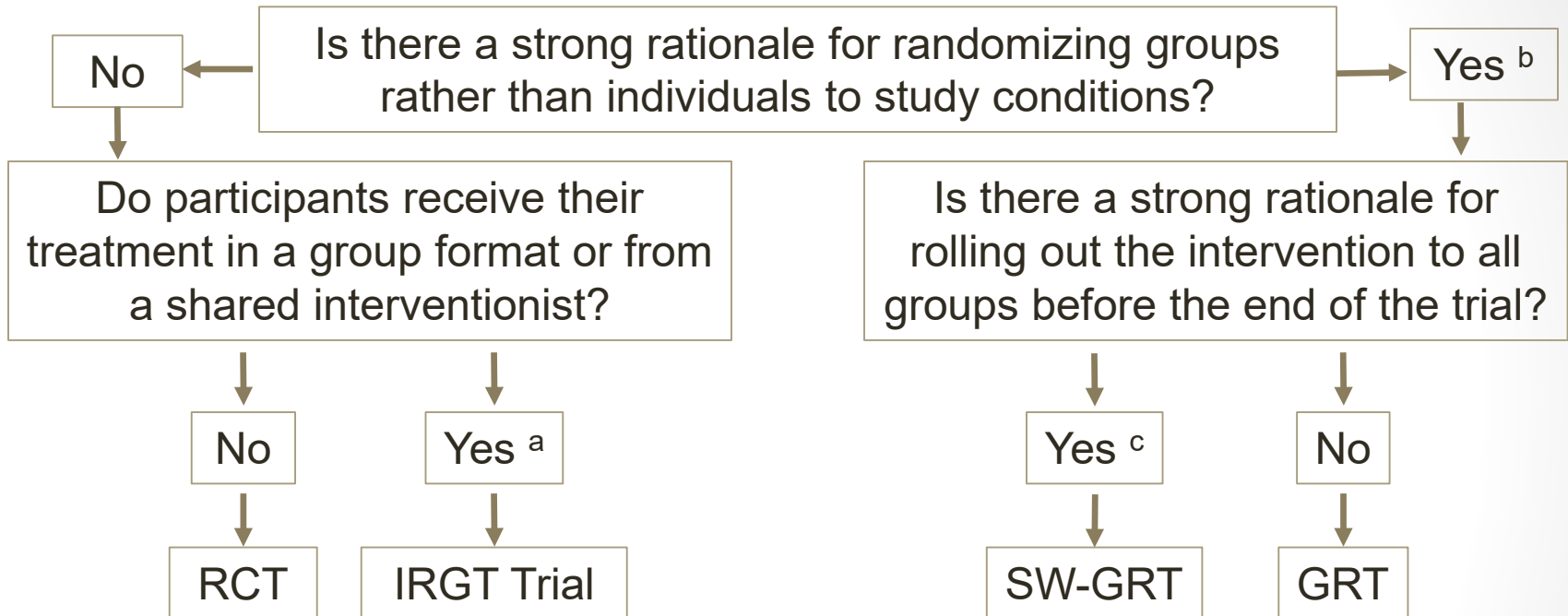
# Impact on the design

- 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 trial progresses.
  - SW-GRTs are inherently less rigorous than parallel GRTs and should be considered only when a parallel GRT is not appropriate.

# The need for these designs

- An RCT is the best comparative design whenever...
  - Individual randomization is possible without post-randomization interaction.
- An IRGT is the best comparative design whenever...
  - Individual randomization is possible but there are reasons to allow post-randomization interaction.
- A GRT is the best comparative design whenever the investigator wants to evaluate an intervention that...
  - Cannot be delivered to individuals without risk of contamination.
- An SW-GRT is an alternative to a parallel GRT if...
  - It is unethical to withhold the intervention from any groups.
  - It is impossible to implement the intervention in many groups simultaneously.
  - External events are unlikely to affect the outcomes.

# Choosing among these designs



<sup>a</sup> 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.

<sup>b</sup> 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.

<sup>c</sup> There may be legitimate political or logistical reasons to roll out the intervention to all groups before the end of the trial.

# 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

# 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.
  - 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

# 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*

in and out migration

group change

recruitment costs

less powerful?

full dose?

*Cohort*

mortality

individual change

tracking and follow-up costs

more powerful?

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. *Statistics in Medicine*. 2007;26(9):2036-51.

# Constrained randomization

- Stratification or 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) evaluated constrained randomization for power and type 1 error.

