Pragmatic and Group-Randomized Trials in Public Health and Medicine
Part 3: Analysis Approaches

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

A free, 7-part, self-paced, online course from NIH
with instructional slide sets, readings, and guided activities

Target Audience

- Faculty, post-doctoral fellows, and graduate students interested in learning more about the design and analysis of group-randomized trials.
- Program directors, program officers, and scientific review officers at the NIH interested in learning more about the design and analysis of group-randomized trials.
- Participants should be familiar with the design and analysis of individually randomized trials (RCTs).
  - Participants should be familiar with the concepts of internal and statistical validity, their threats, and their defenses.
  - Participants should be familiar with linear regression, analysis of variance and covariance, and logistic regression.

Learning Objectives

- And the end of the course, participants will be able to…
  - Discuss the distinguishing features of group-randomized trials (GRTs), individually randomized group-treatment trials (IRGTs), and individually randomized trials (RCTs).
  - Discuss their appropriate uses in public health and medicine.
  - For GRTs and IRGTs…
    - Discuss the major threats to internal validity and their defenses.
    - Discuss the major threats to statistical validity and their defenses.
    - Discuss the strengths and weaknesses of design alternatives.
    - Discuss the strengths and weaknesses of analytic alternatives.
    - Perform sample size calculations for a simple GRT.
  - Discuss the advantages and disadvantages of alternatives to GRTs for the evaluation of multi-level interventions.

Organization of the Course

- Part 1: Introduction and Overview
- Part 2: Designing the Trial
- **Part 3: Analysis Approaches**
- Part 4: Power and Sample Size
- Part 5: Examples
- Part 6: Review of Recent Practices
- Part 7: Alternative Designs and References

A Classification Scheme for Statistical Models

	Gaussian Distribution	Non-Gaussian Distribution
One Random Effect	General Linear Model	Generalized Linear Model
Two Or More Random Effects	General Linear Mixed Model	Generalized Linear Mixed Model

- Fixed effect: the investigators want to draw inferences only about the levels used in the study.
- Random effect: the investigators want to draw inferences about some larger population of levels that are only represented by the levels used in the study.

Preferred Models for 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.

Preferred Models for 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.
  - 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.

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. Statistics in Medicine. 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. Statistics in Medicine. 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.

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 (beginning with SAS 9.1.3).

- Methods that employ the corrected empirical sandwich estimator may have broad application in GRTs.

What About Methods Developed for Analysis of Complex Survey Samples?

- Methods developed for analysis of complex survey samples perform well given a large number of primary sampling units.
  - These methods do not perform well when the number of primary sampling units is limited (<40).
  - The standard normal approximation that often accompanies these methods is not appropriate given limited df.
  - Those methods for analysis of complex survey samples may have limited application in GRTs.

- Many survey analysis programs have adopted empirical sandwich estimation, and if one of the small-sample correction factors is employed, such methods would be applicable to 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? Evaluation 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.

What About Studies Based on Only One Group per Condition?

- Cannot separately estimate variation due to the group and variation due to condition.
- Must rely on a strong assumption:
  - Post hoc correction: external estimate is valid
  - Subgroup or batch analysis: subgroup captures group variance
  - Fixed-effects analysis: group variance is zero
- Varnell et al. (2001) found the second and third strategies are likely to have an inflated Type I error rate.
- This design should be avoided if statistical evidence is important for causal inference.
- It may still be helpful for preliminary studies.

- Varnell SP, Murray DM, Baker WL. An evaluation of analysis options for the one group per condition design: can any of the alternatives overcome the problems inherent in this design? Evaluation Review. 2001;25(4):440-53.

Will Kish's Effective df Help?

- Some have suggested evaluating the intervention effect against effective df = (individual df) / DEFF.
  - This approach was tested in simulations, varying the magnitude of the ICC and the number of groups per condition.
  - Effective df performed no better than df based on the members -- the Type I error rate was still inflated, often badly (Murray et al., 1996).
- Kish's effective df is not likely to have broad application in GRTs.

- 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? Evaluation Review. 1996;20(3):313-37.

What About Unbalanced Designs?

- Group-level imbalance can create analytic problems (Gail et al., 1996; Murray et al., 2006).
- Member-level imbalance can create Type I error inflation and the risk increases with the level of imbalance.
- Johnson et al. (2015) compared 10 model-based approaches to member imbalance.
  - A one-stage mixed model with Kenward-Roger df and unconstrained variance components performed well for g≥14.
  - A two-stage model weighted by the inverse of the estimated theoretical variance of the group means and with unconstrained variance components performed well for g≥6.

- Johnson JL, Kreidler SM, Catellier DJ, Murray DM, Muller KE, Glueck DH. Recommendations for choosing an analysis method that controls Type I error for unbalanced cluster sample designs with Gaussian outcomes. Statistics in Medicine. 2015;34(27):3531-45.