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

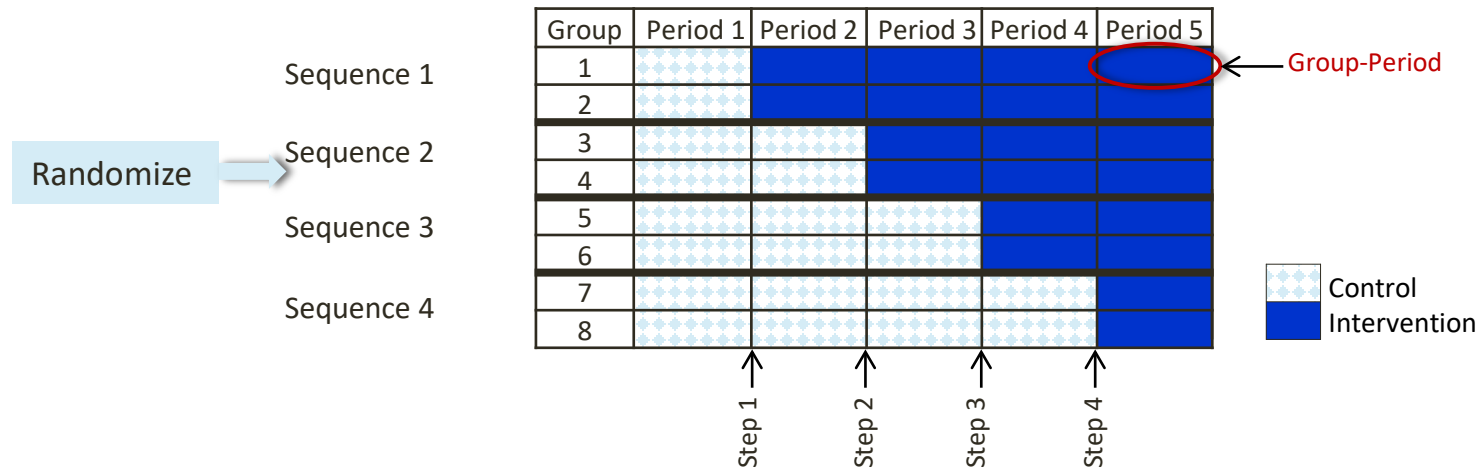
Li F, Lokhnygina Y, et al. An evaluation of constrained randomization for the design and analysis of group-randomized trials. *Stat Med*. 2016;35(10):1565-79. PMID26598212.

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 type of stepped wedge designs
- Key methodological considerations
  - Confounding by time
  - Contamination
  - Time-varying intervention effects
  - Effect heterogeneity
  - Complex correlations

# Stepped wedge group-randomized trial



- 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 SW-GRT 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

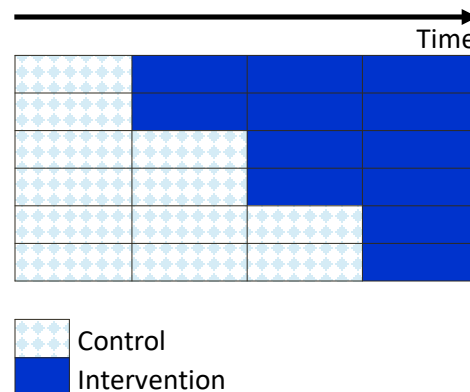
# Key methodological considerations

- 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.

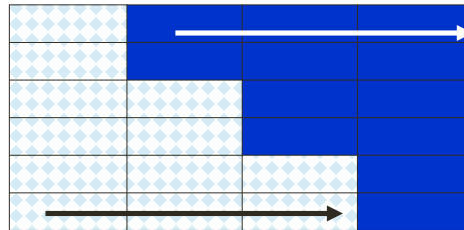
# Contamination

- 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.



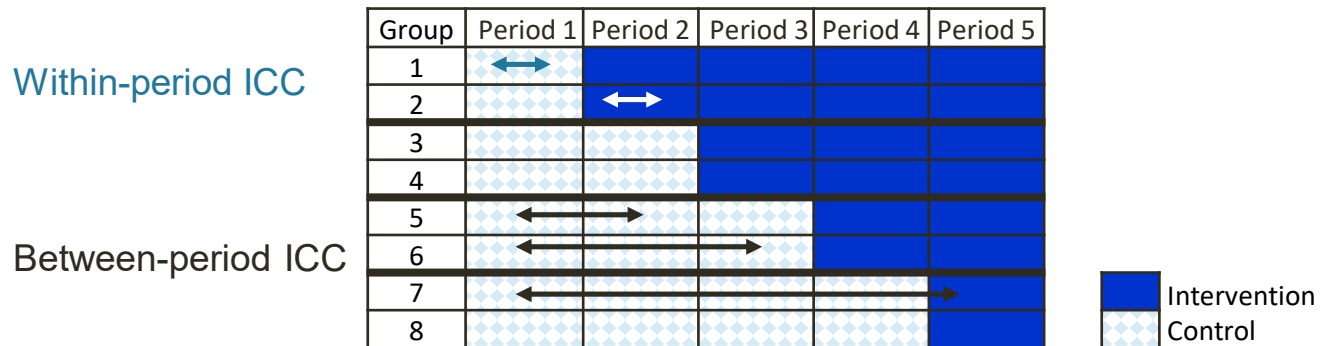
# Effect heterogeneity

- 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. *Contemporary Clinical 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 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.



# Summary of design issues

- All the design features common to RCTs are available to GRTs and IRGTs, with the added complication of an extra level of nesting:
  - Nested cohort and nested cross-sectional designs;
  - Post only, pre-post, and extended designs;
  - Single factor designs and factorial designs;
  - A priori matching or stratification;
  - Constrained randomization
- The primary threats to internal and statistical validity are well known, and defenses are available.
  - Plan the study to reflect the nested design, with sufficient power for a valid analysis, and avoid threats to internal validity.

# Summary of design issues

- Many of the design features common to RCTs are available to SW-GRTs:
  - Cohort and cross-sectional designs;
  - Single factor designs and factorial designs;
  - A priori matching, stratification, or constrained randomization to create comparable sequences.
- The primary threats to internal and statistical validity are well known, and defenses are available.
  - Plan the study to reflect the nested design, with sufficient power for a valid analysis, and avoid threats to internal validity.

# Impact on the analysis

- 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 or SW-GRT will share some physical, geographic, social or other connection.
- The participants who interact in an IRGT will develop similar connections.
- Those connections will create a positive intraclass correlation that reflects extra variation attributable to the group.

$$ICC_{m:g:c} = \text{corr}(y_{i:kl}, y_{i':kl})$$

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

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

# Impact on the analysis

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

$$\sigma_{g:c}^2 = \sigma_y^2 \left( \text{ICC}_{m:g:c} \right)$$

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

$$\sigma_y^2 = \sigma_e^2 + \sigma_{g:c}^2$$

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

$$\text{ICC}_{m:g:c} = \frac{\sigma_{g:c}^2}{\sigma_e^2 + \sigma_{g:c}^2}$$

# Impact on the analysis: GRT, IRGT

- Given m members in each of g groups...

- When group membership is established by random assignment,

$$\sigma_{\bar{y}_g}^2 = \frac{\sigma_y^2}{m}$$

- When group membership is not established by random assignment,

$$\sigma_{\bar{y}_g}^2 = \frac{\sigma_e^2}{m} + \sigma_g^2$$

- Or equivalently,

$$\sigma_{\bar{y}_g}^2 = \frac{\sigma_y^2}{m} (1 + (m-1) ICC)$$

# Impact on the analysis: GRT, 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. *Journal of the American Statistical Association*. 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 <sub>y</sub> =ICC <sub>x</sub> |      |       | ICC <sub>y</sub> =1 |      |      | ICC <sub>y</sub> =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:

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

# Impact on analysis: SW-GRTs

- Groups are crossed rather than nested within conditions.
- Crossing of groups with study conditions reduces the impact of the ICC compared to a parallel GRT or IRGT, either improving power or allowing a smaller study.
- Even so, there are other potential sources of bias:
  - The intervention is confounded with time.
    - SW-GRTs face a greater risk from external events that can affect the outcomes.
  - The intervention effect may vary over time.
- Any analysis that assumes that there is no secular trend, that the intervention effect is constant over time, and that the intervention is the only event that could affect the outcome may be biased.

# 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):100-102.

# The need for these designs

- 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.