What About Constrained Randomization?

- Li et al. (2015) evaluated model-based and randomization tests in the context of constrained randomization in a GRT.
  - The unadjusted randomization test maintained the nominal Type I error rate; the unadjusted model-based test was conservative.
  - Adjusted model-based and randomization tests were similar.
    - Both maintained the nominal Type I error rate.
    - Both had better power under constrained randomization.
  - Correct specification of the permutation distribution is essential under constrained randomization.
- Constrained randomization can improve power if used well.

- Li F, Lokhnygina Y, Murray DM, Heagerty PJ, DeLong ER. An evaluation of constrained randomization for the design and analysis of group-randomized trials. Statistics in Medicine. 2015;35(10):1565-79. PMC4826850.

Is the Non-Negativity Constraint OK?

- Software based on maximum likelihood routinely constrains variance estimates to be non-negative.
  - Combined with traditional methods for calculating df, this constraint introduces a positive bias in the variance component estimates and depresses the Type I error rate, often dramatically (Swallow & Monahan, 1984; Murray et al., 1996).
  - Earlier advice was to avoid the non-negativity constraint.
- Recent evidence suggests that the Kenward-Roger method for df addresses this problem (Andridge et al., 2014).

- Swallow WH, Monahan JF. Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components. Technometrics. 1984;26(1):47-57.
- Andridge RR, Shoben AB, Muller KE, Murray DM. Analytic methods for individually randomized group treatment trials and group-randomized trials when subjects belong to multiple groups. Statistics in Medicine. 2014;33(13):2178-90. PMC4013262.

What About Individually Randomized Group Treatment Trials (IRGTs)?

- Many studies randomize participants as individuals but deliver treatments in small groups (cf. Pals et al., 2008).
  - Psychotherapy, weight loss, smoking cessation, etc.
  - Participants nested within groups, facilitators nested within conditions
  - Little or no group-level ICC at baseline.
  - Positive ICC later, with the magnitude proportional to the intensity and duration of the interaction among the group members.

- Pals SP, Murray DM, Alfano CM, Shadish WR, Hannan PJ, Baker WL. Individually randomized group treatment trials: a critical appraisal of frequently used design and analytic approaches. American Journal of Public Health. 2008;98(8):1418-24. PMC2446464
- Pals SL, Murray DM, Alfano CM, Shadish WR, Hannan PJ, Baker WL. Erratum. American Journal of Public Health. 2008;98(12):2120.

What About Individually Randomized Group Treatment Trials (IRGTs)?

- 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 small groups.
    - Base df on the number of small groups, not the number of members.

- Baldwin SA, Bauer DJ, Stice E, Rohde P. Evaluating models for partially clustered designs. Psychological Methods. 2011;16(2):149-65. PMC3987820.

What About IRGTs In Which Members Belong to More than one Group or Change Groups?

- The IRGT literature assumes that each member belongs to a single group and that group membership does not change.
  - That pattern is not likely to hold in practice.
  - Andridge (2014) found that failure to account for multiple group membership can inflate Type I error for the methods described thus far.
  - Roberts (2013) found that multiple membership multilevel models address this problem.
    - They require data on membership time in each group, which is not routinely collected in IRGTs.

- Andridge RR, Shoben AB, Muller KE, Murray DM. Analytic methods for individually randomized group treatment trials and group-randomized trials when subjects belong to multiple groups. Statistics in Medicine. 2014;33(13):2178-90. PMC4013262.
- Roberts C, Walwyn R. Design and analysis of non-pharmacological treatment trials with multiple therapists per patient. Statistics in Medicine. 2013;32(1):81-98.

Summary

- GRTs require analyses that reflect the nested designs inherent in these studies.
- Used alone, 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.
  - For designs having one or two time intervals, mixed-model ANOVA/ANCOVA is recommended.
  - For designs having three or more time intervals, random coefficients models are recommended.
- Other methods can be used effectively, with proper care, including randomization tests, GEE, and two-stage methods.

Summary

- Other approaches are not appropriate, including analysis at a subgroup level, deleting the unit of assignment if it or the ICC is not significant, designs with one group per condition, and Kish's effective df.
- Unbalanced designs can create analytic problems and an inflated Type I error rate; special methods are required.
- Constrained randomization can be helpful.
- IRGTs face similar problems to GRTs and the solutions are similar: model the small groups or common change agents as nested random effects, with implications for df and testing.

Pragmatic and Group-Randomized Trials in Public Health and Medicine

Visit **https://prevention.nih.gov/grt** to:

- Provide feedback on this series
- Download the slides, references, and suggested activities
- View this module again
- View the next module in this series:
  Part 4: Power and Sample Size

Send questions to:

GRT@mail.nih.gov

National Institutes of Health
Office of Disease Prevention