# 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.

# Factors that affect precision in a GRT

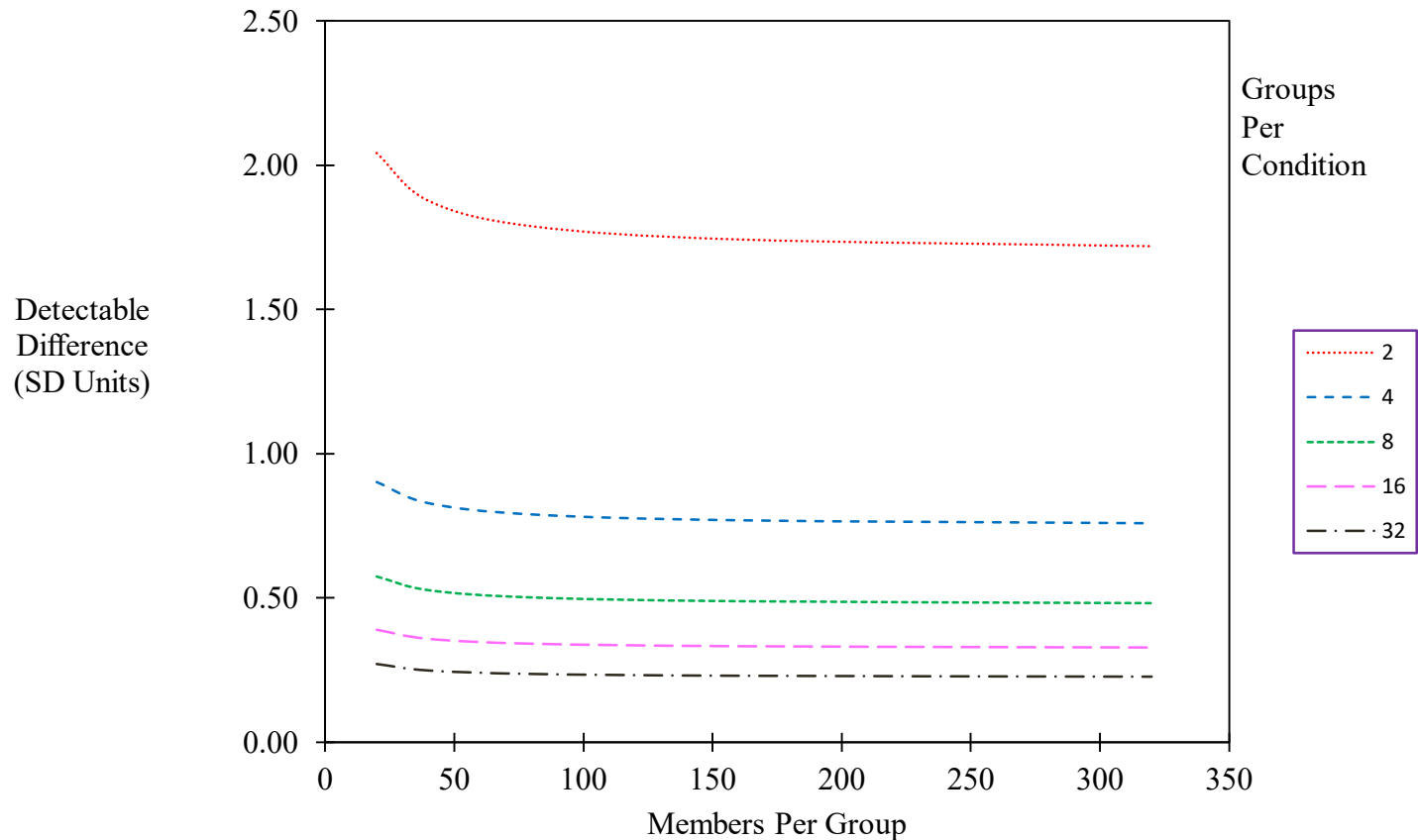
- The variance of the condition mean in a GRT is:

$$\sigma_{\bar{y}_c}^2 = \frac{\sigma_y^2}{mg} (1 + (m-1)ICC)$$

- 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 in a GRT

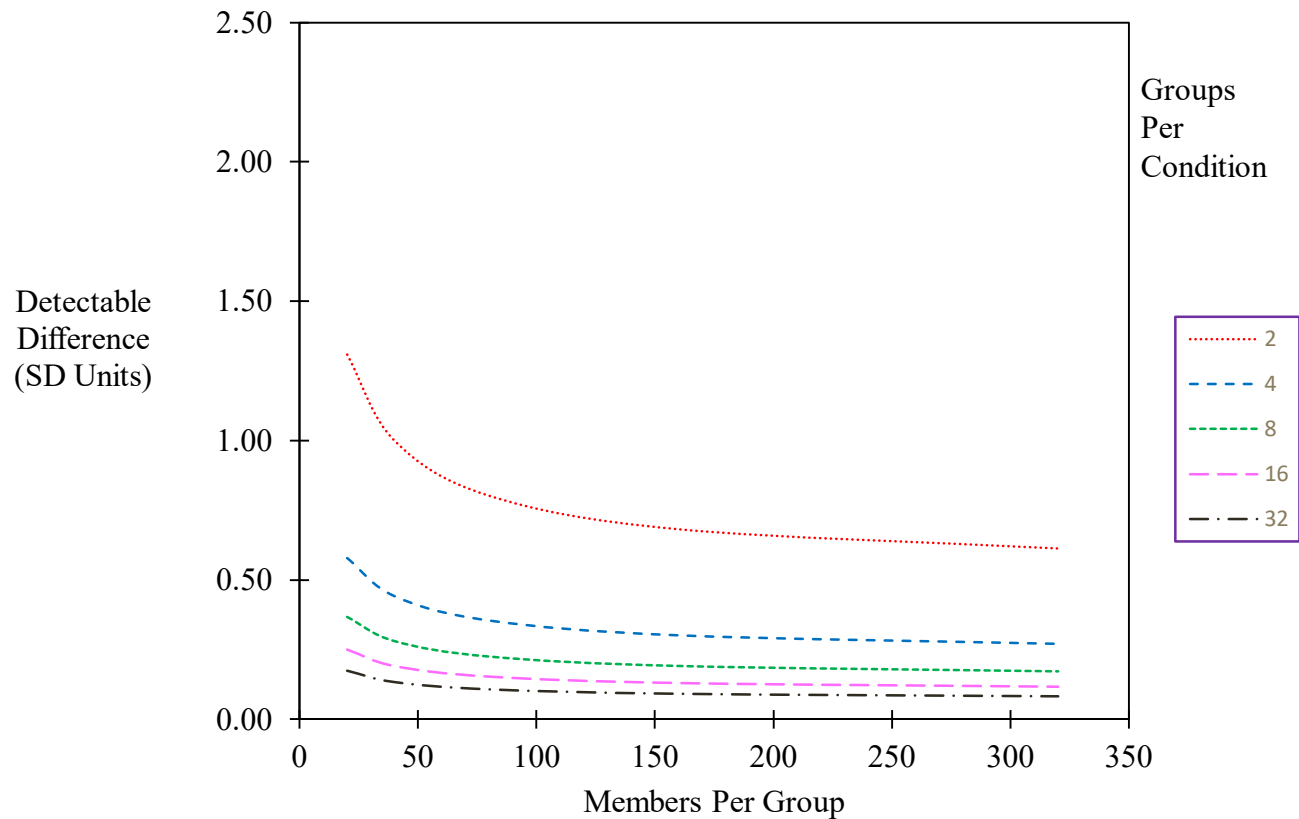
- Increased replication (ICC=0.100)





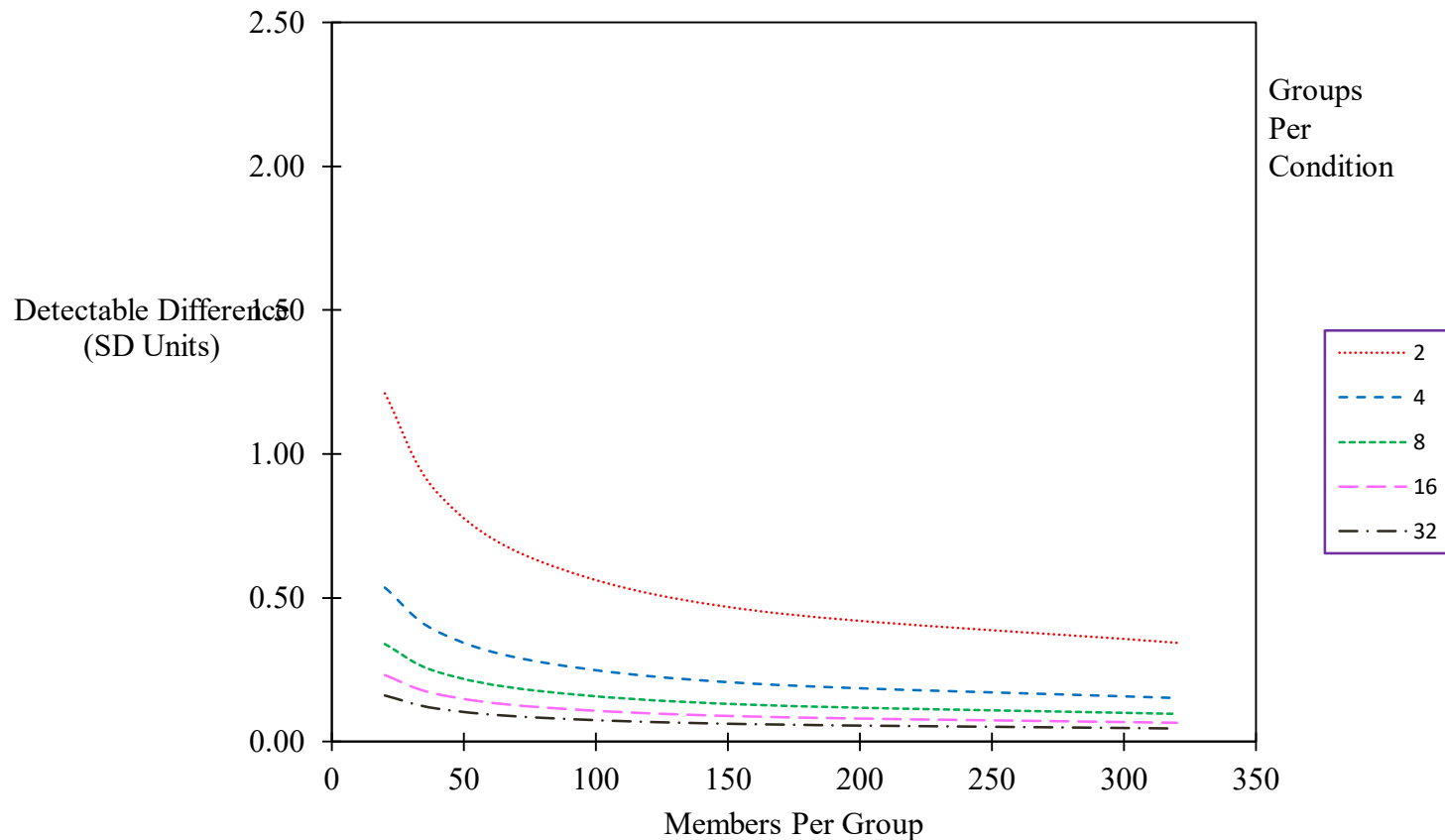
# Improving precision in a GRT

- Reduced ICC (ICC=0.010)



# Improving precision in a GRT

- The law of diminishing returns (ICC=0.001)



# Preferred analytic models for GRTs: 1 or 2 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.

# Preferred analytic models for GRTs: 3 or more 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.

# 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.

# 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 multilevel, 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 in 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.

# 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 Clinical 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 Clinical 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.

# 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.



# 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.

# Power for group-randomized trials

- The usual methods must be adapted to reflect the nested design
  - The variance is greater in a GRT due to the expected ICC.
  - df should be based on the number of groups, not the number of members.
- Power depends heavily on the ICC and the number of groups per condition, less on the number of members per group.
- Many papers now report ICCs and show how to plan a GRT.
- Power in 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 and the number of groups per 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 in 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 in 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. *Statistics in Medicine*. 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. *Statistics in Medicine*. 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. *Statistics in Medicine*. 2016;35(12):2000-15.

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

# Analysis and power for SW-GRTs

- Most of the methods and software assumes that the intervention effect develops fully during the step in which it is introduced and persists at a steady state thereafter.
- That pattern may not hold in practice.
- Use of sample size estimation methods that wrongly assume this pattern may greatly overestimate power.
- Use of data analytic methods that wrongly assume this pattern may yield a substantially diluted intervention effect estimate.
- Much work is needed to flesh out methods for sample size estimation and data analysis for studies in which the intervention effect is expected to develop more gradually or to fade over time.

# 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



# Important things to do

- Read - Turner EL, Li F, Gallis JA, Prague M, Murray DM. 2017. Review of Recent Methodological Developments in Group-Randomized Trials: Part 1-Design. Am J Public Health 107: 907-15
- Read - Turner EL, Prague M, Gallis JA, Li F, Murray DM. 2017. Review of Recent Methodological Developments in Group-Randomized Trials: Part 2-Analysis. Am J Public Health 107: 1078-86
- Read - Hemming K, Taljaard M, McKenzie JE, Hooper R, Copas A, et al. 2018. Reporting of stepped wedge cluster randomised trials: extension of the CONSORT 2010 statement with explanation and elaboration. BMJ 363: k1614
- Read - Murray DM, Pals SL, George SM, Kuzmichev A, Lai GY, et al. 2018. Design and analysis of group-randomized trials in cancer: A review of current practices. Prev Med 111: 241-47